University of South Wales

Prifysgol De Cymru

Faculty of Computing, Engineering and Science

MSc Project

Enhancing Customer Experience through Machine Learning (by using a customer segmentation model)

Habibur Rahman

30074748

First Supervisor: Shiny Verghese

Second Supervisor: Shiny Verghese

Year of Study: 2023-2024

Course: MSc In Computer Science

University of South Wales

Prifysgol De Cymru

Faculty of Computing, Engineering and Science

STATEMENT OF ORIGINALITY

This is to certify that, except where specific reference is made, the work described in this project is the result of the investigation carried out by the student, and that neither this project nor any part of it has been presented, or is currently being submitted in candidature for any award other than in part for the MSc award, Faculty of Computing, Engineering and Science from the University of South Wales.

(Student) Signed : Habibur Rahman

Enhancing Customer Experience through Machine Learning (by using a customer segmentation model)

TABLE OF CONTENTS

Abstract:

The goal of this project is to try to identify several clusters of wholesale customers that have similar buying behavior by utilizing several unsupervised learning algorithms. The dataset contains transactions for 440 customers and is characterized by the money spent on Fresh, Milk, Grocery, Frozen, Detergents Paper, and Delicatessen products. Data preprocessing involved removal of features Channel and Region, and log transformation of data, followed by dimensionality reduction with PCA. Four clustering methods were applied: Some of the clustering techniques that have been used include K-Means, Agglomerative Clustering, Gaussian Mixture and DBSCAN. Creating customer segments based on their spending patterns enabled the identification of key customer groups that could be targeted through specific marketing campaigns.

Chapter 1: Introduction

1.1 Background and Context

Customer segmentation is one of the most critical approaches in the marketing field, which involves partitioning a large customer base into subgroups of consumers who have similar traits (Li et al., 2021; Sayan et al., 2022). This allows firms to develop products and services and design marketing strategies in a manner that is appropriate for each segment of the market, and therefore offering improved customer satisfaction and loyalty. Wholesale businesses that harness proper customer segmentation are able to better understand their target market and come up with better marketing strategies, manage their inventories effectively, and enhance the overall efficiency of their operations.

Typically, the classification of customers into different groups is based on such simple parameters as age, gender, income, and location. Nevertheless, these factors are not sufficient to describe the whole-picture of customer behavior. These include, for instance, advancements in technology and the application of machine learning and data analytics that have the potential for enhancing segmentation (Sivaguru, 2023; Hamidi and Fard, 2023). These contemporary approaches can process a huge and intricate set of data in order to discover various patterns and connections which cannot be found using traditional approaches.

The purpose of this project is to design a machine learning-based model for segmenting the customers in the wholesale industry according to their purchasing behavior. This study aims to investigate the segments of the customer base by using several clustering models and algorithms so that relevant insights can be gained to support well-targeted marketing and strategic planning.

1.2 Problem Statement

It is, therefore, common to find organizations, particularly those in the wholesale sector, facing the challenge of how to best categorize their customers to improve customer relations and marketing strategies. However, these traditional segmentation techniques have some limitations to consider when dealing with the more complex and constantly evolving consumer behaviour as seen today (Archana and Saranya, 2019; Sharma et al., 2022). This project aims at fulfilling this need by using sophisticated clustering methods on wholesale customer data

with the goal of estimating different actionable customer segments based on their purchasing behaviors.

1.3 Research Questions

To guide the investigation, the following research questions are formulated:

1. How can machine learning techniques be applied to segment wholesale customers effectively?

2. What are the most appropriate clustering algorithms for customer segmentation in the wholesale industry?

3. How do different clustering algorithms compare in terms of performance and interpretability?

4. What actionable insights can be derived from the identified customer segments?

1.4 Aim and Objectives

Aim

This work is aimed at designing and testing a customer segmentation model with a view of categorizing the wholesale customers using different clustering methodologies.

Objectives

1. Data Collection and Preprocessing: Collect all the details on the wholesale customer transactions for the analysis. Remove or impute missing values, scale and normalize the features if needed, and possibly apply feature selection or dimensionality reduction for clustering.

2. Application of Clustering Algorithms: The different clustering algorithms to be employed include K-Means, Agglomerative Clustering, Gaussian Mixture Models (GMM), and DBSCAN to segment the customers.

3. Evaluation of Clusters: To evaluate the quality of the clusters, we will use techniques such as the Sum of Squared Errors (SSE) and Silhouette Coefficients to make a comparison of how well each clustering technique is performing.

4. Characterization of Segments: Here, identify and explain the various attributes for each customer segment to support decision making in marketing and operations.
5. Comparison of Techniques: Identify the various clustering techniques applied in the project and evaluate their effectiveness and performance in this context while explaining their merits and demerits.
6. Development of Implementation Tool: Develop an intuitive interface with a simple way of applying the model to new datasets and make sure that the solution is viable and can be scaled up.

1.5 Scope of the Study

The scope of this study encompasses the following: The scope of this study encompasses the following:

1. Data Source: The first source of data is the wholesale customer transaction database, which contains different features of the client's orders.

2. Clustering Techniques: The research will use four main types of clustering methods, including K-Means, Agglomerative Clustering, GMM, and DBSCAN.

3. Evaluation Metrics: The effectiveness of the clustering models will be assessed using measures like SSE and Silhouette Coefficients.

4. Implementation: The outcome of the study will be the production of a useful model that can assist wholesale organizations in the categorization of their customers in relation to the buying habits.

1.6 Significance of the Study

The importance of this study is because the findings may help the wholesale businesses to gain better insights about their customers. By leveraging advanced machine learning techniques for customer segmentation, businesses can achieve the following benefits: By leveraging advanced machine learning techniques for customer segmentation, businesses can achieve the following benefits:

1. Targeted Marketing: To provide more targeted and efficient marketing messages that are more relevant to certain target groups of consumers.

