Leveraging Clustering for Improved Marketing Strategy in E-commerce: A Customer Lifetime Value Approach

By:

Kanini Kagendo Gichuyia

149810

Master of Science in Data Science and Analytics

2024

Leveraging Clustering for Improved Marketing Strategy in E-commerce: A Customer Lifetime Value Approach

By:

Kanini Kagendo Gichuyia

149810

Submitted in Partial fulfilment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June, 2024

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Student's Name: Kanini Kagendo Gichuyia

Sign: _____ _____ Date: _____02.04.2024_____

Approval

The thesis of Kanini Kagendo Gichuyia was reviewed and approved for examination by the following:

Dr. Dickson Odhiambo Owuor,

Institute of Mathematical Sciences,

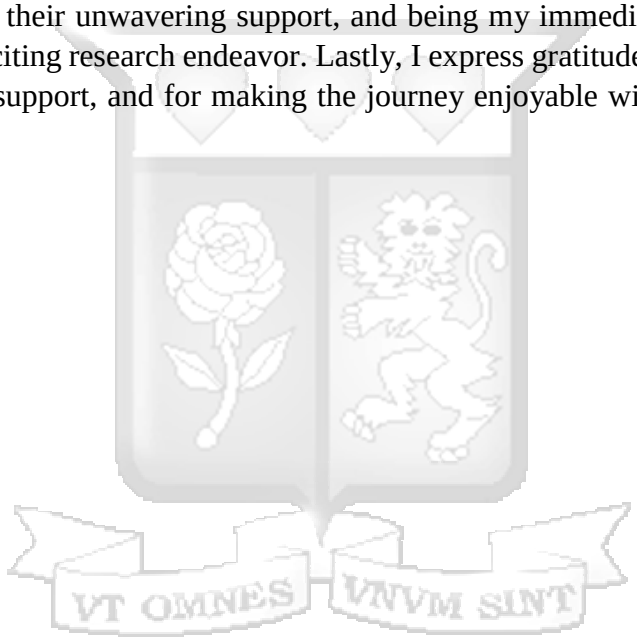
Strathmore University

02 April, 2024

Sign:_____ Date: _____

Acknowledgment

I extend my gratitude to God for His guidance and direction throughout this journey. I am deeply thankful to my supervisor, Dr. Dickson Owuor, for his invaluable assistance with every aspect of the thesis, from availing himself for meetings to offering crucial insights on data extraction and modeling decisions, and providing constructive feedback on the report. I am also indebted to Rhoda Mukami for granting me the opportunity to explore this fascinating research topic and providing access to pertinent data. Applying the knowledge acquired from my studies to real-world applications has been immensely fulfilling. My heartfelt appreciation goes to my family for their unwavering support, and being my immediate audience for their engagement in this exciting research endeavor. Lastly, I express gratitude to my classmates for their encouragement, support, and for making the journey enjoyable with their invitations to various activities.



ABSTRACT

In today's dynamic business landscape, characterized by a shift towards service-focused economies, companies are experiencing a transformative paradigm. They are proactively adapting to a new era, emphasizing the cultivation of enduring customer relationships as the linchpin of sustainable profitability. This strategic shift underscores the pivotal role of marketing, which extends beyond traditional paradigms to serve as the cornerstone for enhancing a company's financial performance. Within this context, marketing endeavors are geared towards augmenting what we refer to as "Customer Lifetime Value" (CLV), a multifaceted concept akin to a mosaic, encapsulating the cumulative value derived from loyal customers over time.

Various models, including the commonly used RFM (Recency, Frequency, Monetary) model, have been utilized in predicting customer lifetime value (CLV). The RFM model evaluates customers based on the recency, frequency, and monetary value of their transactions. Additionally, conventional methods like the widely used Elbow approach have been employed to determine the optimal number of clusters in CLV models. However, this study aims to explore CLV, particularly within the E-Commerce sector, by leveraging the analytical power of the Single Value Decomposition (SVD) clustering method. The paper underscores the critical significance of CLV models in navigating this intricate domain. These models serve as potent instruments for segmenting the market intelligently and optimizing resource allocation for marketing activities. In E-Commerce, where strategic decision-making is vital, businesses deploy these resources judiciously to acquire, retain, and cross-sell to customers, epitomizing the astute acumen required for E-Commerce success.

In the realm of E-Commerce, it has been customary to assess Customer Lifetime Value through the prism of Recency, Frequency, and Monetary (RFM) variables. However, it is essential to recognize that the relative importance of these variables undergoes dynamic interactions influenced by product or service attributes and industry-specific idiosyncrasies within the E-Commerce domain.

To encapsulate, this paper delves into the intricate facets of CLV, unveiling the potential of various CLV models empowered by the clustering method, using Single Value Decomposition approach to determine the most optimal clusters, as strategic assets for modern E-Commerce. These models serve as a compass for market segmentation and resource allocation, thereby sculpting the trajectory towards success for E-Commerce enterprises in the ever-evolving landscape of customer-centric commerce.

Key words: Customer Lifetime Value, Recency, Frequency and Monetary Model, Customer Relationship Management, Prediction

TABLE OF CONTENTS

List of Figures

List of Tables

List of Abbreviations

AOV - Average Order Value

API - Application Programming Interface

CRISP-DM - Cross Industry Standard Process for data Mining

CRM - Customer relationship management

CLV - Customer Lifetime Value

ML - Machine Learning

RFM - Recency , Frequency and Monetary

SVD-Singular Value Decomposition

1. INTRODUCTION

1.1 Background of the study

The rapid expansion of consumer data in recent decades has led to the widespread integration of data science across various sectors. In the e-commerce sector, understanding and predicting customer behavior is crucial for optimizing marketing strategies and maximizing profitability. Customer lifetime value (CLV or LTV) refers broadly to the revenue that a company can attribute to one or more customer over the length of their relationship with the company (Philip E Pfeifer, 2004). Predicting CLV involves forecasting future revenue, aiding budgeting, and proactive customer engagement. Initially centered on direct marketing, CLV research expanded across marketing and customer relationship management, propelled by digital technology. Retaining customers is cost-effective, and CLV influences e-commerce decisions, including profitability analysis, cross-selling, and personalized marketing.

Various models have been employed in the prediction of customer lifetime value (CLV), such as the widely used RFM (Recency, Frequency, Monetary) model. Classification via the RFM model assesses recency, frequency, and monetary attributes of customer transactional behaviour which denotes how recently a customer made a purchase, how frequently they engage with the business, and the monetary value of their transactions. Maintaining satisfaction is crucial, with the proposed machine learning models enhancing customer behavior analysis. Customer segmentation is a vital element in the e-commerce data analysis process, aiding in the categorization of customer groups based on similarities. Due to the abundance of transactions, achieving customer group classification through traditional methods can be challenging. Therefore, this project employs customer segmentation using clustering to categorize customers into various clusters, ranging from high to low value, based on their shared characteristics. Also, customer retention is maintained by the firm using marketing techniques that would bring in more profit thereby minimizing the investment risk (Koul, 2021).

Calculating customer lifetime value (CLV) sets off a paradigm-shifting event that gives companies a tactical edge in identifying the financial possibilities that are there in every customer. This means developing targeted marketing campaigns, managing client relationships expertly, and making prudent use of available resources. In the context of e-commerce, where retaining long-term customers is critical, CLV prediction is a vital tool for improving understanding of latent customer value and, in turn, optimizing strategy.

Performance marketing campaigns are a major component of client acquisition efforts in the e-commerce industry, as advertisers compete to attract prospective customers. To determine the profitability of their endeavors, these campaigns mostly rely on CLV projections. Through precise projection of prospective income and customer acquisition expenses, businesses can assess the efficacy of their marketing campaigns. Moreover, precise CLV predictions play a crucial role in budgeting for costly and high-risk marketing ventures such as television advertisements.

Although CLV has garnered substantial attention for its potential to enhance customer-focused decision-making, its precise estimation remains complex due to the intricate interplay of various factors such as customer behavior, usage patterns, and subscription tenures (Sien Chen,

2020). Hence, a gap persists between the potential insights CLV prediction can offer and the practical implementation within the ecommerce industry's services. This dissertation aims to bridge this gap by developing a data-driven approach that effectively predicts CLV and subsequently facilitates the optimization of marketing strategies, thereby addressing the inherent challenges in customer engagement and retention in the realm of ecommerce. This work is also motivated by the observation that determining the number of clusters in advance is a challenging task for many clustering algorithms, primarily due to the intricate distributions of financial data. Additionally, the pervasive correlations among features lead to diverse shapes in data distributions, posing difficulties for numerous clustering algorithms that struggle to adapt to such variations. For example, k-Means can only separate space with hyperspheres, which are not suitable for rotated strip-shaped distributions derived from linear correlations (VanderPlas, 2019). The Elbow Method is also technique used to determine the optimal number of clusters that effectively represent distinct segments of customers however, the Elbow test is sometimes unreliable because the score curves may be smooth and it is hard to find an elbow (T. Li, 2022).

The primary goal of this dissertation is to formulate a comprehensive data-driven approach for precise prediction of Customer Lifetime Value (CLV) within the e-commerce sector, and subsequently leverage these predictions to enhance marketing strategies. More specifically, it seeks to assess different data-driven methodologies and advanced analytics techniques that can effectively predict CLV, including clustering techniques and strategies for integrating data to derive precise CLV estimates. Furthermore, it aims to develop and implement a robust CLV prediction model tailored specifically to the e-commerce industry, and employ the projected CLV values to optimize marketing strategies within this sector. Additionally, the dissertation aims to investigate how these insights can inform decisions related to customer acquisition, retention efforts, personalized marketing campaigns, and resource allocation. The paper aims to utilize Gradient boosting, KNN, MLP, Linear Regression and Decision Trees as predictive tools for CLV estimation, leveraging their capabilities in handling complex datasets and capturing non-linear relationships. Gradient boosting is chosen for its ability to iteratively combine weak learners, leading to high prediction accuracy. KNN excels in identifying patterns by classifying customers based on similarity to others. MLP neural networks are adept at learning complex relationships and capturing intricate patterns in data, contributing to accurate predictions. Linear Regression serves as a fundamental method for understanding the linear relationship between independent variables and CLV. Decision Trees offer interpretability and capture nonlinear relationships, making them suitable for tasks where feature importance is crucial. By evaluating model performance through measures such as accuracy and root mean square error, the study identifies the best-fit model for CLV prediction, thereby enhancing marketing strategy effectiveness and customer retention efforts. By integrating these models with clustering techniques and data integration strategies, the paper seeks to provide a comprehensive and accurate approach to CLV prediction in the ecommerce industry, facilitating informed decision-making in marketing strategies.

Traditional approaches in customer behavior analysis, such as the RFM model, have limitations in accurately forecasting non-contractual customer behavior. However, with the advent of big data and advanced analytics techniques, there has been a shift towards more sophisticated models that can incorporate various factors, including social activities and purchase behavior, to improve the accuracy of Customer Lifetime Value predictions. It is for this reason that we

explore the utilization of Singular Value Decomposition as a method that offers a novel approach to enhance the accuracy of identifying the elbow point, providing a more robust and adaptable solution for determining the ideal number of clusters. This methodological advancement contributes to the overall goal of refining marketing strategies in the e-commerce sector by ensuring that customer segments based on Customer Lifetime Value are accurately identified, allowing for targeted and effective marketing campaigns tailored to distinct customer groups.

1.2 Problem statement

In the realm of E-commerce, where businesses increasingly depend on the digital landscape to attract and retain customers, the formulation of effective marketing strategies has become paramount. As companies strive to optimize their market presence, they grapple with the challenge of targeting customers more efficiently and sustaining their loyalty over time. The key to overcoming this challenge lies in deciphering Customer Lifetime Value (CLV), a multifaceted concept representing the long-term value a company derives from its customers.

Despite the importance of CLV in E-commerce, there exists a critical gap in the application of advanced analytical techniques, particularly SVD (Single Value Decomposition), to enhance the precision and efficacy of marketing strategies. E-commerce businesses are often faced with the arduous task of segmenting their diverse customer base and allocating resources judiciously to optimize CLV.

This dissertation aims to address this gap by investigating the application of SVD for improving marketing strategies in E-commerce with a specific focus on enhancing the accuracy and utility of CLV models. By leveraging SVD, this study seeks to offer innovative insights into market segmentation, resource allocation, and customer targeting, ultimately equipping E-commerce businesses with data-driven tools to enhance customer retention, acquisition, and cross-selling initiatives.

1.3 Research objectives

1.3.1 General objective

The main objective of this dissertation is to develop a comprehensive data-driven methodology for accurately predicting Customer Lifetime Value (CLV) within the E-commerce sector, and to subsequently utilize these predictions to optimize marketing strategies.

1.3.2 Specific objectives

1. To explore data-driven methodologies and advanced analytics techniques that can be employed to predict Customer Lifetime Value (CLV) effectively.
2. To explore the clustering as an unsupervised approach to enhance the accuracy and effectiveness of CLV models for segmenting customers in E-commerce.
3. To propose, design and develop a clustering approach based on Singular Value Decomposition (SVD) to further enhance the understanding and prediction of CLV.
4. To implement and test the developed model to optimize E-commerce marketing strategies, guiding decisions on customer acquisition, retention efforts, personalized campaigns, and resource allocation.

1.4 Justification of the research

The proposed research holds significant justification owing to its potential to address crucial challenges within the E-Commerce sector. As the industry witnesses escalating competition and customer dynamics, the accurate prediction of CLV emerges as a paramount necessity. This study introduces Single Value Decomposition as an approach for determining the most optimal number of clusters to address this need.

By bridging the gap between conventional CLV estimation methods and the unique demands of the E-Commerce landscape, this study aims to empower companies with actionable insights for marketing strategy optimization. Customer Lifetime Value (CLV) plays a critical role in E-commerce, directly impacting revenue and profitability. By leveraging clustering, businesses can gain deeper insights into CLV, thereby improving customer retention, acquisition, and cross-selling efforts. This approach is especially valuable as traditional marketing strategies often fall short in addressing individual customer needs.

Moreover, in the age of big data, businesses are inundated with vast datasets. SVD offers an efficient method to extract meaningful patterns and insights from this data, enabling informed decision-making in marketing strategies. This research contributes to the academic field of Data Science and Analytics by exploring the practical application of clustering while utilizing the SVD approach in the E-commerce context, providing valuable insights for researchers.

Effective marketing strategies are vital for an E-commerce company's success and profitability, and leveraging clustering to optimize these strategies is expected to enhance revenue and long-term sustainability. Furthermore, personalized marketing strategies developed through clustering are more likely to resonate with customers, leading to improved satisfaction, loyalty, and advocacy.

1.5 Scope of the research

This study will explore the application of clustering and utilizing Single Value Decomposition (SVD) as a data-driven method to segment customers in the E-commerce sector. It will investigate the process of using SVD to identify hidden patterns and relationships among customers based on their behaviors, preferences, and interactions with the platform. The research will delve into the concept of Customer Lifetime Value (CLV) and its significance in the E-commerce industry. It will assess how CLV is calculated, its role in revenue generation, and its impact on marketing strategies. The study will also consider the challenges and limitations of traditional CLV measurement approaches such as RFM. The core objective of this research is to evaluate how leveraging clustering can enhance marketing strategies in E-commerce. It will investigate how personalized marketing approaches, derived from clustering insights, can lead to improved customer acquisition, retention, and cross-selling. The study will also assess the potential for revenue growth and long-term sustainability through optimized marketing.

The scope can be broken down into the following aspects:

1. Literature review: Identify the gaps in the literature and provide a foundation for the objectives and methodology.

2. Data collection and Analysis: This research will involve the collection and analysis of extensive data from E-commerce platforms, including customer transaction history, browsing behavior, and purchase patterns. It will explore data preprocessing techniques, feature selection, and the application of clustering algorithms specifically looking into SVD. The study will also consider ethical considerations related to data privacy and security.
3. Selection of machine learning techniques: Applying Clustering for customer lifetime value prediction.
4. Evaluation and testing of the Clustering technique: This research will include a comprehensive evaluation of the performance and impact of the clustering-based marketing strategies. Key performance indicators (KPIs) such as customer conversion rates, CLV improvement, and revenue growth will be used to measure the success of the approach.
5. Application of findings: Demonstrate the practical application of Clustering technique for forecasting and highlight its benefits.