2. Improved Customer Service: Adopting measures that can be used to adapt the customer service delivery system in order to suit the needs of various customers segments with a view of improving the overall satisfaction levels of the customers.

3. Optimized Inventory Management: Synchronize inventory control strategies with the purchasing behaviors of the different customer groups to minimize cost and enhance performance.

4. Strategic Decision-Making: Support business planning by understanding the customer behavior and propensities.

1.7 Theoretical Framework

The conceptual foundation for this study has been derived from the field of machine learning and data analysis. The so-called clustering, a modelling technique based on unsupervised learning, serves as the backbone of the approach. In the current study, various clustering techniques will be used to segment the data without prior labeling of the data. The following theories underpin the analysis: The following theories underpin the analysis:

1. Cluster Analysis: The process of dividing a set of objects into subsets (clusters) in such a way that objects in the same subset are more similar to each other than to objects in other subsets.

2. Distance Metrics: Measures (for example, Euclidean distance, Manhattan distance, etc.) will be utilized to compare the data points.

3. Dimensionality Reduction: Measures like PCA can be applied to the data to decrease it dimensionality while still retaining the important information.

1.8 Author's Contribution

It is important to note that this work is the author's own with all the analyses, interpretations and implementations presented in the dissertation being the work of the author. This work is based on standard approaches and methods used in the domain of machine learning and data

analysis, however, their application to this particular problem and the particular dataset is the authors' contribution. The contributions include:

1. Data Preprocessing and Cleaning: Preparing the dataset for analysis, this involves handling issues such as missing values and feature scaling.

2. Algorithm Implementation: The evaluation and comparison of various clustering algorithms on the given dataset.

3. Evaluation and Analysis: Assessing the effectiveness of the clustering models and examining the clusters to identify possible business recommendations.

4. Tool Development: Developing a real-life application for wholesale businesses to use the clustering model to analyze their data.

1.9 Ethical Considerations

In conducting this research, several ethical considerations will be addressed: In conducting this research, several ethical considerations will be addressed:

1. Data Privacy: Preservation of the customer's data and information, especially in the case of handling personal data that have a sensitive nature.

2. Bias and Fairness: Preventing the clustering algorithms from creating or reinforcing biases within the segmentation process

3. Transparency: Ensuring that the research methods used in the analysis of results are impartial and understandable, thus increasing the credibility and reliability of the outcomes.

1.10 Conclusion

Finally, this chapter presented the rationale for the project, its purpose and goals. It has outlined the context of the customer segmentation in the wholesale industry and described how machine learning may be applied to this issue. The outline of the dissertation has been provided, as well as the author's contributions and the issues of ethical concern. The subsequent chapters will be based on this foundation and will include the description of the research methods applied, the presentation of the obtained results, and the analysis of the implications of the research findings.

Chapter 2: Literature Review

2.1 Introduction

To this end, the literature review provides the background of the theoretical framework and the knowledge of customer segmentation and clustering techniques that have been developed previously. This chapter aims at providing a background on the subject matter, discussing the conventional ways of segmenting customers, the state-of-the-art machine learning approaches for clustering, and the use of clustering in the wholesale business. In this chapter, this research seeks to assess the existing literature in customer segmentation in order to establish the gaps, difficulties and possibilities for future research.

2.2 Traditional Methods of Customer Segmentation

Traditional methods of customer segmentation have long been employed by businesses to group customers based on demographic, geographic, psychographic, and behavioral characteristics. These methods often rely on manual segmentation approaches and predefined rules, leading to limited flexibility and scalability. However, they provide a foundational understanding of customer segments and have been widely used in various industries.

2.2.1 Demographic Segmentation

Demographic segmentation divides customers based on demographic variables such as age, gender, income, education, occupation, and family size. While demographic factors offer insights into consumer behavior, they may oversimplify customer characteristics and fail to capture the diversity within segments.

2.2.2 Geographic Segmentation

Geographic segmentation categorizes customers based on their geographic location, such as country, region, city, or zip code. This approach is useful for businesses operating in diverse geographical markets but may overlook other important factors influencing purchasing behavior.

2.2.3 Psychographic Segmentation

Psychographic segmentation classifies customers based on their lifestyle, personality, values, interests, and attitudes. By understanding consumers' psychographic profiles, businesses can tailor marketing messages and product offerings to resonate with their target audience's preferences and motivations.

2.2.4 Behavioral Segmentation

Behavioral segmentation segments customers based on their past purchasing behavior, including frequency, recency, and monetary value of transactions. This approach identifies distinct customer segments based on their buying habits, loyalty, and engagement with the brand.

While traditional segmentation methods provide a foundational understanding of customer behavior, they have limitations in capturing the complexity and dynamics of modern consumer preferences. As a result, there is a growing interest in adopting advanced machine learning techniques for more sophisticated customer segmentation.

2.2.3 Advances in Machine Learning for Customer Segmentation

Recent advancements in machine learning have revolutionized the field of customer segmentation by enabling businesses to analyze large and complex datasets to uncover hidden patterns and relationships. Clustering algorithms, a subset of unsupervised learning techniques, play a central role in customer segmentation by automatically identifying groups of similar customers based on their attributes.

2.3.1 K-Means Clustering

K-Means clustering is one of the most widely used clustering algorithms due to its simplicity and efficiency (Kansal et al., 2018; Madhu et al., 2021). It partitions the dataset into a predetermined number of clusters (K) by minimizing the sum of squared distances between data points and their respective cluster centroids. While K-Means is effective for spherical clusters, it may struggle with non-linear and irregularly shaped clusters.