1.6 Limitations of the research

As with any research study, there are several limitations that should be considered when interpreting the findings. Some potential limitations to the research are:

1. The research might face limitations in terms of the availability and quality of historical customer data required for accurate CLV prediction. Incomplete or inconsistent data might affect the performance of predictive models and their ability to generate reliable CLV estimates.
2. The research's focus on the E-Commerce sector might limit applicability to differing sectors and customer dynamics thus the findings and methodologies might have limited generalizability to other sectors or customer segments with different dynamics and characteristics.
3. Assumption of stable customer behavior contrasts with real-world variations, affecting CLV prediction accuracy.
4. Rapid changes in consumer behavior, particularly when dealing with complex behaviors. Capturing these dynamic shifts accurately within the predictive models could be challenging and might lead to less accurate CLV predictions.
5. The SVD approach is a sophisticated data analysis technique, and its implementation may require advanced technical expertise. Constraints related to the complexity of algorithms, computational resources, or access to appropriate software tools could affect the research's practical implementation.

6. While the research focuses on data-driven strategies, the use of customer data raises ethical and privacy concerns. The extent to which personal data is utilized and its implications for data protection regulations and customer trust must be acknowledged.

2. LITERATURE REVIEW

2.1 Introduction

The Literature Review segment of this dissertation delves into the rich body of existing research and knowledge pertinent to the prediction of Customer Lifetime Value (CLV) within the context of the E-Commerce sector. As the landscape evolves, with the shift to E-commerce ventures becoming increasingly prevalent, the need to optimize marketing strategies by accurately estimating CLV becomes paramount. This segment critically examines studies, methodologies, and best practices related to CLV prediction, focusing on their applicability to the unique dynamics of the E-commerce sector. By synthesizing insights from various scholarly works, the Literature Review aims to establish a foundation for Clustering using the Singular Value Decomposition (SVD) method, highlighting both the achievements and gaps in existing literature. The integration of these findings will contribute to a holistic understanding of CLV prediction in the E-commerce industry, paving the way for the subsequent analysis and development of an effective data-driven methodology for marketing strategy optimization.

2.2 Customer Lifetime Value

Customer Lifetime Value (CLV) is a fundamental concept in marketing, representing the total value a customer brings to a business over their lifetime as a customer. It quantifies the long-term impact of individual customers on a company's revenue and profitability. Understanding and maximizing CLV is critical for E-commerce companies aiming to thrive in a competitive landscape as it allows them to make data-driven decisions about resource allocation, customer acquisition, and retention strategies.

Research by (Fader, 2018) emphasizes the importance of CLV as it provides valuable insights into customer retention, acquisition, and cross-selling. CLV models have traditionally been evaluated using the Recency, Frequency, and Monetary (RFM) framework, which categorizes customers based on their transaction recency, frequency, and monetary value. However, the relative significance of these variables varies based on specific product or service attributes and industry characteristics.

Various predictive models can be utilized to estimate Customer Lifetime Value, including simple regression models and complex machine learning algorithms. These models consider factors such as purchase history, frequency of purchases, average order value, and customer engagement metrics to provide an accurate estimation of a customer's potential value. Furthermore, customer segmentation plays a crucial role in optimizing marketing strategies. Previous studies have incorporated customer lifetime value in defining customer segments (Radit Rahmadhan, 2022). For example, (Lee, Lee, Chang, & Sano, 2021) utilized the K-Means clustering model and customer lifetime value considering product preferences to predict behavior in buying products.

One approach to customer segmentation is the use of Single Value Decomposition method for clustering. SVD is a technique that enables businesses to group customers based on similar patterns and behaviors.

This approach utilizes Singular Value Decomposition, a matrix factorization technique, to analyze customer data and identify underlying patterns. SVD in customer segmentation has

gained popularity due to its ability to handle high-dimensional data and capture non-linear relationships between variables effectively. Using SVD for customer segmentation can provide valuable insights into different customer segments and their characteristics. It enables businesses to tailor their marketing strategies and offerings to specific customer groups, ultimately leading to higher customer satisfaction and increased customer lifetime value.

The concept of customer lifetime value has a long-standing history and is widely recognized as an essential metric in marketing. Driven in part by the interest in moving from transaction-oriented/product-centric marketing strategies to relationship-oriented/customer-centric marketing strategies, the 1990s saw the emergence of customer lifetime value as a key metric in the field of marketing.

Since then, numerous studies have focused on enhancing the accuracy and applicability of customer lifetime value models.

These studies have explored various approaches to predicting customer lifetime value, including regression models and machine learning algorithms such as Bayesian Inferences, Moving Averages, Regressions, and Pareto/NBD models. Moreover, segmentation techniques have been applied to customer lifetime value analysis.

For instance, researchers believe that CLV works as the basis for firms to segment their customers and allocate marketing resources. Segmenting customers based on their lifetime value and designing corresponding marketing schemes can bring more profits to retailers than segmenting customers based on sociodemographic characteristics. In retail banking, market segmentation can also estimate customer future value as part of Customer Lifetime Value (Ridloah, 2016). Furthermore, previous studies have incorporated customer lifetime value in defining customer segments (Radit Rahmadhan, 2022).

In addition, (Hyunseok Hwang, 2004) investigated the moderation effect of customer lifetime value in segmenting customers. Their study found that customer lifetime value can serve as a moderating factor in the relationship between customer segmentation and marketing effectiveness.

Overall, previous studies have shown that leveraging customer lifetime value in marketing strategy can lead to improved customer segmentation, allocation of marketing resources, and ultimately higher profits for businesses. The literature review draws attention to the significance of customer lifetime value as a metric in marketing and its role in shifting from product-centric to a customer-centric approach. However, it is worth noting that customer segmentation based on lifetime value alone may not capture all relevant factors influencing customer behavior and preferences. Future research could explore the integration of customer lifetime value with other segmentation variables, such as demographic, geographic, and behavioral factors, to create more comprehensive customer segments. In summary, previous studies have demonstrated the importance of incorporating customer lifetime value in marketing strategy and segmentation.

For instance, researchers believe that CLV works as the basis for firms to segment their customers and allocate marketing resources. Segmenting customers based on their lifetime value and designing corresponding marketing schemes can bring more profits to retailers than segmenting customers based on sociodemographic characteristics alone. Furthermore, Mazzoni et al. highlight the importance of customer lifetime value in the banking industry.

They conclude that customer lifetime value includes calculating the past and present value of customers and predicting the future value of customers (Ridloah, 2016). Furthermore, (Oliver Dzobo, 2014) proposed a segmentation model using hierarchical clustering to cluster electricity customers based on similar cost characteristics. Additionally, (Ozgen, 2017) introduced the concept of "customer equity," which is the sum of the lifetime values of a firm's customers and emphasized its importance in understanding and managing customer relationships.

In the realm of retail banking, it is evident that customer segmentation based on lifetime value plays a significant role in improving marketing strategies. It allows banks to identify and allocate resources to high-value customers, develop tailored marketing campaigns, and provide personalized advisory services. Integrating customer lifetime value into marketing strategy and segmentation has been widely recognized as a crucial factor for businesses across various industries, including e-commerce and retail banking. In recent years, there has been a growing interest in leveraging customer lifetime value for segmentation and marketing strategy development in the e-commerce industry. This is evident in the research conducted (Su-yeon Kim, 2006) who explored the use of clustering to improve marketing strategies in e-commerce by incorporating customer lifetime value. Furthermore, previous studies have shown that segmenting customers based on their lifetime value rather than sociodemographic characteristics leads to higher profits for retailers and allows for more targeted marketing efforts. Moreover, (Mazzoni, 2005) applied a multidimensional segmentation approach in the e-commerce industry, considering consumer lifestyles, motivational factors, and product attributes to identify distinct customer segments with varying lifetime value. These studies collectively demonstrate that leveraging customer lifetime value in e-commerce can provide valuable insights for effective marketing segmentation and strategy development. Customer segmentation based on customer lifetime value has been widely recognized as a critical factor in improving marketing strategies and optimizing resource allocation in various industries, including e-commerce and retail banking. By utilizing customer lifetime value as a segmentation criterion, businesses can gain a deeper understanding of their customers' preferences, behaviors, and purchasing patterns. This understanding enables businesses to effectively target and personalize their marketing efforts, leading to increased customer engagement, loyalty, and ultimately, higher profitability. By incorporating customer lifetime value into their marketing strategy, businesses can identify and prioritize high-value customers, allocate resources effectively, and tailor their marketing campaigns to meet the specific needs and preferences of each customer segment.

These strategies not only increase customer satisfaction and retention but also maximize the ROI of marketing initiatives. Furthermore, customer lifetime value can also be utilized in the context of retail banking (Ridloah, 2016). As the retail banking industry undergoes significant changes and embraces data-driven approaches, understanding customer segmentation has become increasingly important. One way to leverage customer lifetime value in the retail banking industry is by using it as a strategy to forecast which customers will be profitable. By calculating the customer’s lifetime value, banks can identify high-value customers and provide personalized advisory services to enhance their experience and increase profitability. Moreover, the emergence of big data and social media has further enhanced the potential for leveraging customer lifetime value in marketing strategies. These platforms enable businesses to gather and analyze vast amounts of customer data, allowing for more accurate identification and segmentation of customers based on their lifetime value. As an illustration, (Shi, 2019)

conducted a comprehensive review of clustering methods tailored for spatiotemporal data, spanning various domains such as social media, human mobility, and transportation analysis.

Segmenting customers based on their lifetime value is crucial for businesses in understanding customer behavior, predicting future value, and allocating marketing resources effectively. By segmenting customers based on their lifetime value, businesses can identify high-value customers who are more likely to make repeat purchases and have a higher potential for upselling and cross-selling. Furthermore, clustering allows businesses to identify potential value customers who may not have reached their full purchasing potential yet but exhibit promising behaviors. These customers can be targeted with personalized marketing campaigns to increase their engagement and loyalty. Previous studies have shown that segmenting customers based on their lifetime value is more effective than segmenting based on sociodemographic characteristics alone.

Another study by (Cvijović Jelenaa, 2017) proposed a customer lifetime value model for classifying individual customers into groups and developing customized products for each of them.

These findings suggest that leveraging customer lifetime value segmentation, specifically through clustering, can significantly improve marketing strategies in the e-commerce industry. By utilizing clustering, businesses can gain deeper insights into customer behavior patterns and preferences. This allows for the development of targeted marketing campaigns and personalized product offerings, leading to increased customer engagement, loyalty, and ultimately, higher profits for retailers. In addition to segmenting customers based on their lifetime value, businesses can also incorporate other factors such as customer types, violation of contract, and purchase frequency to further refine their marketing strategies. Furthermore, the importance of customer segmentation and calculating customer lifetime value is particularly relevant in the retail banking industry. The use of clustering in conjunction with customer lifetime value offers promising potential in improving marketing strategies and overall business profitability. In summary, leveraging clustering for improved marketing strategy in e-commerce, specifically through customer lifetime value segmentation, has been shown to be effective in various industries such as e-commerce and retail banking. These findings highlight the potential benefits of using clustering, through SVD for improved marketing strategy in e-commerce, specifically through customer lifetime value segmentation. However, it is important for future research to further explore the application of clustering in different industries and evaluate its effectiveness in comparison to other clustering algorithms and segmentation techniques.

In the realm of Customer Lifetime Value (CLV) prediction, leveraging diverse machine learning models has garnered significant attention. Notably, Linear Regression, Gradient boosting, KNN, MLP, and decision trees have emerged as prominent choices due to their adeptness in handling intricate datasets and capturing complex relationships between predictor variables and CLV. A study by (Neha Chaudhuri, 2021) investigated the predictive power of various machine learning techniques, including Decision Trees, Random Forest, Support Vector Machines, and Artificial Neural Networks, alongside a deep learning method, to forecast online purchases by retail customers. They utilized two distinct sets of variables, platform engagement, and customer characteristics, as key predictors. The study revealed that the deep learning technique exhibited superior predictive performance compared to traditional

machine learning methods when applied to the same dataset. These findings not only contribute to expanding academic knowledge regarding purchase prediction for online e-commerce platforms but also provide valuable insights for platform designers to enhance platform engagement strategies. Gradient boosting, recognized for its iterative approach in combining weak learners, offers high prediction accuracy, making it particularly well-suited for CLV prediction tasks. (Singh, 2018)conducted Response Regularization to devise a mathematical framework comprising multi-layer models for estimating customer lifetime value. This framework was built upon a robust theoretical taxonomy and assumptions rooted in customer characteristics. A study by (Lami, 2022) transcends the conventional predict-then-optimize paradigm by integrating prediction and prescription processes, thus enhancing the robustness and accuracy of decision-making. By jointly predicting uncertain quantities and prescribing optimal decisions, the framework offers a holistic approach to prescriptive analytics, with wide applicability across diverse domains. Real-world case studies in healthcare, industrial operations, and pandemic management illustrate the transformative potential of predictive and prescriptive analytics methodologies in addressing complex challenges and optimizing resource allocation.

In summary, the dissertation underscores the predictive advantage of advanced machine learning techniques, particularly gradient boosting, in accurately forecasting Customer Lifetime Value (CLV) and optimizing marketing strategies within the ecommerce industry. Through innovative methodologies and empirical validations, the research contributes to the advancement of predictive and prescriptive analytics, offering actionable insights for informed decision-making and value creation in various operational contexts. Similarly, KNN stands out in identifying patterns by classifying customers based on their similarity to others, thereby effectively recognizing behavioral trends (Mazzoni, 2005). Meanwhile, MLP neural networks demonstrate proficiency in learning intricate relationships within data, contributing to accurate CLV predictions. Additionally, decision trees, with their interpretability and ability to capture nonlinear relationships, are deemed suitable for CLV prediction tasks where feature importance and interpretability are paramount considerations. The literature underscores the significance of incorporating these models alongside clustering techniques and data integration strategies to provide a comprehensive and accurate approach to predicting CLV in the ecommerce industry, thereby facilitating informed decision-making in marketing strategies.

2.3 Managing failed customer interactions (customer churn)

Customer churn occurs when customers who previously purchased goods or services from a business stop doing so and switch to a competitor (Wu, 2022). In e-commerce, customer churn occurs in a non-contractual relationship where it can be challenging for businesses to detect the termination of the relationship in advance (Shao, 2016). To reduce customer churn, e-commerce companies must accurately predict which high-value customer groups are likely to churn and study the purchasing habits of customers who have not churned to retain them. The key to predicting e-commerce customer churn is to combine customer data over a period and analyze customer purchase behavior to establish e-commerce customer churn prediction models (Zhang, 2015). These models can be used to provide e-commerce customer churn retention measures and identify high-value non-churn e-commerce customers for retention. According to (Shao, 2016) research, retaining existing customers is less costly than acquiring new ones in e-commerce. Thus, identifying the reasons for customer loss is crucial. (Lu, 2018)

emphasized the need to analyze customer loss, predict potential customer loss, and take appropriate measures to retain customers in the e-commerce sector. E-commerce companies typically use various methods and technologies, such as data mining, to establish and study customer churn prediction models based on customer basic characteristics and transaction behavior data (Xiahou, 2022). Data mining technologies, including customer segmentation, customer churn prediction, and fraud analysis, are widely used in e-commerce customer relationship management.

2.4 Customer segmentation

Customer segmentation involves identifying the value of customer relationships, which is crucial for efficient targeted marketing (Yue Li, 2021). According to the Pareto principle, 80% of a company's profits come from 20% of its customers, while the bottom 30% of non-profit customers account for a 50% loss of profits (Kanyinda, 2021). Therefore, to perform effective customer segmentation, it's important to first recognize and leverage customer value. By focusing on customer value and allocating resources towards targeted marketing, companies can improve their core competitiveness.