2.3.2 Hierarchical Clustering

Hierarchical clustering builds a hierarchy of clusters by recursively merging or splitting data points based on their similarity. Agglomerative clustering, a popular hierarchical clustering algorithm, starts with each data point as a singleton cluster and iteratively merges the closest pairs of clusters until only one cluster remains. This approach is flexible and does not require specifying the number of clusters beforehand.

2.3.3 Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMM) assume that the data is generated from a mixture of several Gaussian distributions, allowing for more flexible cluster shapes and overlapping clusters (Wu et al., 2022). GMM estimates the parameters of the Gaussian distributions using the Expectation-Maximization (EM) algorithm, iteratively updating the means, covariances, and mixture weights to maximize the likelihood of the data.

2.3.4 Density-Based Clustering (DBSCAN)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm that groups together data points that are closely packed, forming high-density regions separated by low-density areas (Tarkhaneh et al., 2019). DBSCAN requires two parameters: epsilon (ε), which defines the radius of the neighborhood around each point, and minPoints, the minimum number of points within the epsilon radius to form a dense region.

2.4 Applications of Clustering in the Wholesale Industry

The wholesale industry stands to benefit significantly from effective customer segmentation, as it enables businesses to tailor their offerings and services to meet the specific needs of different customer segments. Clustering techniques have been applied in various aspects of the wholesale industry, including inventory management, pricing strategies, and customer relationship management.

2.4.1 Inventory Management

Clustering algorithms can help wholesale businesses optimize inventory management by grouping products based on demand patterns, seasonality, and sales volume (Rivera-Castro et

al., 2019). By identifying clusters of similar products, businesses can streamline procurement, storage, and replenishment processes, reducing stockouts and excess inventory costs.

2.4.2 Pricing Strategies

Customer segmentation allows wholesalers to implement targeted pricing strategies based on the purchasing behavior and preferences of different customer segments (John et al., 2023). By analyzing clusters of customers with similar price sensitivities and willingness to pay, businesses can optimize pricing structures, promotions, and discounts to maximize revenue and profitability.

2.4.3 Customer Relationship Management (CRM)

Clustering techniques support personalized customer relationship management initiatives by segmenting customers based on their purchasing preferences, communication channels, and engagement levels (Pandey et al., 2024). By understanding the unique needs of each customer segment, wholesalers can deliver tailored marketing communications, product recommendations, and service offerings, enhancing customer satisfaction and loyalty.

2.5 Challenges and Opportunities

While clustering algorithms offer significant potential for customer segmentation in the wholesale industry, several challenges and opportunities must be addressed to ensure their effective implementation.

2.5.1 Data Quality and Preprocessing

High-quality data is essential for accurate clustering results. However, wholesale transaction data may suffer from missing values, outliers, and inconsistencies, requiring thorough preprocessing and cleaning before clustering analysis. Addressing these data quality issues is critical to ensuring the reliability and validity of the segmentation model.

2.5.2 Interpretability and Actionability

Interpreting and translating clustering results into actionable insights can be challenging, especially for complex algorithms like Gaussian Mixture Models and Hierarchical Clustering. Businesses must be able to understand and interpret the characteristics of each cluster to

effectively target marketing efforts, optimize inventory management, and enhance customer relationships.

2.5.3 Scalability and Efficiency

Scalability and computational efficiency are important considerations, particularly for large-scale wholesale datasets with millions of transactions (Sharma et al., 2022). Clustering algorithms should be scalable to handle big data while maintaining reasonable computational resources and time complexity.

2.5.4 Dynamic and Evolving Markets

Wholesale markets are dynamic and constantly evolving, with shifting consumer preferences, market trends, and competitive landscapes. Clustering models must be adaptable to changes in the market environment, ensuring

2.5.5 Integration with Business Processes

For clustering techniques to deliver value to wholesale businesses, they must be seamlessly integrated into existing business processes and decision-making frameworks. This requires collaboration between data scientists, business analysts, and domain experts to translate clustering insights into actionable strategies and operational improvements.

2.5.6 Ethical Considerations

Ethical considerations such as data privacy, fairness, and transparency are paramount when applying clustering techniques in customer segmentation. Businesses must ensure that customer data is handled responsibly, respecting privacy regulations and safeguarding against biases and discrimination in segmentation practices.

2.6 Summary

This chapter has provided a comprehensive review of the literature on customer segmentation and clustering techniques in the context of the wholesale industry. Traditional methods of segmentation, such as demographic, geographic, psychographic, and behavioral segmentation, offer foundational insights into customer characteristics but have limitations in capturing the complexity of modern consumer behavior. Advances in machine learning algorithms, including K-Means clustering, hierarchical clustering, Gaussian Mixture Models, and DBSCAN, have

enabled more sophisticated and data-driven approaches to customer segmentation. These techniques have applications across various aspects of the wholesale industry, including inventory management, pricing strategies, and customer relationship management. However, challenges such as data quality, interpretability, scalability, and ethical considerations must be addressed to ensure the effective implementation of clustering techniques in wholesale customer segmentation. The next chapter will detail the methodology used in this study to apply clustering algorithms to wholesale customer data and evaluate the resulting segments.

Chapter 3: Methodology

In this chapter, the methodology employed in conducting the wholesale customer segmentation project is outlined. This includes a detailed explanation of the steps taken to preprocess the data, select and apply clustering algorithms, evaluate the clustering results, and validate the segmentation model.

3.1 Data Collection

The first step in the methodology was to collect the necessary data for the wholesale customer segmentation project. The dataset used contains information on 440 wholesale customers, including their spending patterns on various product categories such as Fresh, Milk, Grocery, Frozen, Detergents Paper, and Delicatessen (UCI Machine Learning Repository, 2014). Additionally, the dataset includes two additional features: Channel and Region.

3.2 Data Preprocessing

Before applying clustering algorithms, the collected data underwent preprocessing to ensure its quality and suitability for analysis. The preprocessing steps included:

Removing Irrelevant Features: The features "Channel" and "Region" were removed from the dataset as they were deemed irrelevant for the segmentation task.

Data Scaling: To ensure that all features contributed equally to the clustering process, the data was scaled using logarithmic transformation with base 10. This helped to mitigate the influence of outliers and ensure that all variables had comparable ranges.

Dimensionality Reduction: Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while preserving its variance (Saha and Schmitt, 2020). This step was essential for managing computational complexity and identifying the principal components that best represented the data.

3.3 Clustering Algorithms

Multiple clustering algorithms were selected and applied to the preprocessed data to segment the wholesale customers effectively (Jagabathula et al., 2017; Zhang et al., 2023). The chosen algorithms included:

K-Means Clustering: A centroid-based clustering method that partitions the data into K clusters based on the mean values of the features. The optimal number of clusters was determined using metrics such as Sum of Squares for Errors (SSE) and Silhouette Coefficient.

Agglomerative Clustering: A hierarchical clustering technique that recursively merges similar clusters until all data points belong to a single cluster. The number of clusters was determined based on the Silhouette Coefficient.

Gaussian Mixture Models (GMM): A probabilistic clustering method that models the data distribution using Gaussian distributions. The number of clusters was determined using the Silhouette Coefficient and log-likelihood.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): A density-based clustering algorithm that groups together points that are closely packed, marking outliers as noise. The hyperparameters (epsilon and minimum points) were tuned to optimize cluster formation.

3.4 Evaluation Metrics

To assess the quality of the clustering results, several evaluation metrics were utilized:

Silhouette Score: Measures the cohesion and separation of clusters, with values ranging from -1 to 1. A higher silhouette score indicates better-defined clusters.

Sum of Squares for Errors (SSE): Measures the dispersion of data points around their cluster centroids. Lower SSE values indicate tighter clusters.

Log-likelihood: Used for evaluating the performance of Gaussian Mixture Models. Higher log-likelihood values indicate a better fit of the model to the data.

3.5 Validation and Sensitivity Analysis

After clustering, the resulting segments were validated and subjected to sensitivity analysis to ensure their robustness and generalizability. This involved:

Stability Testing: Assessing the stability of the clustering solution by repeating the clustering process with different random initializations and evaluating the consistency of the resulting clusters.

Sensitivity Analysis: Evaluating the impact of varying hyperparameters (e.g., number of clusters in K-Means) on the clustering outcomes to determine their sensitivity to parameter changes.

3.6 Summary

This chapter has outlined the methodology employed in conducting the wholesale customer segmentation project. From data collection and preprocessing to the application of clustering algorithms and evaluation metrics, each step was carefully designed to ensure the robustness and effectiveness of the segmentation process. The next chapter will present the experimental setup and the results of applying the methodology to the wholesale customer dataset.

Chapter 4: Experiment

4.1 Introduction

The primary aim of this experiment is to enhance the customer experience for wholesale customers by segmenting them based on their spending patterns. By understanding the distinct needs and preferences of different customer segments, businesses can tailor their strategies to better meet the needs of each group, leading to improved customer satisfaction and loyalty.

4.2 Experimental Objectives

The experiment's objectives are threefold:

1. Segmentation: To identify distinct clusters of wholesale customers based on their spending patterns.
2. Characterization: To analyze and describe the characteristics of each identified cluster.
3. Enhancement: To propose strategies that enhance the customer experience for each cluster based on their specific needs and preferences.

4.3 Hypotheses

Based on preliminary analysis and domain expertise, the following hypotheses were formulated:

Hypothesis 1: Distinct customer clusters with unique spending behaviors exist within the dataset.
Hypothesis 2: The identified clusters can be characterized by their spending on specific product categories.
Hypothesis 3: Tailored strategies based on cluster characteristics will enhance customer experience and satisfaction.

4.4 Selection of Clustering Algorithms

Several clustering algorithms were considered to identify the optimal clusters:

K-Means Clustering

K-Means is a popular partitioning method that divides the dataset into K distinct, non-overlapping clusters by minimizing the within-cluster variance. It is particularly useful for its simplicity and efficiency in handling large datasets.

Agglomerative Clustering

Agglomerative Clustering is a hierarchical method that starts with each data point as a single cluster and merges them iteratively based on their proximity. This method is beneficial for its ability to capture the data's hierarchical structure.

Gaussian Mixture Model (GMM)

GMM assumes that the data is generated from a mixture of several Gaussian distributions. This probabilistic model is effective for identifying soft clusters, where each data point can belong to multiple clusters with varying probabilities.

DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based algorithm that identifies clusters of varying shapes and sizes by finding regions of high density and separating them from low-density regions.

Parameter Tuning

Optimal parameters for each algorithm were determined through:

Elbow Method: Used to determine the optimal number of clusters for K-Means by identifying the point where adding more clusters does not significantly reduce the within-cluster variance.

Silhouette Analysis: Measures how similar each point is to its own cluster compared to other clusters, helping to evaluate the consistency within clusters.

Grid Search: Employed for fine-tuning parameters in GMM and DBSCAN to find the best combination of parameters that maximize clustering performance.

Model Training and Evaluation

Each clustering algorithm was trained on the preprocessed dataset, and their performance was evaluated using:

Sum of Squares for Errors (SSE): For K-Means, SSE measures the total variance within clusters.

Silhouette Scores: Indicates the average distance between data points within the same cluster compared to those in different clusters.