As e-commerce continues to develop rapidly, business activities carried out online are becoming more real-time and interactive, transforming the traditional product-centric model into a customer-centric approach (Alshamsi, 2022). Customers have become a crucial resource for competitive advantage and profitability, according to (Agrawal, 2018). However, the churn rate of e-commerce customers is high, and companies must establish long-term alliances and continuous customer relationships to retain them. The e-commerce customer base is extensive and complex, and their value varies. (Saghir, 2019) raised the question of accurately identifying high-value customers, predicting churn, and retaining them in advance, which has become a hot topic in the e-commerce field. (Wu S. Y., 2021) found that evaluating customer value helps companies identify valuable customers among many consumers and implement different customer management strategies based on their values, maximizing the impact of limited resources. Purchasing value is the main explicit value of e-commerce customers and is selected as the primary measure of customer value. The amount of purchases is directly related to the sales volume of enterprise products or services and is a guarantee for achieving enterprise profit targets and customer value. Historical transaction data is mainly used to identify customer value and form the basis for customer segmentation.

2.4.1 RFM Model

The RFM model, introduced by Arthur Hughes in 1994, serves as a foundational concept for calculating customer lifetime value. The model incorporates three key parameters: Recency (R), Frequency (F), and Monetary (M). The Recency parameter reflects the time since the customer's last purchase, with smaller values indicating recent activity and higher values signifying longer intervals since the last purchase. A smaller R implies a customer's continued interest and potential responsiveness to marketing efforts. Conversely, a larger R suggests decreased competitiveness of the enterprise's offerings.

Frequency (F) represents how often a customer purchases within a given period. Higher F values indicate greater customer loyalty and overall customer value. The combination of purchase frequency and recency helps predict future customer behavior. Customers with high frequency but distant recency may have been valuable in the past but are likely to be lost.

Monetary (M) signifies the total amount spent by customers on enterprise products or services within a specified timeframe. Higher monetary values are associated with greater customer loyalty and value. To mitigate collinearity issues between monetary and recency/frequency, the average consumption amount is often used instead of the total amount. Overall, the RFM model provides a comprehensive framework for understanding and predicting customer behavior based on these three crucial dimensions.

2.4.2 K-means and Ward's method

Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method (P. P. Pramono, 2019) focuses on enhancing customer segmentation by aligning it with customer lifetime value (CLV). The primary objective is to identify customer segments sharing similar lifetime values, enabling companies to tailor targeted strategies for each segment. In the context of the competitive beauty industry in Indonesia, effective Customer Relationship Management (CRM) is crucial for companies to thrive. This study addresses this challenge by employing a clustering method to identify customer segments with similar lifetime value, enabling targeted and strategic approaches. The two-stage clustering method involves Ward's method for initial cluster number selection and K-Means for segmentation analysis. The study compares the LRFM (Length, Recency, Frequency, Monetary) model with an extended version, LRFM-Average Item (AI), using validity indices. The results indicate that introducing the Average Item variable does not significantly enhance clustering outcomes. The ranking process, based on Customer Lifetime Value (CLV) scores, incorporates weighted LRFM model variables, determined through the Fuzzy AHP method. The study provides valuable insights into customer characteristics, aiding in the formulation of effective sales and marketing strategies.

2.4.3 Fuzzy c-means cluster and the RFM model

The study conducted by (Prasetyo, 2020) underscores the significant growth and popularity of e-commerce as a preferred platform for online shopping, driven by its accessibility to internet users and a diverse range of offered products. With intense competition in the business landscape, effective marketing strategies are crucial for e-commerce companies to attract, retain, and engage customers. This study focuses on customer segmentation in e-commerce, employing the Fuzzy C-Means clustering and the RFM method. The clustering process, executed six times with varying cluster numbers, revealed that the optimal number of clusters, identified through the Xie-Beni validity index, is four clusters. These clusters represent distinct customer segments based on recency, frequency, and monetary value. Notably, Segment 4 emerges as the most valuable with highly loyal customers. The identified segments provide valuable insights for tailoring targeted marketing strategies to enhance customer engagement and satisfaction.

2.4.4 Tree clustering, RFM model

The research conducted by (Ming, 2017) addresses the limitation of traditional customer segmentation methods by introducing the Recency Frequency Monetary Purchase Tree (RFMPT) approach, which incorporates the value of goods in transaction data. The methodology involves building an RFM purchase tree based on the product categories, and subsequently proposing a fast clustering algorithm termed Based Recency Frequency Monetary Purchase Tree Clustering (BRFMPTC). The BRFMPTC algorithm utilizes a CoverTree (CT)

index structure, enabling the rapid identification of k densest purchase trees as cluster centers and efficient assignment of other objects to their nearest cluster center. Experimental results highlight the enhanced performance of this method, particularly when incorporating distance weighting, as compared to conventional clustering algorithms. This innovative approach offers a promising solution for more nuanced and effective customer segmentation.

2.4.5 RFM model, K-Means and Fuzzy C-Means algorithms

The study conducted by (Christy, 2021) focuses on the efficient segmentation of enterprise customers based on RFM (Recency, Frequency, and Monetary) values, categorizing them into groups with similar behavior. Transactional data is analyzed over a specific period to gain insights into customer needs and identify potential customers. Customer segmentation not only provides a deeper understanding but also contributes to increased revenue. The study emphasizes the importance of customer retention over acquiring new ones, suggesting tailored marketing strategies for each segment. The research conducts RFM analysis and employs traditional K-means and Fuzzy C-Means algorithms for clustering. A novel approach for choosing initial centroids in K-Means is introduced. The results are compared based on iterations, cluster compactness, and execution time, providing valuable insights into the effectiveness of different segmentation methodologies.

2.4.6 K-means clustering method, K-medoids method and Fuzzy RFM model

Customer segmentation can be effectively conducted using the K-means clustering method, K-medoids method, and the Fuzzy RFM model. Among these, the K-means clustering method partitions a dataset into K distinct clusters based on the least squared Euclidean distance to the mean. In the context of customer segmentation, K-means can identify clusters of customers exhibiting similar behaviors, such as recency, frequency, and monetary value. On the other hand, the K-medoids method, a variant of K-means, is more robust to outliers as it selects actual data points as representatives of clusters. This method can be advantageous when dealing with datasets containing outliers or when using non-Euclidean distance metrics. Additionally, the Fuzzy RFM model integrates fuzzy logic into the traditional RFM model, allowing for the representation of uncertainty in customer segmentation. This model assigns degrees of membership to each cluster, providing a more nuanced understanding of customer behavior, especially when ambiguity is present. The overall process involves data preparation, feature selection, parameter tuning, model application, interpretation, and strategy formulation based on the identified customer segments. These methods collectively empower businesses to tailor marketing strategies, enhance customer engagement, and improve overall satisfaction by catering to the distinct needs and behaviors of different customer segments. The study conducted by (Muningsih, 2018) explores the effectiveness of two clustering methods, namely the widely used K-Means method and the efficient K-Medoids method, in grouping customers based on their characteristics. The objective is to categorize customers into three clusters: very potential (loyal) customers, potential customers, and non-potential customers. Utilizing online sales transaction data and employing the Fuzzy RFM (Recency, Frequency, and Monetary) model, the study compared the performance of the two clustering methods. The results indicated that the K-Means method outperformed the K-Medoids method, achieving an accuracy value of 90.47%. The clusters formed consisted of 16 members in Cluster 1, 11 members in Cluster 2, and 15 members in Cluster 3. This research contributed valuable insights

into customer segmentation techniques, providing a foundation for understanding customer behavior in the context of online sales transactions.

2.4.7 Determining Cluster size using the elbow method

A study conducted by (Marisa, 2023) aims to identify potential customer clusters using the K-Means clustering approach, leveraging the elbow method to determine the optimal number of clusters. Analyzing a dataset of 100 customers from a minimarket based on gender, age, and purchase retention, the initial clusters are set to 5. Through K-Means calculation, the SSE value indicates the lowest value, and the elbow angle graph highlights cluster 4. Therefore, the optimal number of clusters is determined as four (4) for further analysis. The results show that high-potential customers, particularly females with high purchase retention, dominate in three (3) clusters, mainly in the age range above 35 years. Customers with lower potential are dispersed across clusters, showing varied gender and age but no dominant trends. This insight can guide management in formulating targeted promotion strategies.

2.4.8 Determining cluster size using Singular Value Decomposition (SVD)

One clustering technique that has gained prominence is Singular Value Decomposition clustering. Singular Value Decomposition is a matrix factorization method that decomposes a matrix into its constituent parts, providing valuable insights into the underlying structure and relationships within the data. The use of Singular Value Decomposition in recommender systems has been extensively explored in the literature emphasizing its use to streamline data representation and enable predictions through linear regression.

SVD is a powerful mathematical technique that has demonstrated success in various data analysis applications. In E-commerce, SVD offers the potential to segment customers with a higher degree of precision and personalization. By leveraging this technique, businesses can uncover hidden patterns and relationships within their data, which can guide more targeted marketing efforts.

In the contemporary ecommerce sector, retaining customer loyalty and sustaining their attention poses significant challenges, necessitating the periodic reinforcement of marketing strategies. Research by (K. Bhade, 2018) introduces a systematic approach to target customers effectively, optimizing profits for organizations. Initial steps involve analyzing sales data derived from purchase history to identify parameters with maximum correlation. Machine learning algorithms, specifically K-Means clustering for customer segmentation and Singular Value Decomposition for tailored recommendations, guide the allocation of resources to maximize profitability. The paper also addresses challenges in recommender systems, such as the cold start problem and sparsity, proposing strategies to overcome these drawbacks.

Pooling efforts to maintain customer loyalty and attention due to fierce competition, organizations must continuously reinforce their marketing strategies. A paper by (Yogesh Jadhav, 2020) presents a systematic approach to target customers effectively and maximize profits through technological advancements. The initial step involves analyzing sales data, particularly focusing on the parameters of recency and frequency in customer purchases for clustering. Utilizing machine learning algorithms, resources can be directed toward profitable customers. The paper also addresses drawbacks in recommender systems, such as the cold start problem and sparsity, proposing solutions. Customer segmentation employs K-Means

clustering, while Singular Value Decomposition (SVD) is applied for personalized recommendations. SVD proves efficient in matrix factorization, simplifying data storage and processing by removing redundant features from the utility matrix. The recommender system model's equation facilitates the calculation of predicted scores for user-product interactions and subsequent recommendations.

Sustainable tourism, particularly in the form of eco-friendly hotels, is gaining prominence globally. A paper by (Mehrbakhsh Nilashi, 2021) delves into the intersection of eco-friendly accommodations, Electronic Word-of-Mouth (e-WOM), and advanced data analysis techniques. Focusing on online reviews of eco-friendly hotels, the study employs multi-criteria decision-making and machine learning methods for preference learning. A novel approach incorporates the Expectation-Maximization (EM) algorithm for unsupervised learning to cluster travellers' reviews. Higher-Order Singular-Value Decomposition and a similarity measure are utilized to identify customers with similar preferences. The Adaptive Neuro-Fuzzy Inference System (ANFIS) is employed as a supervised learning technique to predict travellers' preferences for eco-friendly hotels. Criteria importance is determined through the entropy-weight approach in each segment. The effectiveness of this hybrid approach is validated through experiments using data from eco-friendly hotels in the Czech Republic on the TripAdvisor platform, showcasing its efficacy in customer segmentation and preference prediction for sustainable accommodation.

Previous research highlights the effectiveness of SVD in uncovering complex customer behaviors. This approach goes beyond traditional RFM categorizations to create customer segments based on intricate patterns in behavior, purchase history, and preferences. SVD allows E-commerce companies to optimize their marketing strategies by aligning them with these detailed customer segments.

The connection between SVD and customer lifetime value lies in its ability to identify distinct customer segments based on their predicted future value. By applying SVD to customer data, businesses can group customers with similar characteristics and behaviors together, allowing for more targeted marketing strategies. These strategies can be aimed at increasing customer retention, cross-selling and upselling, and maximizing overall customer lifetime value. Moreover, SVD can help businesses understand the factors that contribute to a higher customer lifetime value. For example, clustering can uncover patterns and associations between different customer attributes such as purchase history, demographic information, and browsing behavior. By identifying these patterns, businesses can determine which factors have a significant impact on customer lifetime value and tailor their marketing efforts accordingly.

Utilizing SVD clustering can help businesses identify these segments and implement targeted marketing strategies to convert them into high-value customers, thereby increasing customer lifetime value and ultimately driving profitability.

2.5 Marketing strategies in e-commerce

E-commerce marketing strategies aim to connect with customers, understand their needs, and deliver personalized experiences. Traditional strategies often rely on broad categorizations, such as demographic or geographic segmentation, which may not effectively address individual customer behaviors.

Literature reveals that a more granular approach is required to address complex customer behaviors. Customers now exhibit intricate behaviors influenced by multiple factors, including past purchase history, website interactions, and responses to marketing campaigns (Lemmens, 2007). Therefore, it is imperative for E-commerce businesses to adopt more sophisticated and personalized marketing approaches that resonate with customers.

Another case study by (Sommella, 2023) focused on leveraging clustering for customer segmentation in the fashion e-commerce industry. They applied clustering to customer data, considering factors such as purchase history, browsing behavior, and demographic information. The results showed that clustering successfully identified distinct customer segments with different preferences and shopping behaviors. By tailoring marketing strategies to these segments, the e-commerce businesses were able to improve customer satisfaction and increase the effectiveness of their marketing efforts.

In the context of advancing computer technology and the era of artificial intelligence, a study conducted by (Benfang Yang, 2022) addresses the crucial task of analyzing user demand bias to optimize the operations of e-commerce platforms. Leveraging CS domain signaling data, PS domain IP packet data, and customer CRM data provided by operators, the research delves into various dimensions of operator user portraits. The operator user portrait platform is segmented into individual subunits, and data mining techniques are applied to each subunit. The system adeptly processes and mines multidimensional data related to operators' users, forming comprehensive user portraits through data aggregation. The paper concludes by demonstrating the application value of this research in enhancing precision marketing and personalized services offered by operators, particularly by analyzing user data encompassing mobile phone usage and consumption behavior.

2.6 Data-driven decision-making

E-commerce businesses are inundated with vast datasets. Effectively harnessing this data is crucial for informed decision-making in marketing strategy. Utilizing clustering techniques facilitates the extraction of meaningful patterns and insights from intricate datasets, establishing a foundation grounded in data-driven approaches for optimizing strategies.

The exploration of targeted marketing strategies, particularly through market segmentation, has garnered significant attention in both industry and academia. In the realm of retail, traditional segmentation models often fall short in providing deep market insights and identifying smaller segments. A study by (Yoseph, 2020) addresses these limitations by harnessing the capabilities of the Hadoop distributed file system for processing extensive datasets. Employing modified best-fit regression, specifically Expectation-Maximization (EM) and K-Means++ clustering algorithms, the research conducts three distinct market segmentation experiments, assessing their performance through cluster quality evaluation. The outcomes contribute to data-driven decision-making in two key aspects that include revealing insights into customer purchase behavior within each Customer Lifetime Value (CLTV) segment and evaluating the clustering algorithm's effectiveness in generating accurate market segments. Notably, the analysis highlights a brief average customer lifetime of two years and a high churn rate of 52%. Subsequently, a targeted marketing strategy based on these findings is implemented for departmental store sales, resulting in a significant increase in the sales growth rate from 5% to

9%. This underscores the role of data-driven insights in shaping and optimizing marketing strategies for enhanced business outcomes.

2.7 Challenges and limitations

It is imperative to address the potential hurdles and constraints associated with applying clustering techniques within the context of e-commerce and Customer Lifetime Value (CLV) analysis. These challenges may encompass issues related to data privacy and security, as e-commerce platforms often deal with sensitive customer information. Furthermore, the limitations of clustering, such as the curse of dimensionality and lack of the comprehensive capability to model specific temporal patterns commonly observed in practical scenarios, such as the periodic purchasing behavior of customers may impact the practicality of implementing these techniques. Additionally, the need for substantial computational resources and expertise may present challenges for businesses with limited technical capabilities. Acknowledging these challenges and limitations is crucial for developing a comprehensive understanding of the proposed research and providing a foundation for addressing these issues in the subsequent research methodology and discussion sections.

2.8 Conclusion

This literature review provides a foundation for the research on leveraging SVD for improved marketing strategy in E-commerce. It highlights the significance of CLV, the evolving landscape of marketing strategies, the potential of SVD, and the importance of data-driven decision-making. This study aims to contribute to the existing body of knowledge by exploring how SVD can enhance marketing strategies in the context of complex customer behaviors in the E-commerce sector.