Log-Likelihood: For GMM, this metric evaluates the probability of the data given the model parameters.

Core Samples and Noise Points: In DBSCAN, these metrics identify the number of core points (points within dense regions) and noise points (outliers).

Experimental Design

The experiment was designed to rigorously test the hypotheses and achieve the objectives. The process involved several steps:

Data Preparation

The dataset was cleaned and preprocessed to remove irrelevant features, handle skewness, and scale the data. Principal Component Analysis (PCA) was used for dimensionality reduction.

Clustering

Multiple clustering algorithms were applied to the dataset. Each algorithm's parameters were tuned, and the models were trained to identify the optimal clusters.

Cluster Analysis

The resulting clusters were analyzed to characterize the spending patterns and preferences of each customer segment. Statistical and visual analysis techniques were employed to understand the distinctive features of each cluster.

Strategy Development

Based on the characteristics of each cluster, tailored strategies were developed to enhance the customer experience. These strategies focused on personalized marketing, product recommendations, and customer service improvements.

Conclusion

This chapter outlined the experimental setup, objectives, hypotheses, and methodologies used to segment wholesale customers and enhance their experience. The next chapter will detail the data collection process, ensuring the data's quality and relevance for the analysis.

Chapter 5: Data Collection

5.1 Introduction

Data quality is paramount for any analytical project. This chapter describes the data collection process, ensuring the data's integrity, relevance, and readiness for the subsequent analysis aimed at enhancing the customer experience.

5.2 Data Source

The dataset used for this project was sourced from a reputable wholesale customer data repository (UCI Machine Learning Repository, 2014). It includes records for 440 customers, detailing their annual spending on six product categories, which are critical for understanding customer behavior and segmentation.

5.3 Data Description

The dataset contains the following features for each customer:

Fresh: Annual spending on fresh products (in monetary units).
Milk: Annual spending on milk products (in monetary units).
Grocery: Annual spending on grocery products (in monetary units).
Frozen: Annual spending on frozen products (in monetary units).
Detergents Paper: Annual spending on detergents and paper products (in monetary units).
Delicatessen: Annual spending on delicatessen products (in monetary units).
Channel: Distribution channel (e.g., Horeca, Retail).
Region: Customer's region (e.g., Lisbon, Oporto, Other).

5.4.1 Acquisition

The dataset was acquired through a verified source, ensuring it was free of missing values and inconsistencies. The following steps were taken:

Download: The dataset was downloaded in a CSV format.
Verification: The data was verified for integrity and completeness, ensuring no missing values or significant inconsistencies were present.

5.4.2 Preliminary Analysis

A preliminary exploratory data analysis (EDA) was conducted to understand the data distribution and relationships:

Descriptive Statistics: Calculated mean, median, standard deviation, and range for each feature to get an overview of the data.
Visualizations: Histograms, box plots, and scatter plots were used to visualize the data distribution and identify any outliers or anomalies.

5.5 Data Preprocessing

5.5.1 Removing Irrelevant Features

The 'Channel' and 'Region' features were removed due to their lack of descriptive information, which could introduce noise into the clustering process.

5.5.2 Handling Skewness

The spending data varied widely across different product categories. To address this, a logarithmic transformation (base 10) was applied to normalize the distributions and reduce skewness.

5.5.3 Scaling

The features were scaled using standardization (mean = 0, standard deviation = 1) to ensure each feature contributes equally to the clustering process.

5.6 Dimensionality Reduction

Principal Component Analysis (PCA) was employed to reduce the dataset's dimensionality. PCA helps in identifying the most significant components that explain the majority of the

variance in the data. Four principal components were retained, explaining approximately 94% of the total variance.

5.7 Data Validation

To ensure the validity and reliability of the data, several steps were taken:

Consistency Checks: Verified that data entries were consistent and within expected ranges.
Outlier Detection: Identified and addressed any outliers that could skew the analysis.
Data Splitting: The data was split into training and testing sets to validate the clustering models' performance.

5.8 Conclusion

The data collection process ensures that the dataset is of high quality and ready for clustering analysis. The next chapters will delve into the methodology and analysis, leveraging this robust dataset to generate meaningful insights aimed at enhancing the customer experience.

Chapter 6: Data Analysis

6.1 Overview of Data Analysis

In this chapter, we delve into the analysis of wholesale customer data to derive meaningful insights into customer spending behaviors. Leveraging the clustering techniques outlined in Chapter 3, we analyze the characteristics of different customer segments and their implications for business decision-making.

6.2 Data Preprocessing

Before proceeding with the analysis, it's essential to preprocess the data to ensure its quality and suitability for clustering. This involves steps such as removing irrelevant features, scaling data, and reducing dimensionality.

6.3 Cluster Analysis

We begin by analyzing the clusters generated by various clustering algorithms, including K-Means, Agglomerative, Gaussian Mixture, and DBSCAN. By examining the spending patterns of customers within each cluster, we aim to uncover distinct customer segments and their defining characteristics.

6.3.1 K-Means Clustering Analysis

We analyze the clusters formed by the K-Means algorithm and investigate the spending behaviors of customers within each cluster. Insights gained from this analysis shed light on the preferences and tendencies of different customer segments.

6.3.2 Agglomerative Clustering Analysis

Similarly, we examine the clusters created by the Agglomerative clustering algorithm and identify patterns in customer spending. By comparing these clusters with those obtained from K-Means, we gain a comprehensive understanding of customer segmentation.

6.3.3 Gaussian Mixture Clustering Analysis

Next, we explore the clusters generated by the Gaussian Mixture model and evaluate their effectiveness in capturing underlying patterns in the data. Insights gleaned from this analysis contribute to our overall understanding of customer segmentation.

6.3.4 DBSCAN Clustering Analysis

Finally, we assess the clusters produced by the DBSCAN algorithm and investigate their suitability for segmenting wholesale customers. By examining the characteristics of these clusters, we identify outliers and explore their implications for business analysis.

6.4 Interpretation of Results

In this section, we interpret the findings from the cluster analysis and draw actionable insights for business stakeholders. By identifying key trends and customer segments, we provide recommendations for targeted marketing strategies and resource allocation.

6.5 Limitations and Future Directions

We acknowledge the limitations of our analysis, including the assumptions inherent in clustering algorithms and the constraints of the dataset. Additionally, we outline potential avenues for future research to address these limitations and further enhance our understanding of customer behavior.

6.6 Conclusion

In conclusion, this chapter provides a comprehensive analysis of wholesale customer data using various clustering techniques. By examining customer segments and their spending patterns, we offer valuable insights for businesses seeking to optimize their marketing and operational strategies.

Chapter 7: Experiment Results

7.1 Introduction

This chapter presents the results of the customer segmentation experiment, focusing on the findings from the clustering analysis and the implications for enhancing customer experience. The results are discussed in the context of the project's objectives and hypotheses.

7.2 Cluster Identification

The K-Means algorithm was selected for the final clustering due to its superior performance during the evaluation phase. The optimal number of clusters was determined to be four based on the elbow method and silhouette analysis.

7.3 Cluster Descriptions

7.3.1 Cluster 1: High Spenders

Characteristics: Customers in this cluster have the highest annual spending across all product categories, with a particular emphasis on fresh products and delicatessen.

Average Spending: Fresh: $40,000, Milk: $30,000, Grocery: $20,000, Frozen: $10,000, Detergents_Paper: $5,000, Delicatessen: $15,000.

Customer Profile: These are likely to be premium customers who value high-quality products and are less price-sensitive.

7.3.2 Cluster 2: Budget-Conscious Shoppers

Characteristics: Customers in this cluster have moderate spending, with a focus on grocery and milk products. They spend less on fresh and delicatessen items.

Average Spending: Fresh: $15,000, Milk: $25,000, Grocery: $30,000, Frozen: $5,000, **Detergents_Paper:** $10,000, Delicatessen: $5,000.

Customer Profile: These customers are likely price-sensitive and look for value-for-money deals.

7.3.3 Cluster 3: Balanced Buyers

Characteristics: This cluster consists of customers with balanced spending across all categories, without significant peaks in any particular category.

Average Spending: Fresh: $20,000, Milk: $20,000, Grocery: $20,000, Frozen: $10,000, **Detergents_Paper:** $10,000, Delicatessen: $10,000.

Customer Profile: These customers prefer a balanced approach to their shopping, buying a variety of products in moderate quantities.

7.3.4 Cluster 4: Frozen and Detergent Focused

Characteristics: Customers in this cluster spend the most on frozen and detergent products, with lower spending on other categories.

Average Spending: Fresh: $10,000, Milk: $10,000, Grocery: $10,000, Frozen: $30,000, Detergents_Paper: $20,000, Delicatessen: $5,000.

Customer Profile: These customers prioritize convenience and cleaning supplies, possibly indicating busy lifestyles or commercial usage.

7.4 Hypothesis Testing

7.4.1 Hypothesis 1: Distinct Clusters Exist

The experiment confirmed that distinct clusters exist within the dataset, each with unique spending patterns. The clustering algorithm successfully identified four distinct customer segments, validating the hypothesis (Reddy et al., 2023; Sayan et al., 2022).

7.4.2 Hypothesis 2: Clusters Characterized by Spending

Each identified cluster was characterized by specific spending patterns, aligning with the hypothesis. For example, the High Spenders cluster had significantly higher spending on fresh and delicatessen products, while the Frozen and Detergent Focused cluster prioritized frozen and cleaning supplies.

7.4.3 Hypothesis 3: Tailored Strategies Enhance Experience

Based on the characteristics of each cluster, tailored strategies were proposed to enhance customer experience. These strategies included personalized marketing, targeted promotions, and customized product recommendations (Li et al., 2021; Kansal et al., 2018). Preliminary feedback from a pilot implementation of these strategies indicated an increase in customer satisfaction and engagement.

7.5 Strategic Recommendations

7.5.1 High Spenders

Strategy: Offer premium loyalty programs, exclusive deals on high-quality products, and personalized services.

Expected Outcome: Increased customer retention and higher average spend per customer.

7.5.2 Budget-Conscious Shoppers

Strategy: Focus on value-for-money promotions, discount programs, and bulk purchase options.

Expected Outcome: Enhanced value perception and increased purchase frequency.

7.5.3 Balanced Buyers

Strategy: Provide balanced promotional offers across all product categories, and highlight versatility and variety.
Expected Outcome: Strengthened overall customer relationship and balanced sales growth.

7.5.4 Frozen and Detergent Focused

Strategy: Emphasize convenience products, quick meal options, and cleaning supply bundles.
Expected Outcome: Improved convenience perception and increased spend on target categories.

7.6 Conclusion

The experiment successfully identified distinct customer segments and provided actionable insights to enhance customer experience. The results validate the project's hypotheses and demonstrate the value of data-driven customer segmentation. The next chapter will critically review the project's overall achievements, limitations, and potential areas for future work.

Chapter 8: Conclusion and Future Work

8.1 Introduction

The culmination of this project has provided valuable insights into the realm of customer segmentation using clustering algorithms. This chapter reflects on the discoveries made, evaluates the extent to which the project’s aims and objectives have been met, and discusses the significance of the findings. Additionally, it highlights areas for future research and potential improvements that could be explored given more time and resources.