In the following chapters, we will delve into the research methodology, data collection, analysis, and the practical implementation of SVD clustering for marketing strategy optimization in E-commerce.

3. METHODOLOGY

3.1 Introduction

The e-commerce market has become fiercely competitive due to recent industry growth. Therefore, retaining customers is crucial for the success of companies in this field. This methodology chapter aims to explain the plan and methods for forecasting Customer Lifetime Value in e-commerce. Customer Lifetime Value is a key metric for e-commerce companies, helping them understand the long-term value of their customers and make decisions about marketing, resources, and customer groups. We will use a mix of machine learning models to achieve this. This section outlines our approach for predicting Customer Lifetime Value (CLV) in e-commerce to improve marketing strategies. Our method includes steps like collecting and preparing data, developing models using statistical and machine learning methods, and carefully evaluating and validating these models.

Over the past thirty years, there has been an evolution of diverse models and methodologies aimed at computing Customer Lifetime Value (CLV). These models are tailored to suit various types of businesses, managerial viewpoints, and business-to-consumer (B2C) interactions. A common classification divides CLV models into three categories: contractual, semi-contractual, and non-contractual relationships between customers and companies (Estrella-Ramón, 2014). In the context of online shopping, it is crucial for CLV models to acknowledge the non-contractual aspect of the business-to-consumer (B2C) relationship. This includes factors such as the absence of formal customer membership, the inherent sharing dynamics, the ongoing nature of purchases, and the variability in spending patterns (Jasek, 2018). Notably, the existing literature, as indexed by the Web of Science, lacks comprehensive comparative analyses of a wide array of CLV models, focusing on their predictive capabilities.

Some prior research has explored the performance of CLV models in specific business contexts. For instance, (Nie, 2022) identified the superior performance of creation of a unified customer record, categorization of customers into ranked groups, interpolation of missing parameters, and concludes with the calculation and validation of individual CLV values. (Takhun Kim, 2022) conducted a study that validates the performance of stochastic customer base models in noncontract settings where the time at which a customer becomes dormant is not observable. The study employs four buy-'til-you-die (BTYD) models: a) the original Pareto/NBD model, b) the Pareto/GGG model, c) the BG/CNBD-k model, and d) the MBG/CNBD-k model. Across one-month to twelve-month forecasting horizons, these BTYD models effectively classify active customers, yielding receiver operating characteristic curve areas ranging from 0.82 to 0.86. The findings underscore the BTYD framework's utility in customer base analysis, serving as an immediate heuristic approach to complement existing customer relationship management tools. The integration of probabilistic models is evident in the work of (Jan Valendin, 2022), who applied the LSTM approach as a forecasting tool for sequential data and to show its flexibility in incorporating covariates to datasets from diverse environments.

This paper's primary focus is on harnessing the power of Singular Value Decomposition (SVD) to enhance marketing strategies in the e-commerce sector, with a unique emphasis on the Customer Lifetime Value (CLV) framework in the e-commerce environment, a topic of current relevance on both local and global scales. Companies in this category typically maintain

extensive databases containing customer and transaction data, which they utilize for various financial and marketing purposes. The online retail market has witnessed continuous growth, leading to increased competitive pressure in the realm of online shopping. Given these dynamics, our research remains highly current and pertinent to the present business environment.

3.2 Research Design

The study employs the CRoss Industry Standard Process for Data Mining (CRISP-DM) model, a well-established methodology for data mining projects and a foundational model for a data science process. It consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. These phases involve defining project objectives, acquiring and exploring data, preparing and transforming data, selecting and applying appropriate modelling techniques, evaluating model performance, and integrating models into business processes. Ultimately, adherence to the CRISP-DM methodology ensures effective data mining, accurate model construction, and informed business decisions, contributing to overall success.

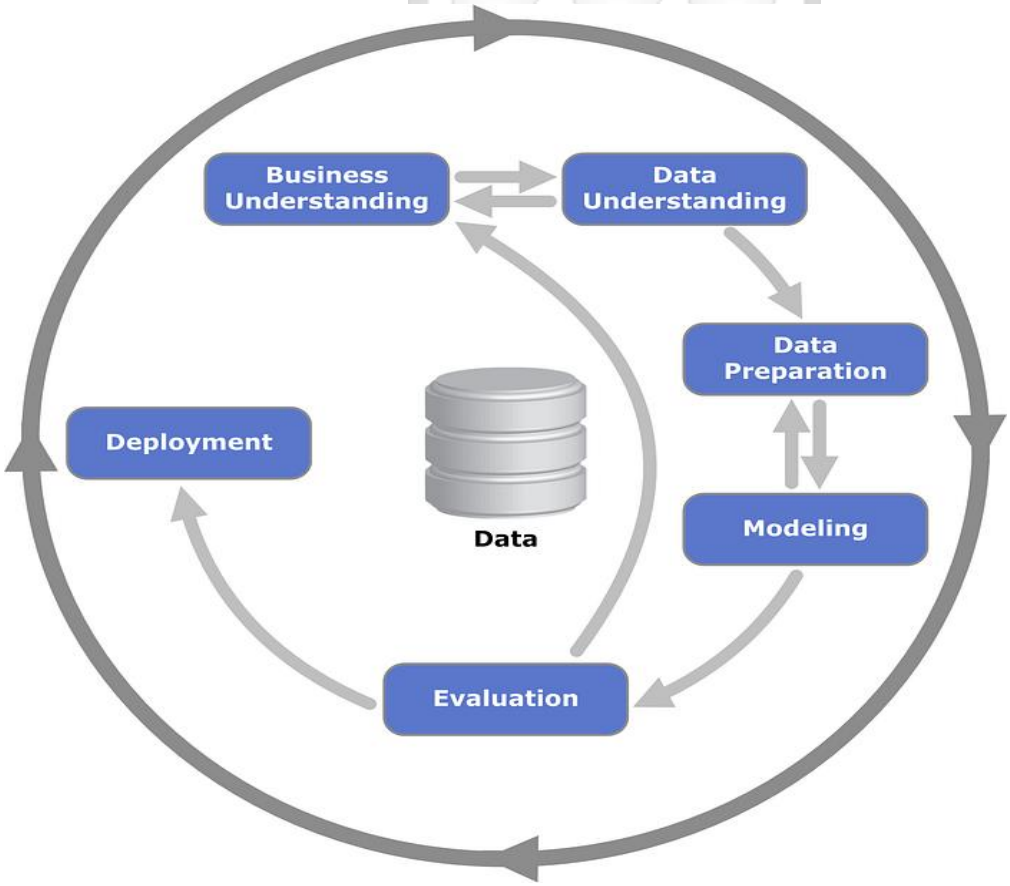


Figure 1 CRISP-DM Model by Hotz, 2018

3.3 Data collection

In this study, data was obtained from a Kenyan Franchise selling baby products from 2020-2023 point of sale data in excel format. This data provides the necessary information on customer behavior characteristics and value. This data was used to construct a customer value

model, which will help in identifying the high-value customers of the enterprise and predicting the customer lifetime value. This data collection process was crucial as it forms the foundation of the entire analysis and prediction process. Below is the breakdown of the dataset used:

Table 1 Data features description

Column Name	Description
Order ID	A unique identifier is assigned to each customer order for tracking and reference purposes.
Sale ID	An identifier for the specific sale transaction, linking it to other relevant records and details.
Date	The date on which the sale transaction or order was processed or completed.
Order	Details related to the customer's order linked to a Salesperson for commission processing.
Transaction type	Specifies the nature of the transaction, such as a product purchase or shipping.
Sale type	Indicates the type of sale, which might include purchase or return.
Sales channel	Specifies the channel through which the sale was conducted, whether in-store, online, or through a specific platform – e.g. Mobile money platform such as Mpesa, card payments, cash etc.
Product	Identifies the specific product(s)
Product type	Refers to the classification system for organizing products based on common attributes.
Net quantity	Represents the quantity of products sold, excluding any returns or discount.
Gross sales	The total revenue generated from sales before accounting for discounts, returns, or other adjustment.
Discounts	The total value or percentage of discounts applied to the sales transactions.
Returns	Represents any products that customers returned, including details such as quantity and value.
Net sales	The total revenue generated from sales after accounting for discounts and returns.
Shipping	The cost associated with shipping the products to the customer, if applicable.
Taxes	Total taxes incurred as part of the sales transaction. – they're zero since the sales are tax inclusive (VAT).
Total sales	The overall sum of net sales, shipping, taxes, and any other relevant charges, providing the total revenue for the sale.They are the same as the Gross sales, Net sales and Total sales.

3.4 Data Preprocessing

In this stage, the raw data is cleaned, transformed, and prepared for analysis. This involves tasks such as handling missing or inconsistent data, performing feature engineering to create relevant variables, and integrating data from various sources. The goal is to ensure the quality

and suitability of the data for modeling purposes, laying the groundwork for accurate and effective analysis in subsequent stages of the data mining process.

3.4.1 Cleaning the data

Data cleaning was conducted meticulously using Python's pandas module, involving a thorough examination of the dataset. Our analysis focused on identifying missing values and categorizing them into three distinct types: Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR). Depending on the category, appropriate methods were applied to address the missing values. For MCAR, imputation with the mode was performed, while multiple imputation was utilized for MAR scenarios. MNAR instances were tackled using domain-specific imputation techniques. Additionally, redundant columns and those with zero or negative data points were dropped. Class imbalances and outliers were also identified and treated accordingly, ensuring the robustness and reliability of the dataset for further analysis.

3.4.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a vital process in understanding dataset structures, attributes, and underlying patterns. It involves univariate analysis, focusing on individual variables to assess their distributions and summary statistics like mean, median, and standard deviation for anomaly detection. Bivariate analysis explores relationships between pairs of variables, revealing correlations and dependencies through techniques like scatter plots and correlation matrices. Multivariate analysis extends this exploration to multiple variables simultaneously, providing insights into complex interactions and relationships using methods such as principal component analysis and cluster analysis. Throughout the EDA, the goal is to identify data quality issues like missing values and outliers, which could introduce noise and affect analysis outcomes, enabling informed decisions in subsequent data analysis steps.

3.5 Data Transformation

This is a crucial stage in preparing the dataset for analysis which will involve converting the data into a format suitable for clustering analysis. This will include standardizing variables to ensure comparability, transforming skewed distributions to achieve normality, and scaling features to bring them within a similar range. Additionally, the RFM (Recency, Frequency, Monetary) methodology is employed to segment customers based on their transactional behavior allowing for the creation of meaningful customer segments that can inform targeted marketing strategies and predict customer lifetime value.

3.5.1 Categorical variable encoding

The dataset comprised categorical variables, which, due to their label representation, are incompatible with direct utilization in machine learning algorithms. To address this, the One-Hot-Encoding method was employed, transforming these categorical variables into numerical values. This technique involves creating separate binary variables, assigning a value of 0 or 1 to each category within the original categorical variable.

3.5.2 Feature Scaling

Feature scaling is a preprocessing technique aimed at adjusting the range of features to a common scale, often within a predefined range like [0, 1] or [-1, 1]. This is crucial because datasets frequently contain features with differing scales, which can impact the performance of machine learning algorithms. In this study, standard scaling was employed, which involves transforming the features to have a mean of zero and a standard deviation of one. This transformation was achieved using the Scikit Learn library in Python.

3.5.3 Feature Engineering

Feature engineering is an essential preprocessing step in data analysis, aimed at enhancing the predictive power and interpretability of the analysis by creating new features from existing data.

3.5.4 Recency, Frequency and Monetary Value

By leveraging techniques like Recency, Frequency, and Monetary (RFM) modelling, customer data can be segmented based on their behavior. Recency measures the time since a customer's last purchase, while Frequency indicates how often a customer makes purchases, and Monetary value represents the total amount spent. Weighted RFM scores are derived by assigning different weights to each component based on their relative importance.

3.5.5 Average Order Value

Average Order Value (AOV) plays a crucial role in assessing customer satisfaction and identifying high-value customers. It indicates the average amount of revenue generated from each order. It's calculated using the formula:

$$\boldsymbol{AOV = \frac{Total\ Revenue}{Total\ Number\ of\ Orders}}$$

AOV provides valuable insights into customer purchasing behavior and helps businesses make strategic decisions regarding pricing, marketing, and customer acquisition. It also plays a crucial role in estimating Customer Lifetime Value (CLV), which represents the total revenue a business can expect from a single customer over their entire relationship. CLV is often computed by multiplying AOV by the average customer lifespan and the average purchase frequency. The relationship between AOV and CLV is significant because a higher AOV generally leads to a higher CLV, indicating greater profitability and customer value. We can define a threshold value for the Average Order Value (AOV) and compare it against the AOV of each customer. Customers whose AOV exceeds this threshold can be classified as high-value customers. Here's the formula:

High-value customers $=\{\text{Customers} \mid AOV_{customer} > \text{Threshold}\}$ where

- High-value customers represents the set of customers classified as high-value customers.
- $AOV_{customer}$ denotes the Average Order Value for a specific customer.
- Threshold is the predetermined threshold value set to identify high-value customers.

3.5.6 Sales Turnover

Another important feature is sales turnover, representing the total revenue generated from sales within a specific period, which provides insights into customer engagement and revenue generation potential. It's an essential metric for understanding the financial performance of a business and plays a crucial role in determining CLV. Here's the equation for computing sales turnover:

$\boldsymbol{Sales\ Turnover} = \sum_{i=1}^{n}$ Total Sales i where

- Total Sales$_i$ represents the total sales for each individual transaction or order within the specified time period.
- $\boldsymbol{n}$ denotes the total number of transactions or orders within the specified time period.

Understanding sales turnover on a monthly, quarterly, and yearly basis provides insights into revenue trends, seasonality, and overall business performance. Analyzing sales turnover alongside other customer-related metrics can help businesses identify patterns, forecast future revenues, and optimize marketing strategies to enhance customer retention and increase CLV.

By visualizing sales turnover over different time periods, we can gain a comprehensive understanding of revenue fluctuations and make informed decisions to improve customer satisfaction, loyalty, and ultimately, CLV.

3.5.7 Customer Lifetime Value

Customer Lifetime Value (CLV) reflects the duration of a customer's relationship with the company and is vital for understanding customer loyalty and designing tailored retention strategies. A basic Customer Lifetime Value (CLV) model, as outlined by (Gupta, 2006) involves incorporating revenues, costs, discount rates, repeat purchase probabilities, and acquisition costs over a specified time frame at an individual customer level and is denoted by the following formula:

$$\boldsymbol{CLV} = \sum_{t=0}^{T} \frac{(pt-ct)rt}{(1+i)^t} - \mathbf{AC}$$

where

$pt = price\ paid\ by\ a\ customer\ at\ time\ t,$

$ct = direct\ cost\ of\ servicing\ the\ customer\ at\ time\ t,$

$i = discount\ rate\ or\ cost\ of\ capital\ for\ the\ firm,$

$rt = probability\ of\ customer\ repeat\ buying\ or\ being$ alive $at\ time\ t,$

$AC = acquisition\ cost\ of\ the\ customer,\ and$

$T = time\ horizon\ for\ estimating\ CLV.$

Through these we gain valuable insights into customer behavior.

3.6 Data analysis and machine learning models

To construct a customer lifetime value prediction model, this project will then proceed through several stages. Following that, data visualization will be conducted to gain insights into the dataset beyond the typical characteristics of the customers. Finally, the best performing model will be employed to establish a customer lifetime value prediction model.

3.6.1 Determining optimal clusters through singular value decomposition

This technique involves decomposing the data matrix into three matrices: **U**, ∑, and **V**. These matrices represent the latent factors, singular values, and singular vectors, respectively. Singular Value Decomposition (SVD) decomposes an $(\boldsymbol{nxm})$ matrix $\boldsymbol{A}$ into three matrices: $\boldsymbol{U}$, ∑, and $\boldsymbol{V}$ such that $\boldsymbol{A = U\Sigma V^T}$. Here, $\boldsymbol{U}$ is an **(*n*×*k*)** orthogonal matrix with its column vectors known as the left singular vectors of $\boldsymbol{A}$, $\boldsymbol{V}$ is a **(*k*×*m*)** orthogonal matrix with its column vectors termed the right singular vectors of $\boldsymbol{A}$, and Σ is a **(*k*×*k*)** diagonal matrix containing the singular values of $\boldsymbol{A}$ ordered in decreasing order.