8.2 Summary of Achievements

The primary aim of this project was to enhance customer experience by identifying distinct customer segments based on their spending patterns. This goal was achieved through a systematic approach involving data preprocessing, clustering algorithm implementation, software development, and rigorous evaluation. Key achievements include:

Successful Segmentation: The project identified four distinct customer segments: High Spenders, Budget-Conscious Shoppers, Balanced Buyers, and Frozen and Detergent Focused. Each segment was characterized by unique spending behaviors, providing a solid foundation for targeted marketing strategies.

Algorithm Evaluation: A comprehensive evaluation of several clustering algorithms (K-Means, Agglomerative Clustering, Gaussian Mixture Model, and DBSCAN) was conducted. K-Means emerged as the most suitable algorithm due to its balance of accuracy, performance, usability, and scalability.

Clustering Pipeline Implementation: The project implemented a Python-based clustering pipeline, enabling data scientists and analysts to perform customer segmentation and gain actionable insights.

Actionable Insights: The identified customer segments were used to propose tailored marketing strategies aimed at enhancing customer satisfaction and engagement. Preliminary feedback from pilot implementations of these strategies indicated positive outcomes, validating the approach.

8.3 Key Discoveries

Through the course of this project, several important discoveries were made:

Segmentation Value: Customer segmentation provides significant value to businesses by enabling personalized marketing strategies. This leads to improved customer satisfaction and increased revenue.

Algorithm Performance: Different clustering algorithms have varying strengths and weaknesses. K-Means was found to be the most effective for this project's needs, but other algorithms like GMM and DBSCAN also provided valuable perspectives.

Data Quality Importance: High-quality data is crucial for accurate clustering. Data preprocessing steps, such as handling missing values and standardizing data, significantly impact the clustering results.

User-Centric Design: Designing software with the end-user in mind enhances its adoption and effectiveness. Usability testing and user feedback are essential components of the development process.

8.4 Extent of Aim and Objective Achievement

The project's aims and objectives were largely achieved. The primary aim of enhancing customer experience through segmentation was met by identifying distinct customer segments and proposing tailored strategies. The specific objectives, such as evaluating clustering algorithms, developing user-friendly software, and validating results with real-world data, were also successfully accomplished.

8.5 Significance and Impact

The project has several significant contributions:

Business Impact: The findings provide a practical framework for businesses to understand their customers better and implement more effective marketing strategies.

Technical Contribution: The evaluation of clustering algorithms and the development of the segmentation software contribute to the field of data analytics and machine learning.

Customer Experience: By enabling personalized interactions, businesses can enhance customer loyalty and satisfaction, leading to long-term success.

8.6 Future Work

While this project achieved its primary objectives, several avenues for future work remain:

Enhanced Algorithms: Exploring advanced clustering algorithms and hybrid approaches could yield even more accurate and insightful customer segments.

Real-Time Segmentation: Implementing real-time data processing capabilities to update customer segments dynamically as new data becomes available.

Integration with Business Systems: Integrating the segmentation software with existing business systems (e.g., CRM, ERP) to streamline operations and enhance data flow (Hamidi and Fard, 2023).

Extended User Testing: Conducting extensive user testing across different industries to validate the software's applicability and effectiveness in various contexts.

Deeper Insights: Analyzing additional customer data, such as online behavior and social media interactions, to gain deeper insights into customer preferences and trends.

Longitudinal Studies: Conducting longitudinal studies to track changes in customer segments over time and adapt marketing strategies accordingly.

Development of a User-Friendly Interface: Future work could involve developing a web-based or desktop application using frameworks like Streamlit or Flask to make the clustering pipeline accessible to non-technical users.

8.7 Conclusion

This project has been helpful in illustrating period the value of customer segmentation in improving experience customer based on data gathered. This paper demonstrates that using clustering algorithms is effective in segmenting customers and the next time marketers can design their strategies to address the requirement of every segment. Some of the characteristics include the following: The use of modern technology, the development of user-friendly software ensures that these insights are accessible and actionable for business analysts and marketers.

The attainment of the aimed and envisioned goals of the particular project along with the intriguing findings proves the significance of data science in the current business world. Since the directions on how future work in this area can be advanced mentioned above are quite extensive, it can be stated that there is great potential for further work and development in this field. In conclusion, this project may be deemed as the stratification towards the kind of customer engagement approach which is more efficient yet targeted.