The columns of $\boldsymbol{U}$ constitute an orthogonal basis for the column space of $\boldsymbol{A}$. SVD is particularly suitable for sparse matrices. Since only the first $\boldsymbol{k}$ concepts are considered semantically important (indicated by high singular values), we approximate the decomposition as:

$$\boldsymbol{A = U_k\Sigma_k V_k^T}$$

where $\boldsymbol{U_k}$ contains the first $\boldsymbol{k}$ most important concept vectors, $\boldsymbol{\Sigma_k}$ contains the respective singular values, and $\boldsymbol{U_k\Sigma_k V_k^T}$ contains the pseudo-document vectors represented using the first *k* concept vectors. In essence, SVD projects the original $\boldsymbol{m}$-dimensional vectors into a vector space of dimension $\boldsymbol{k}$ **(where $k \ll m$).** The SVD approximation, known as rank-$\boldsymbol{k}$ SVD, can be obtained either by "trimming" the full-SVD matrices or by utilizing a specialized method designed to directly perform the rank-$\boldsymbol{k}$ SVD.

How SVD works:

Let $\boldsymbol{A}$ be an $\boldsymbol{nxm}$ rank-r matrix. Let $\sigma_1,\ldots, \sigma_r$ be the eigen values of the matrix $\sqrt{AA^T}$. There exists orthogonal matrices $\mathbf{U}=[u_1,\ldots,u_r]$ and $\mathbf{V}=[v_1,\ldots,v_r]$ whose column vectors are orthonomal and a diagonal matrix $\sum = \text{diag}(\sigma_1,\ldots, \sigma_r)$. The decomposition $\mathbf{A}=U\sum V^T$ is known as a singular value decomposition (SVD) of matrix A, and $\sigma_1,\ldots, \sigma_r$ are the singular values of $\boldsymbol{A}$. The column values of **U** or (**V**) are referred to as the left (or right) singular vectors of matrix $\boldsymbol{A}$. This decomposition provides a representation of the original matrix $\boldsymbol{A}$. It's important to note that the left and right singular vectors are generally not sparse. At most, there are **r** nonzero singular values, where **r** is the smaller of the two dimensions of the matrix. As the singular values typically decrease rapidly, we only consider the $\boldsymbol{k}$ greatest singular values and their corresponding singular vector coordinates to create a $\boldsymbol{k}$- reduced SVD of matrix $\boldsymbol{A}$.

Definition: Let $\boldsymbol{k}$, $\boldsymbol{0<k<r}$ and SVD be the SVD of A.

$$A = \boldsymbol{U\Sigma V^T} = (U_k U_0)\begin{pmatrix} \sum \boldsymbol{k} & \mathbf{0} \\ \mathbf{0} & \sum \mathbf{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{Vk^T} \\ \boldsymbol{V0^T} \end{pmatrix}$$

The decomposition A = $\boldsymbol{U_k\Sigma_k V_k^T}$ is referred to as a k-reduced SVD (or rank-k SVD). The below shows the k-reduced singular value decomposition, where the grey areas indicate the first k coordinates from singular vectors that are utilized.

$(\boldsymbol{Ak})$ = $(\boldsymbol{Uk})$ $(\boldsymbol{\Sigma k})$ $(\mathbf{V}\boldsymbol{k^T})$

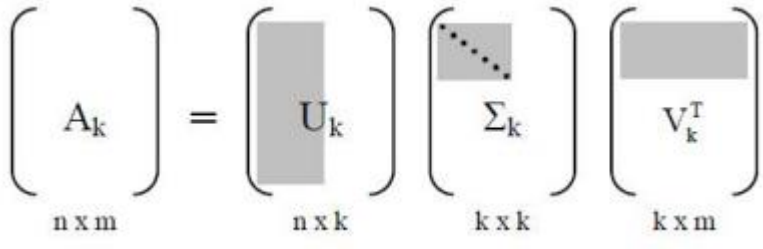


Figure 2 Using SVD as a solution for search result clustering. Hassam D, Abdella S, Vaclav S, PoznanUniversity of Technology Academic Journals, 2014

Among all **n x m** matrices **C** of rank at most ***k*** the one that minimizes $||A\text{-}C||_f^2$ = $\sum_{i,j}(Ai,j - Cw,j)^2$ is the rank ***k*** SVD. This is because the rank-k SVD provides the best rank-k estimation of the original matrix A, and any other decomposition will increase the sum of squares $||A\text{-}C||_f^2$.

Theorem (SVD):

Step1: Generating gradual items

Given dataset **A** represented as an **n x m** matrix, construct ***2m*** gradual items, each containing$(\mathbf{1} - \boldsymbol{\epsilon})\left(\frac{n^2}{2}\right)$ columns. Each column represents a pairwise ranking respecting the gradual variation of the item. Encode pairwise rankings of gradual item ***g*** as a ranking vector ***Rg***. Aggregate ranking vectors to form a ***2m x*** $(\mathbf{1} - \boldsymbol{\epsilon})\left(\frac{n^2}{2}\right)$ ranking matrix ***R.***

Step 2: Constructing the net-win vectors

Define **S =** $\frac{\mathbf{1}}{\boldsymbol{n}}$ $\mathbf{RA^T}$ as the net-win matrix, where the *g-th* row of **S** represents the net-win vector of gradual item ***g.*** This step denoises the $\frac{n^2}{2}$ x (n^2)- dimensional **R** matrix by projecting it onto an n-dimensional **S** matrix.

Step 3: Clustering Gradual Items

Compute $\mathbf{S_e}$, the low-rank **r** approximation of **S** via Singular Value Decomposition (SVD). Form clusters $\boldsymbol{C_1,\ldots C_n}$. Initialize each cluster randomly with a different gradual item ***g***. Then, assign unclustered gradual items to clusters such that each cluster $\boldsymbol{C_k}$ = $\{\boldsymbol{g,:}|| \mathbf{S_{eg}} - \mathbf{S_{eg}}, ||2 \leq \boldsymbol{p}\}$ where ***g*** is an unclustered gradual item and ***p*** is the specified threshold.

For each gradual item ***g*** in cluster $\boldsymbol{C_k}$, estimate the score vector derived, and a cluster defined, which estimates a gradual pattern.

The transformation of the net-win matrix provides a more efficient representation compared to using an **n x m** matrix, especially for sparse data, as it reduces computational memory

requirements. After generating and projecting the ranking matrix $\boldsymbol{R}$ onto the net-win matrix $\boldsymbol{S}$ in Steps 1 and 2, Step 3 computes $\boldsymbol{S_e}$ as a rank $\boldsymbol{r}$ approximation of $\boldsymbol{S}$. Then $\boldsymbol{S_e}$ is used to generate clusters $\{\widehat{\boldsymbol{Ck}}\}$ using a threshold-based algorithm. These clusters $\{\widehat{\boldsymbol{Ck}}\}$ is measured by the number of errors, defined as $\min_{\pi} \sum_{k}|C_k \triangle C_{\pi(k)}|$ where $\triangle$ denotes the symmetric difference between the two sets.

Step 4: Score vector estimation

In cluster $\{\widehat{\boldsymbol{Ck}}\}$, the estimated score vector for each gradual item $\boldsymbol{g}$ is determined by the expression:

$\hat{\theta}_g = \arg\max_{\gamma} \sum_{g,i,j} [\![\{Rgij = 1\}\ \log \frac{e\gamma i}{e\gamma i + e\gamma j}$ for any $\boldsymbol{g}$ and $\boldsymbol{i} < \boldsymbol{j}$

Step 5: Inferring gradual patterns

Estimate a gradual pattern $\widehat{\boldsymbol{Pk}}$ for each cluster $\{\widehat{\boldsymbol{Ck}}\}$. If $\{\widehat{\boldsymbol{Ck}}\}$ comprises $\boldsymbol{g}$ gradual items, then the estimated score vectors $\widehat{\boldsymbol{\theta}}_{\mathbf{k}} = \{\hat{\theta}_{k,1,\ldots,} \hat{\theta}_{k,q}\}$ yields the estimated support value for $\widehat{\boldsymbol{Pk}}$ as:

$\hat{\sigma}_{pk} = \frac{1}{|D'|} * \frac{1}{q} \sum_{x=1,i,j}^{x=q} [\![\{Rgij = 1\}$ for any $\boldsymbol{i} < \boldsymbol{j}$

Where $[\![\{Rgij = 1\}$ represents the set of column indices where column i,j=1 in vector $R_{gx.}$

This decomposition allows us to identify patterns and similarities within the data, effectively grouping customers into clusters based on their characteristics and behaviors.

By leveraging machine learning techniques, companies can proactively identify and address failed customer interactions, leading to increased customer satisfaction, improved brand reputation, and higher revenue. The process-oriented approach presented in this dissertation can guide e-commerce companies in developing effective strategies for managing customer interactions and improving their overall customer experience.

3.6.2 Evaluating cluster performance

After applying SVD, it is important to evaluate the performance of the clusters. This can be done using various metrics such as silhouette score and within-cluster sum of squares using the elbow method. These metrics will help assess the quality and effectiveness of the clustering algorithm in accurately grouping customers with similar characteristics.

3.6.2.1 Silhouette Score

This is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. A silhouette score close to 1 indicates dense, well-separated clusters, while a score close to -1 indicates overlapping clusters. A high silhouette score would imply that the algorithm effectively groups customers based on their characteristics and behaviour. An effective clustering solution should exhibit data points that show greater similarity to their own cluster compared to the nearest neighboring cluster. The Silhouette Score effectively encapsulates this concept by favoring high within-cluster similarity while penalizing low between-cluster similarity.

The silhouette coefficient is calculated using the mean intra-cluster distance (a), which is the average distance of the data point to all other data points in the same cluster and the mean nearest-cluster distance (b) from each sample, which is the minimum average distance of the data point to all other data points in any other cluster. The silhouette score is then given by the formula: $y_i = (bi - ai)/\max(ai, bi)$.

To calculate the silhouette score for a clustering algorithm, we then compute the average silhouette score of all data points in the data set shown in the below formula:

Silhouette score = $\frac{1}{n} * \sum yi$

3.6.2.2 Within cluster sum of squares (WCSS)

WCSS measures the compactness of clusters by summing the squared distances between each data point and its cluster centroid. It quantifies the variability within clusters, with lower values indicating tighter clusters. WCSS is often used as part of the elbow method to determine the optimal number of clusters, where the plot of WCSS against the number of clusters is examined for an "elbow" point representing a significant decrease in WCSS. A lower WCSS value indicates that the algorithm successfully groups customers into segments, facilitating identification of distinct customer profiles. The primary algorithm employed is the Hartigan-Wong algorithm (1979), which computes the total within-cluster variation by summing the squared Euclidean distances between data points and their respective centroids:

$\mathbf{W(Ck)} = \sum_{xi \in Ck}(xi - \mu k)^2$

Where:

- x_i is a data point in cluster C_k
- μ_k denotes the mean value of points assigned to a cluster C_k

Each observation (x_i) is allocated to a specific cluster to minimize the sum of squares distance between the observation and its assigned cluster center (μ_k). We define the total within-cluster variation as follows:

total withinness = $\sum_{k=1}^{k} \boldsymbol{W(Ck)} = \sum_{k=1}^{k} \sum_{xi \in Ck}(xi - \mu k)^2$

The total within-cluster sum of squares measures the compactness (or goodness) of the clustering, with the objective of minimizing it as much as possible.

3.7 Developing the customer lifetime value prediction model

The next step in the methodology is to develop a CLV model for each customer cluster. This can be done by calculating the CLV metric for each customer within a cluster using appropriate methods such as the historic CLV or predictive models.

3.7.1 Gradient Boosting

Gradient Boosting is a machine learning technique that builds a series of weak predictive models (typically decision trees) sequentially, with each model correcting the errors made by its predecessor. The algorithm minimizes a loss function (e.g., Mean Squared Error) by adding

new models to the ensemble, each one focusing on the residuals of the previous model. Mathematically, the prediction $\hat{y}$ of the ensemble is given by:

$\hat{y} = \sum_{i=1}^{n} \alpha i hi(x)$

where

α_i is the weight assigned to the i^{th} model

$h_i(x)$ is the prediction of the i^{th} model.

3.7.2 Neural Networks

Neural networks are computational models inspired by the human brain's structure and functioning. In the context of predicting customer lifetime value, Multilayer Perceptrons (MLPs) are a type of neural network commonly used. An MLP is essentially a feedforward artificial neural network comprising an input layer, one or more hidden layers, and an output layer.

MLPs are versatile in machine learning, capable of tasks such as classification, regression, and time-series forecasting. Each node in an MLP is connected to every node in the subsequent layer through weighted connections. The input layer receives the initial data, while each hidden layer applies nonlinear transformations using activation functions like the sigmoid or ReLU. Finally, the output layer generates predictions, which could be scalar values or vectors.

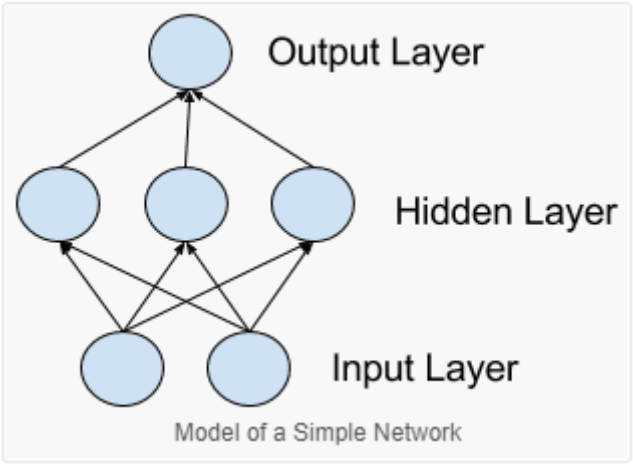


Figure 3When to Use MLP, CNN, and RNN Neural Networks by Jason Brownlee on machinelearningmastery.com

Here's a breakdown of how MLP operates:

- Input Layer: Nodes in this layer correspond to input variables, processing input data by calculating weighted sums.
- Hidden Layers: Each node in hidden layers receives input from nodes in the previous layer, computes weighted sums, and applies activation functions to produce outputs.
- Output Layer: Outputs from the last hidden layer are fed into the output layer, which computes weighted sums and applies an activation function to generate predictions.
- Backpropagation: MLP weights are typically learned through backpropagation, where the error between predicted and actual outputs is propagated backward through the network,

adjusting weights to minimize error. This often involves techniques like stochastic gradient descent.

The purpose of an MLP in predicting customer lifetime value is to discern patterns between input data (such as customer behavior, demographics, etc.) and the target output (lifetime value). By adjusting network weights during training, MLPs capture complex relationships between input and output variables.

MLP Formula:

The mathematical formula for an MLP involves matrix multiplications and activation functions across layers. Denoting the input vector as $\boldsymbol{x}$ weight matrix as $\boldsymbol{w}$ bias vector as $\boldsymbol{b}$ and activation function as $\boldsymbol{f}$ the computations for each layer can be represented as follows:

- For the first hidden layer:

  $\boldsymbol{z1} = \boldsymbol{f}(\boldsymbol{w1} * \boldsymbol{x} + \boldsymbol{b1})$

- **For the subsequent hidden layers:**
  $\boldsymbol{zi} = \boldsymbol{f}(\boldsymbol{wi} * \boldsymbol{zi} - \boldsymbol{1} + \boldsymbol{bi})$

- **For the output layer:**
  $\mathbf{y} = \boldsymbol{f}(\boldsymbol{wo} * \boldsymbol{zlasthidden} + \boldsymbol{bo})$

These computations involve applying activation functions to weighted sums of inputs at each layer. Additionally, these equations can be expressed in vectorized form for efficient computation over entire datasets, particularly using libraries like NumPy. Vectorized equations allow for parallel processing and optimization of computations.

MLPs learn from data by adjusting the weights of connections between neurons to minimize a loss function, such as Mean Squared Error, through techniques like backpropagation. The prediction of a neural network $\hat{\boldsymbol{y}}$ is calculated through forward propagation, which involves passing input data $\boldsymbol{x}$ through the network layers using a series of weighted sums and activation functions. The prediction formula for an MLP, considering $\boldsymbol{L}$ layers, is given as:

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}(\boldsymbol{L}) \cdot \boldsymbol{f}\left(\boldsymbol{W}(\boldsymbol{L}-\boldsymbol{1}) \cdot \ldots \cdot \boldsymbol{f}\left(\boldsymbol{W}(\boldsymbol{1}) \cdot \boldsymbol{x} + \boldsymbol{b}(\boldsymbol{1})\right) + \boldsymbol{b}(\boldsymbol{2})\right) + \cdots + \boldsymbol{b}(\boldsymbol{L})\right)$$

Here,$\boldsymbol{W}(\boldsymbol{i})$ and ,$\boldsymbol{b}(\boldsymbol{i})$ represent the weights and biases of the , $\boldsymbol{i}$'th layer, $\boldsymbol{f}$ is the activation function and $\boldsymbol{x}$ is the input data. This formula illustrates how input data is processed through the network layers to produce predictions. Each layer applies its weights and biases along with an activation function to transform the data, contributing to the final prediction.

3.7.3 K Nearest Neighbours

The K-Nearest Neighbors (K-NN) algorithm is a straightforward supervised learning method used for both classification and regression tasks. It operates on the principle of similarity, assuming that new data points are similar to those in the dataset. K-NN finds the similarity between new data and existing data points to categorize the new data. It stores all available data and classifies new data points based on their similarity to existing ones. K-NN demonstrates flexibility as it can be applied to both regression and classification tasks, although it's more commonly used for classification. Additionally, K-NN is non-parametric, meaning it

doesn't make assumptions about the underlying data distribution. It follows a lazy learning approach, postponing learning until classification, as it stores the entire dataset and classifies new data when needed.

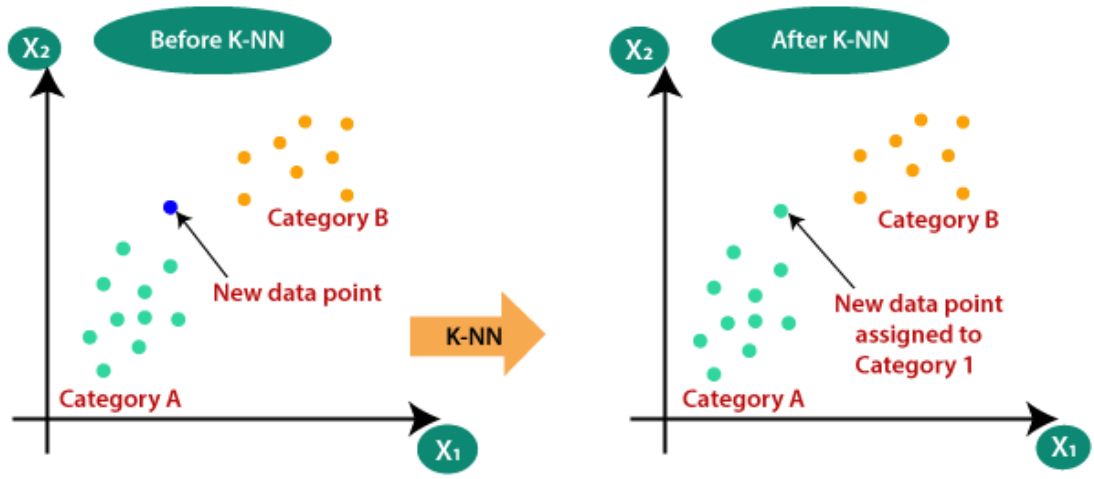


Figure 4 k-nearest-neighbor-algorithm-for-machine-learning, javatpoint.com

The K-NN algorithm's steps include:

- Selecting the value of K, which determines the number of neighbors to consider.
- Calculating the Euclidean distance between the new data point and existing data points.
- Choosing the K nearest neighbors based on the calculated distances.
- Counting the number of data points in each category among the K nearest neighbors.
- Assigning the new data point to the category with the highest number of neighbors.
- Completing the model, making it ready for predictions.

In KNN, the prediction for a new data point is based on the majority class (for classification) or the average of the k nearest data points (for regression). The prediction $\hat{y}$ for a new instance is given by:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} yi$$

where y_i is the value of the i^{th} nearest neighbor.

3.7.4 Decision Trees

Decision Trees serve as the fundamental building blocks of all tree-based models and consist of three key components: nodes, links, and leaves. Nodes represent the features or attributes defining the dataset under study, while links or branches indicate decisions made based on those attributes. Leaves, akin to final-stage nodes, signify the outcomes of the prediction model. Traversal of the Decision Tree begins from the root node, situated at the top, and progresses downward to the leaves at the bottom. To classify an unknown instance, the model commences at the root and follows the branch indicated by the outcome of each test until reaching a final node (leaf node).

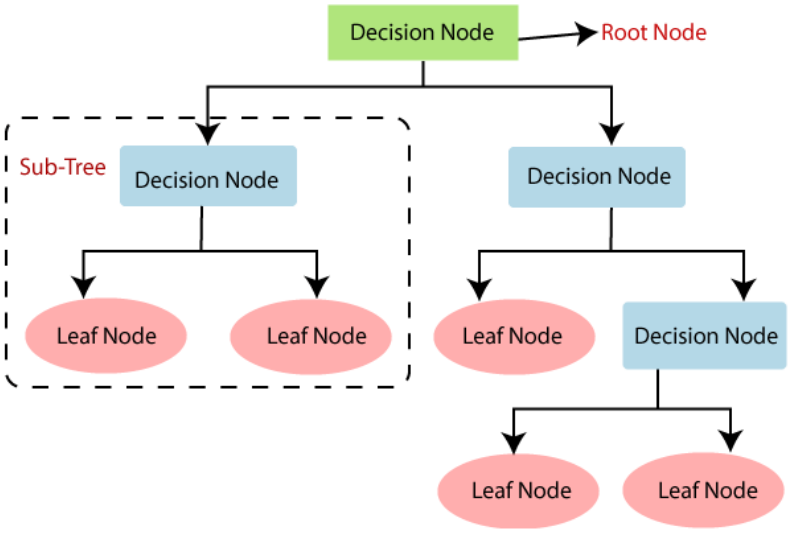


Figure 5The general structure of a decision tree, javatpoint.com, machine-learning-decision-tree -algorithm

Decision Trees can be classified as either Classification Trees (for predicting categorical values) or Regression Trees (for predicting real numbers). In regression trees, the predicted response is determined by the mean response of training observations within the same terminal node, while for classification, the final prediction aligns with the most prevalent class in the node. This adaptability of Decision Trees in handling both discrete and continuous attributes stands as a significant advantage of this algorithm. Decision Trees, being straightforward yet potent, are utilized for both regression and classification tasks. In these trees, datasets are partitioned into subsets based on feature values to minimize impurity measures like Gini impurity or entropy. The prediction $\hat{y}$ for a new instance is determined by traversing the tree from the root node to a leaf node:

$\hat{y}$ = value of leaf node

3.7.5 Linear Regression

Linear regression stands as a straightforward yet powerful regression technique widely utilized in machine learning. It operates under the assumption that the data adheres to the following equation:

$$\boldsymbol{y_i} = \beta 0 \ + xi1\beta 1 + \cdots + xip\beta p + \epsilon i$$

where ***yi*** represents the i-th data point, $x_{i,j}$ signifies the j-th feature of the i-th data point βj denotes the j-th coefficient corresponding to the j-th feature and ϵi symbolizes normally distributed white noise with zero-mean and constant variance. The above equation can be more concisely expressed in vector form as:

$\boldsymbol{y_i} = x_i\beta+\epsilon_i$

where

$x_i = [1, x_{i1}, x_{i2}\ldots,x_ip]$,

$\beta^T= [\beta_0, \beta_1, \beta_2,\ldots, \beta_p]$

Linear regression assumes that the target variable depends on the features, and that the features are independent, denoted as $\boldsymbol{C}[x_{ij}, x_{ik}]=0$ for all $j \neq k$. The regressor is defined as:

$$\widehat{\boldsymbol{yi}} = \mathbf{E}[\mathbf{y_i}] = \mathbf{E}[\mathbf{x_i}\beta+\epsilon_i] = \mathbf{x_i}\beta$$

Utilizing the regressor from the above equation, the residual can be calculated as:

$$\mathbf{r_i} = y_i - \widehat{yi} = \mathbf{x_i}\beta+\epsilon_i - \mathbf{x_i}\beta = \epsilon_i$$

Equation 5.4 suggests that if the deterministic trend of the data is precisely identified, then the residual is expected to be zero-mean white noise with constant variance. It can be demonstrated that the optimal estimation of the β coefficients in the regressor equation above is the solution to the least squares problem:

$$\boldsymbol{\beta} = \text{argmin}_\beta ||y\text{-}X\,\beta||_2^2$$

where y is a vector containing the elements y_i, and X is a matrix where the i-th row contains the features of the i-th element. One drawback of linear regression is its sensitivity to outliers, as the minimization problem in equation 5.5 involves a square operator, amplifying the effects of outliers. Additionally, the requirement for uncorrelated features imposes a stringent constraint that is often unmet. Correlated feature variables introduce uncertainties in the β estimates, resulting in uncertain models. Moreover, even if a regressor exists that describes the data trend, identifying these features is not trivial.

3.7.6 Evaluation criteria

Evaluation criteria for these models typically include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). MAE measures the average absolute difference between predicted and actual values, while MSE measures the average squared difference. RMSE is the square root of MSE, providing an interpretable scale in the same units as the target variable. R^2 represents the proportion of variance in the target variable that is explained by the model, with values closer to 1 indicating a better fit.

- The Mean Absolute Error (MAE) quantifies the magnitude of the average discrepancy and is defined as (Jasek et al., 2018):

  $$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}|\boldsymbol{Ai} - \boldsymbol{Fi}|$$

  where
  Ai *is the actual profit of the customer during the holdout period,*
  Fi *is the predicted profit of the customer during the holdout period and*
  n *is the number of customer*s
- Root Mean Squared Error (RMSE) characterizes the dispersion of prediction errors and is articulated as (Hyndman and Koehler, 2006):

  $$\boldsymbol{RMSE} = \sqrt{\boldsymbol{mean}(\boldsymbol{Ai} - \boldsymbol{Fi})^{\mathbf{2}}}$$

- The Mean Squared Error (MSE) is a measure of the average squared discrepancy between the actual and predicted values. It can be defined using the formula:

  $$\boldsymbol{MSE}= \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{Ai} - \boldsymbol{Fi})^{\mathbf{2}}$$

- The coefficient of determination, $\boldsymbol{R^2}$ also known as R-squared, represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated using the formula:

$$\boldsymbol{R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Ai - Fi)^2}{\frac{1}{n}\sum_{i=1}^{n}(Ai - \bar{A})^2}}$$

Where $\bar{A}$ is the mean of the actual values A_i. This formula compares the MSE of the model to the MSE of a baseline model that always predicts the mean of the dependent variable. A higher $\boldsymbol{R^2}$ value indicates a better fit of the model to the data.

3.8 Deployment

The deployment process for the trained model involves saving it to a file using serialization techniques such as Joblib in Python. This saved model file is then stored in a secure location, such as a cloud storage service or a dedicated server, to ensure accessibility. Integration with applications or services is facilitated by developing APIs or web services that expose endpoints for receiving input data and returning predictions. Regular monitoring and maintenance of the deployed model are essential to ensure continued performance and accuracy over time. This includes tracking performance metrics, detecting drift in input data distribution, and periodically retraining the model. Version control is maintained to track changes and updates to the deployed model, ensuring reproducibility and traceability of predictions. Overall, the deployment process aims to make the trained model accessible, reliable, and scalable for real-world applications, providing valuable insights and predictions to end-users.

3.9 Leveraging SVD for improved marketing strategy in E-commerce

Once the clusters and CLV models are developed, the final step in the methodology is to leverage the SVD and customer lifetime value approach for improved marketing strategy in e-commerce.

This involves using the insights gained from the clustering analysis and CLV models to tailor marketing strategies for each customer cluster.

By understanding the unique characteristics and behaviors of each cluster, marketers can personalize their messaging, offers, and promotions to better resonate with customers in different clusters.

4. SYSTEM DESIGN AND ARCHITECTURE

This chapter delves into the design and architecture of our system, which is centered around leveraging clustering techniques for customer lifetime prediction in the e-commerce sector. Here, we explore the components of the system, its architecture, user registration and authentication processes, and the deployment of machine learning prediction endpoints using Streamlit.

This section also presents the design and architecture of the system, which automates the Extract, Transform, and Load (ETL) process by retrieving Excel files from the Point of Sale (POS) platform database. The system interacts with an API endpoint housing the machine learning models. These models predict customer lifetime value based on the transformed data, which is then appended to the existing dataset. Subsequently, the data is accessed and visualized using Streamlit. This integration with Streamlit allows for an intuitive interface where users can explore and interpret the predictions generated by the machine learning models.

4.1 System components

Our system comprises several key components:

1. **Point of Sale (POS) Platform Integration**: This component involves retrieving Excel files from the Point of Sale platform of the e-commerce venture. These files contain transactional data that will be used for analysis and modeling.
2. **Python automated ETL pipeline:** The Extract, Transform, Load (ETL) pipeline processes the raw data extracted from the POS platform. It involves cleaning, transforming, and preparing the data for analysis and modeling.
3. **Machine Learning Models:** Clustering techniques are applied to the preprocessed data to segment customers based on their behavior and characteristics. These models are trained and evaluated to predict customer lifetime value.
4. **API for User Interaction:** An API is developed to facilitate user registration, authentication, and interaction with the system. Users can register, login, and access machine learning prediction endpoints through this API.
5. **Deployment Platform:** Streamlit is used as the deployment platform for hosting the machine learning prediction endpoints. It provides an intuitive interface for users to interact with the models and view predictions.

4.2 Overview of the System Architecture

The system architecture depicts the interaction between these components. Data flows from the Point of Sale platform through the ETL pipeline to the machine learning models. Users interact with the system through the API and access the deployed prediction endpoints via Streamlit.

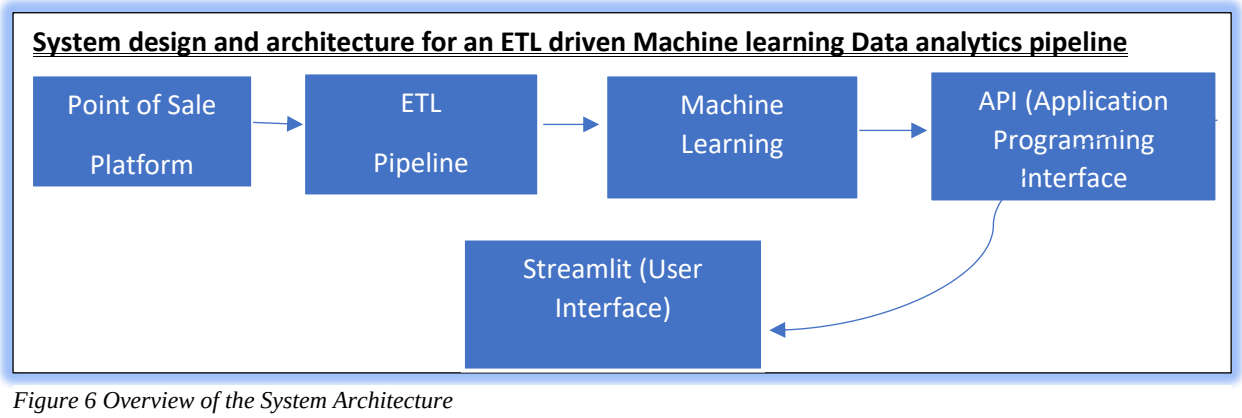


Figure 6 Overview of the System Architecture

4.4 API User Login and Token Generation

After registration, users can log in to the system using their credentials. Upon successful authentication, the API generates a unique access token for the user. This token is required for accessing protected endpoints and ensuring secure communication between the client and server.

4.5 Authentication and Authorization

Authentication ensures that only registered users with valid credentials can access the system. Authorization mechanisms are implemented to control access to different endpoints based on user roles and permissions. This ensures that sensitive functionalities are only accessible to authorized users.

4.6 Machine Learning Prediction Endpoint

The machine learning prediction endpoint is deployed using Streamlit. Users can input relevant data, such as customer attributes, and receive predictions regarding customer lifetime value based on the trained best performing model, in this case, it was Gradient Boosting Model.