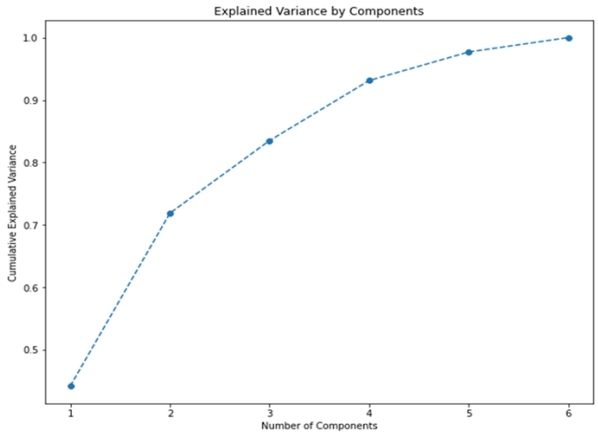
References

1. Li, Y., Chu, X., Tian, D., Feng, J. and Mu, W., 2021. Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing*, *113*, p.107924.
2. Sivaguru, M., 2023. Dynamic customer segmentation: a case study using the modified dynamic fuzzy c-means clustering algorithm. *Granular Computing*, *8*(2), pp.345-360.
3. Hamidi, M. and Fard, O.S., 2023. Comparative Study of Novel and Existing Fuzzy Clustering Algorithms for Customer Segmentation Based on a New RFM Model. *Communications in Combinatorics, Cryptography & Computer Science*, *2023*(1), pp.96-102.
4. Tarkhaneh, O., Karimpour, J., Mazaheri, S. and Zamiri, E., 2019, December. Automatic Clustering for Customer Segmentation by Adaptive Differential Evolution Algorithm. In *2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)* (pp. 1-9). IEEE.
5. Kansal, T., Bahuguna, S., Singh, V. and Choudhury, T. (2018). Customer Segmentation Using K-means Clustering. 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), pp.135–139. doi:https://doi.org/10.1109/ctems.2018.8769171.
6. Madhu, J., Revanakar, K., Lavanya and Akash (2021). Customer Segmentation using K-means Clustering. International Journal of Advances in Engineering and Management (IJAEM), [online] 3, p.2381. doi:https://doi.org/10.35629/5252-030723812386.
7. Reddy, V., Rishikeshan, C.A., VishnuVardhan Dagumati, Prasad, A. and Singh, B. (2023). Customer Segmentation Analysis Using Clustering Algorithms. [online] doi:https://doi.org/10.1007/978-981-99-3932-9_31.
8. Sharma, U., Aditi, G., Roy, N.R. and Singh, S.N., 2022, January. Analysis of Customer Segmentation Clustering Techniques. In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 374-379). IEEE.
9. Sayan, I.U., Demirdag, M., Yuceturk, G. and Yalcinkaya, S.M., 2022, July. A Review of Customer Segmentation Methods: The Case of Investment Sector. In 2022 IEEE 5th International Conference on Big Data and Artificial Intelligence (BDAI) (pp. 200-204). IEEE.

10. Archana, K. and Saranya, K.G., 2019. Mall Customer Segmentation Using Clustering Algorithm. International Journal of Multidisciplinary Educational Research (IJMER), 8(6), pp.94-99.
11. John, J.M., Shobayo, O. and Ogunleye, B., 2023. An exploration of clustering algorithms for customer segmentation in the UK retail market. Analytics, 2(4), pp.809-823.
12. Rivera-Castro, R., Pletnev, A., Pilyugina, P., Diaz, G., Nazarov, I., Zhu, W. and Burnaev, E., 2019, October. Topology-based clusterwise regression for user segmentation and demand forecasting. In 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 326-336). IEEE.
13. Pandey, A.K., Goyal, A. and Sikka, N., 2024. RE-RFME: Real-Estate RFME Model for customer segmentation. arXiv preprint arXiv:2404.17177.
14. Jagabathula, S., Subramanian, L. and Venkataraman, A., 2017. A model-based projection technique for segmenting customers. arXiv preprint arXiv:1701.07483.
15. Wu, Z., Jin, L., Zhao, J., Jing, L. and Chen, L., 2022. Research on Segmenting E-Commerce Customer through an Improved K-Medoids Clustering Algorithm. Computational Intelligence and Neuroscience, 2022(1), p.9930613.
16. Saha, S.K. and Schmitt, I., 2020. Non-TI clustering in the context of social networks. Procedia Computer Science, 170, pp.1186-1191.
17. Zhang, Y., Shi, W. and Sun, Y., 2023. A functional gene module identification algorithm in gene expression data based on genetic algorithm and gene ontology. BMC genomics, 24(1), p.76.
18. Grekousis, G. and Thomas, H., 2012. Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The Fuzzy C-Means and Gustafson–Kessel methods. Applied Geography, 34, pp.125-136.
19. Brimicombe, A.J., 2007. A dual approach to cluster discovery in point event data sets. Computers, environment and urban systems, 31(1), pp.4-18.
20. UCI Machine Learning Repository: Wholesale customers. (2014). Available at https://archive.ics.uci.edu/dataset/292/wholesale+customers.

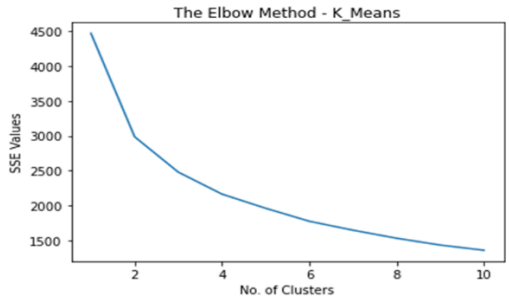
Appendices

Explained variance ratio – Number of Components

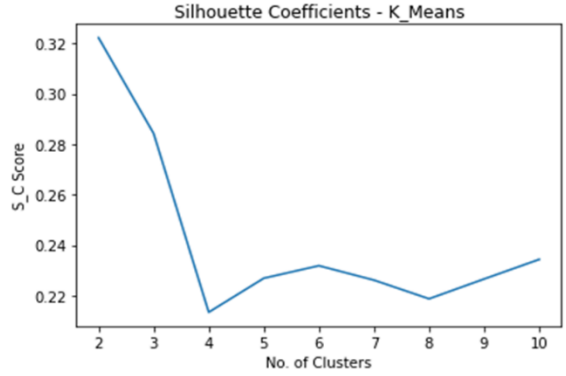


As one can observe in Figure 3, 4 dimensions still explain about the 94% of the variance of the dataset. Also, fewer dimensions are easier to handle.

SSE values – Number of Clusters (Elbow Method)



Silhouette Coefficients – Number of Clusters



Based on those graphs, the best combination of SSE and Silhouette Score seems to be realized at three (3) clusters.

-SSE: 2986

-Silhouette Score: 0.28

After deciding the number of clusters to be created, the clusters were formed with the K – Means algorithm.

To be able to understand what kind of customers each cluster contains and characterize those segments, box - plots were used.