Below is a snippet of the interface before user inputs for prediction:

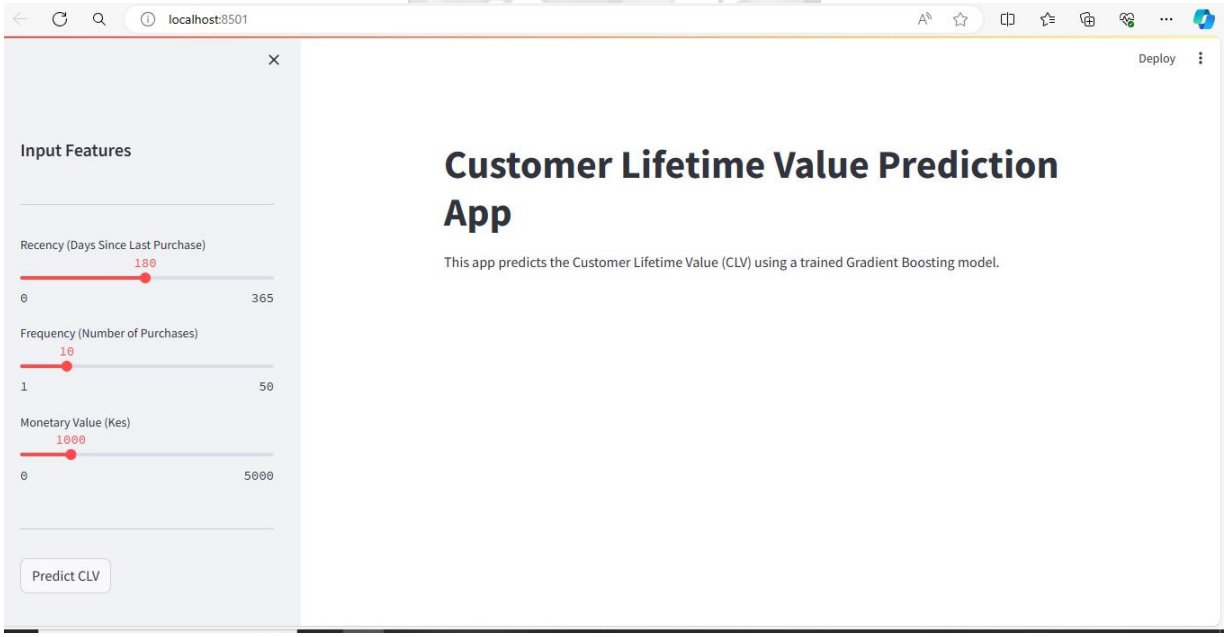


Figure 7 Interface before user inputs for prediction

Below is a snippet of the predicted values after user inputs:

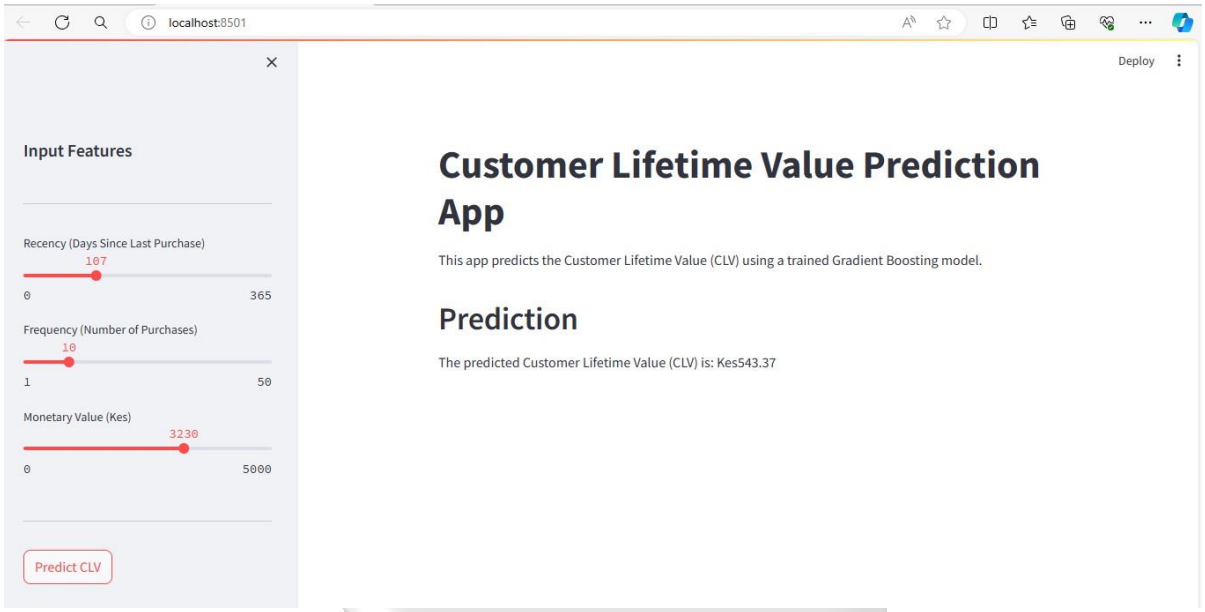


Figure 8 Predicted values after user inputs

5. RESULTS

5.1 Exploratory Data Analysis

The paper employed an exploratory data analysis approach to comprehensively understand the distribution, relationships, and interactions within the dataset. This method allows for a deep dive into the data, uncovering patterns, outliers, and correlations that can provide valuable insights for further analysis and decision-making processes. By adopting an exploratory approach, we can gain a nuanced understanding of the complexities inherent in the dataset, leading to more informed conclusions and potentially uncovering unexpected findings. To start with, the data consists of 28,755 rows and 17 columns.

5.1.1 Univariate exploratory analysis

This involved examining a single variable at a time within the dataset to understand its distribution, central tendency, and variability to gain insights into its characteristics and behavior.

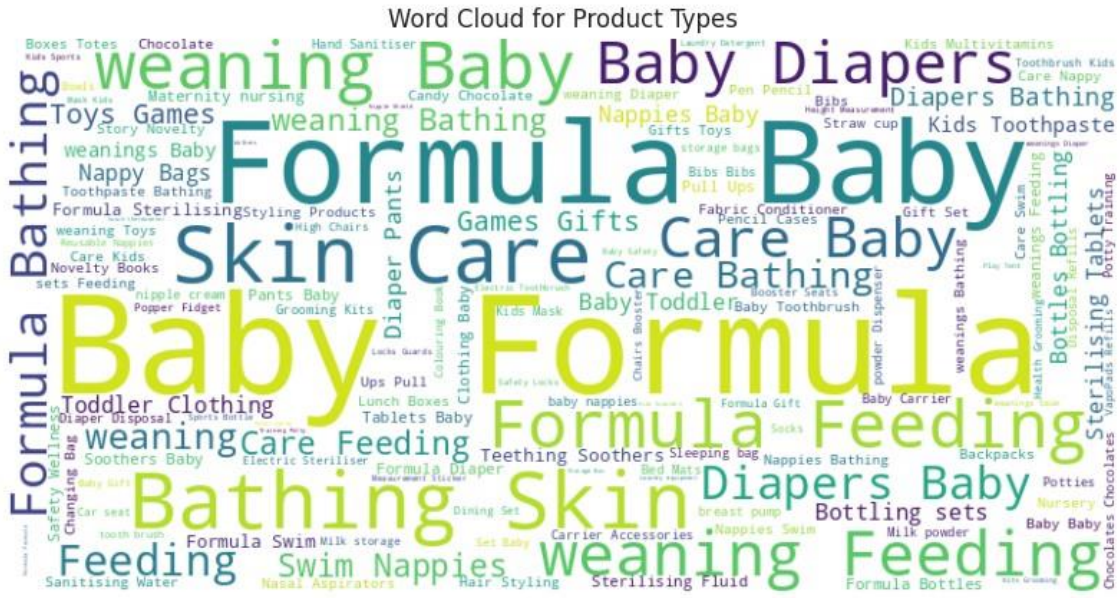


Figure 9 Word Cloud for Product types

From the above wordcloud, it is evident that the most purchased product type is Baby Formula.

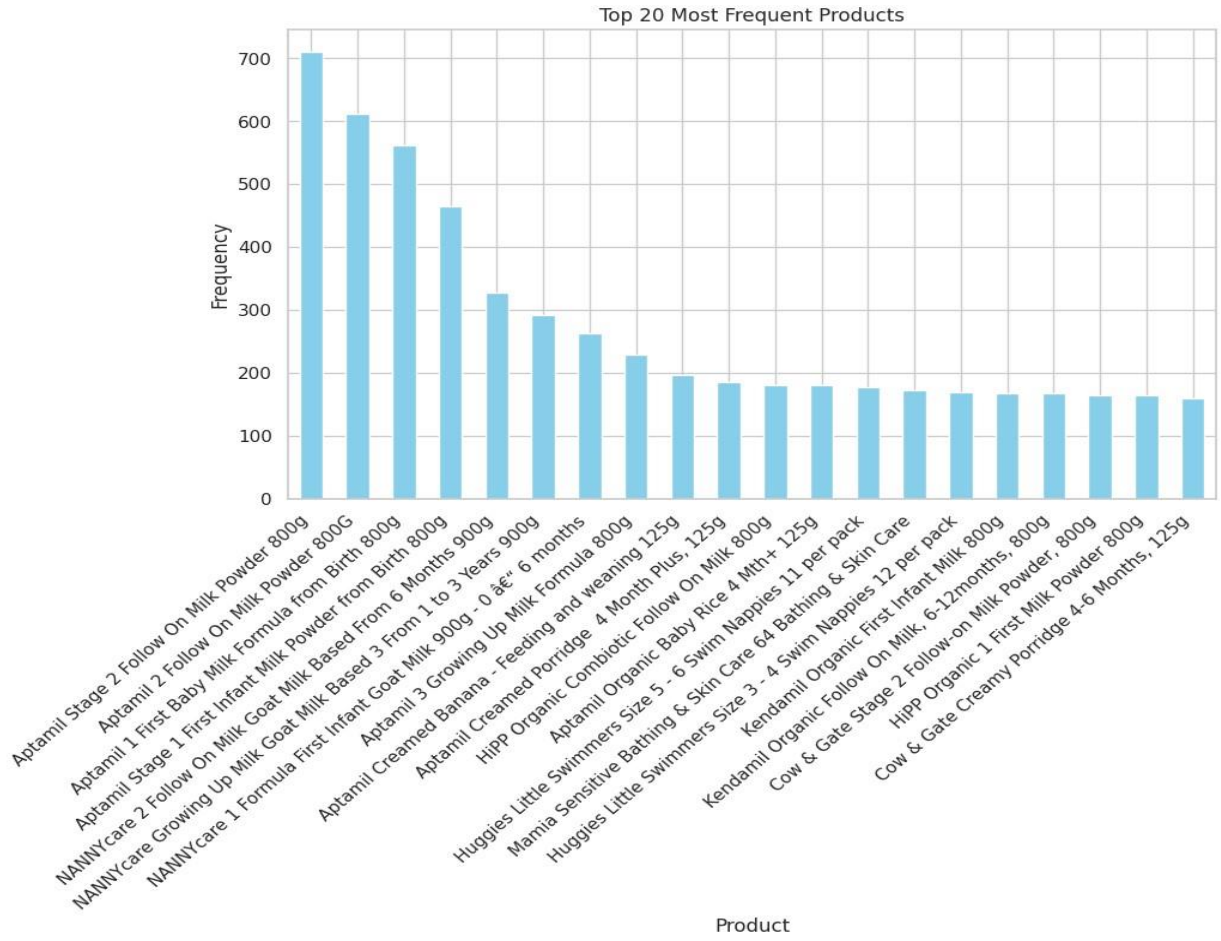


Figure 10 Top 20 most purchased products

The above chart shows the top 20 most purchased products. The most purchased product is the Aptamil Stage 2 Follow on Milk Powder 800g.

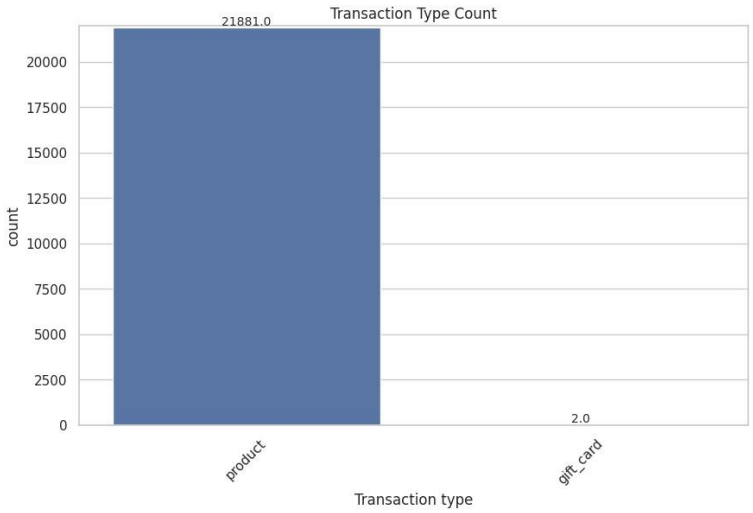


Figure 11 Transaction type count

Based on the provided results, it's evident that the majority of transactions in the dataset are categorized as "product" transactions, comprising a substantial count of 21,881 occurrences. This indicates that product transactions significantly outnumber other transaction types, suggesting that the primary activity within the dataset revolves around the sale or purchase of products. Additionally, there are only two instances of transactions categorized as "gift_card." While this category represents a minor proportion of the overall transactions, it still indicates the presence of gift card transactions within the dataset.

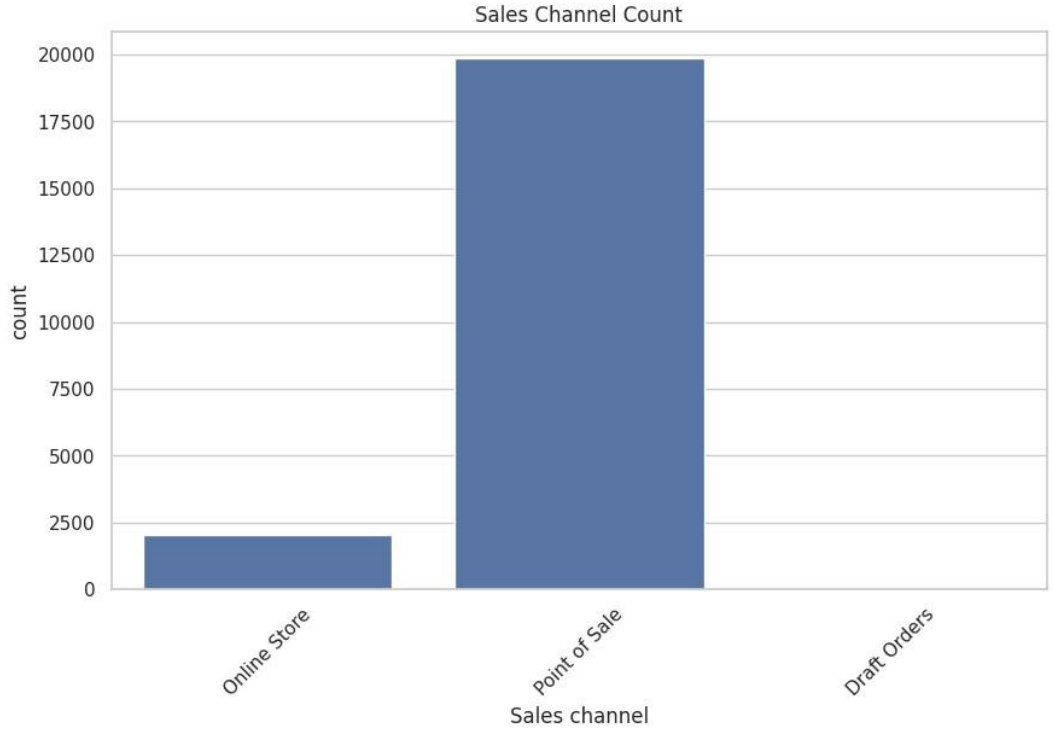


Figure 12 Sales channel count

From the above visual, the online payment sales channel was less popular than the point of sale channel. This could be attributed to the ease and efficiency the point of sale (mpesa) has over the credit cards for online payment.

5.1.2 Bivariate exploratory analysis

Bivariate exploratory analysis serves as a statistical method utilized to examine the relationship between two variables within a dataset. By scrutinizing the association between these variables, researchers can extract valuable insights regarding the interconnections and patterns present among distinct data points. Through this analytical approach, the strength, direction, and significance of the relationship between variables can be assessed, offering a deeper understanding of the underlying dynamics within the data.

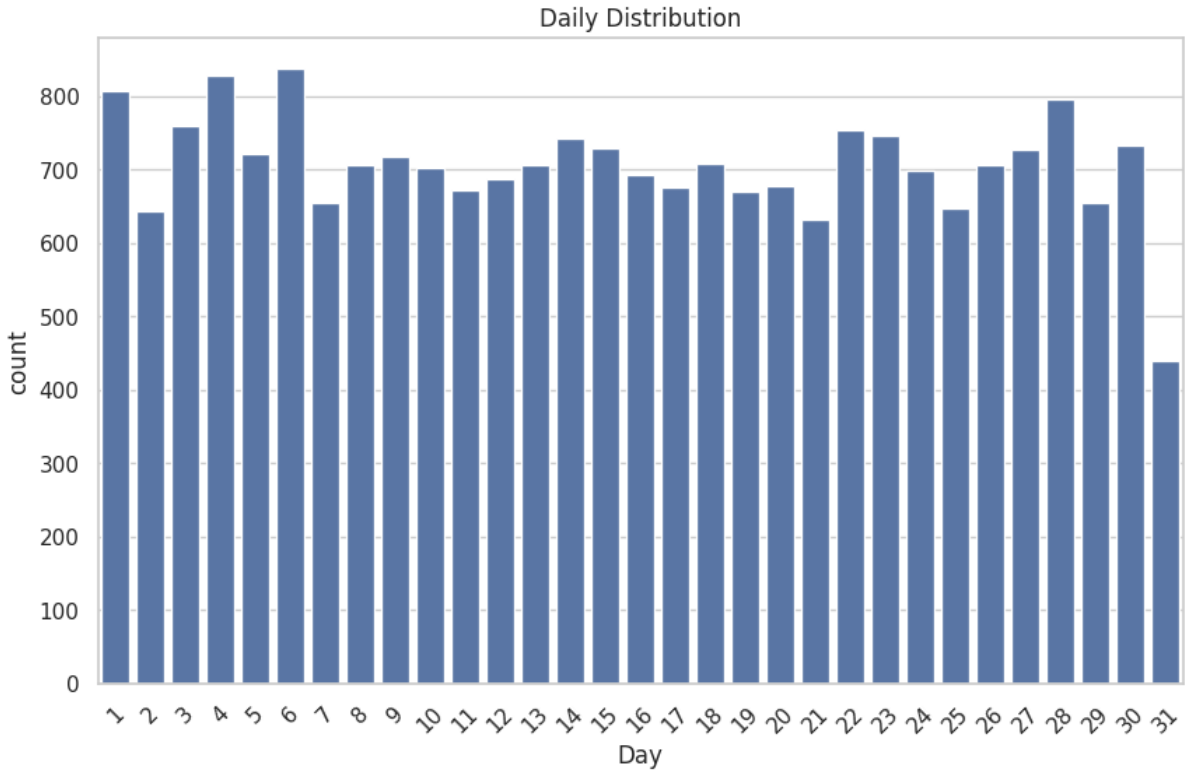


Figure 13 Daily distribution of products purchased

The above shows that most customers prefer to purchase items between the first six days of the month as well as the last four days of the month.

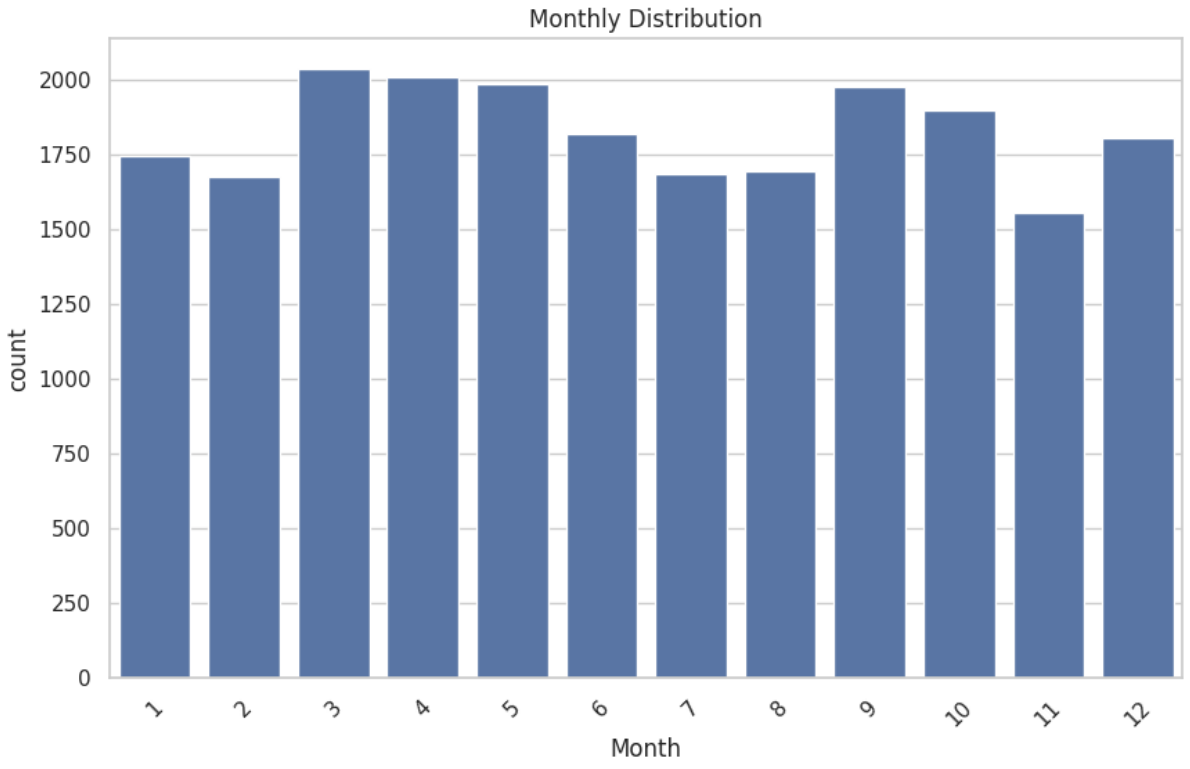


Figure 14Monthly distribution of products purchased

The above shows that most purchases occur between the months of March to May followed closely by purchases made between the months of September and October then December.

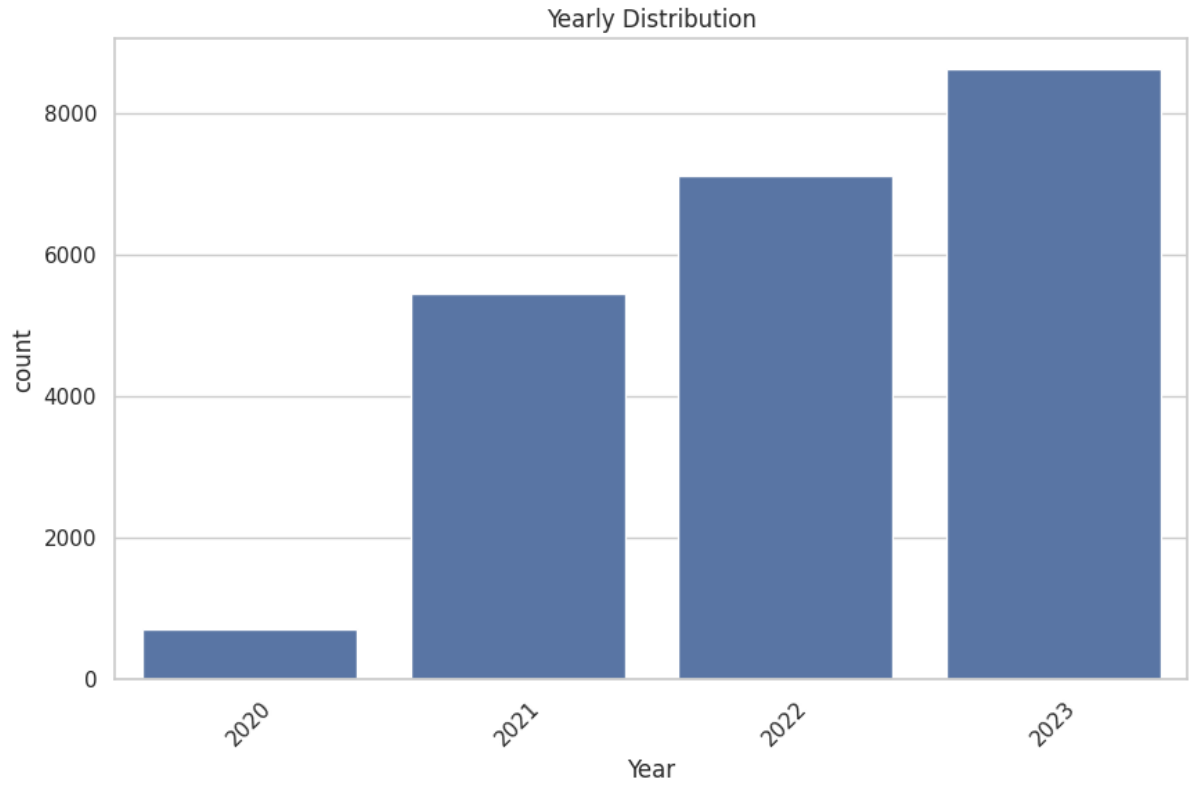


Figure 15Yearly distribution of products purchased

From the graph above, there's a steady increase in the number of customers purchasing products from 2020 to 2023.

5.1.3 Multivariate exploratory analysis

This entailed simultaneously analyzing multiple variables within the dataset to explore their relationships, dependencies, and interactions. By considering the collective behavior of several variables, we were able to discern patterns and detect correlations.

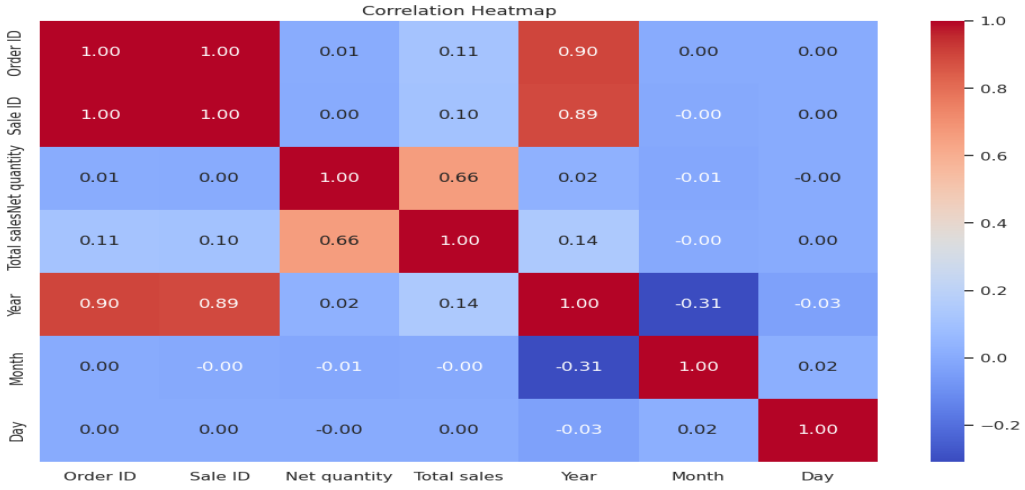


Figure 16 Correlation heatmap

Based on the above heatmap with the correlation coefficients, below is the interpretation of the relationships between the variables:

- Order ID & Sale ID: These variables exhibit a very strong positive correlation of approximately 0.99. This indicates that there is almost a perfect linear relationship between the Order ID and Sale ID, which is expected since they likely represent identifiers for the same transaction.
- Net Quantity & Total Sales: There is a moderate positive correlation of approximately 0.66 between Net Quantity and Total Sales. This suggests that there is a tendency for higher quantities of items to be associated with higher total sales, which is reasonable.
- Year & Order ID/Sale ID: Year exhibits a strong positive correlation with Order ID and Sale ID, indicating that the order/sale IDs are likely assigned sequentially over time.
- Total Sales & Year: There is a weak positive correlation of approximately 0.14 between Total Sales and Year. This suggests a slight tendency for sales to increase over the years, but the correlation is not very strong.

5.2 Model performance evaluation

In this chapter, we delve into the results obtained from employing clustering techniques, Singular Value Decomposition (SVD), to segment customers based on Recency, Frequency, and Monetary (RFM) values. Furthermore, we explore the optimal number of clusters through the Within-Cluster Sum of Squares (WCSS) metric and utilize the elbow method for visualization. Finally, we discuss the application of these segmentation insights in predicting customer lifetime value (CLV).

5.2.1 Clustering analysis

We began by applying Singular Value Decomposition (SVD), a dimensionality reduction technique, to the RFM dataset. SVD enables us to identify underlying patterns in the data by decomposing it into singular vectors and values. Through this process, we aimed to group customers based on their RFM characteristics.

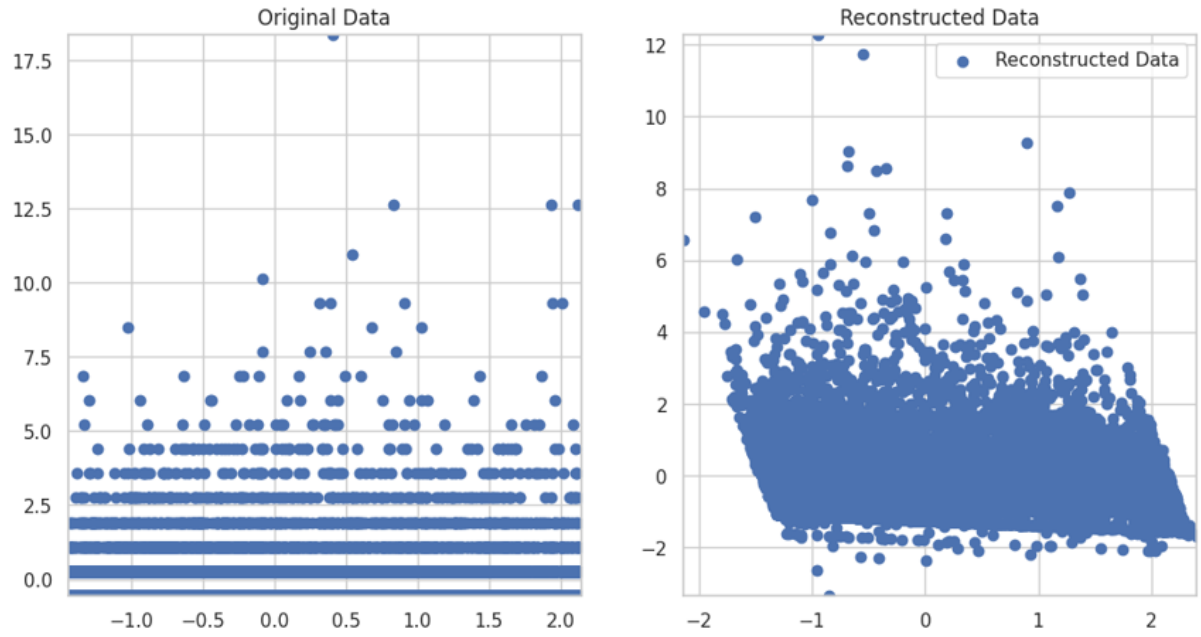


Figure 17 Original vs reconstructed data using SVD

To determine the optimal number of clusters for segmentation, we utilized the Within-Cluster Sum of Squares (WCSS) metric. WCSS measures the compactness of clusters by calculating the sum of squared distances between each data point and its assigned cluster centroid. We employed the elbow method, which involves plotting the number of clusters against the corresponding WCSS values. The point where the rate of decrease in WCSS slows down (the "elbow point") indicates the optimal number of clusters. From the below, the optimal number of clusters is 4.

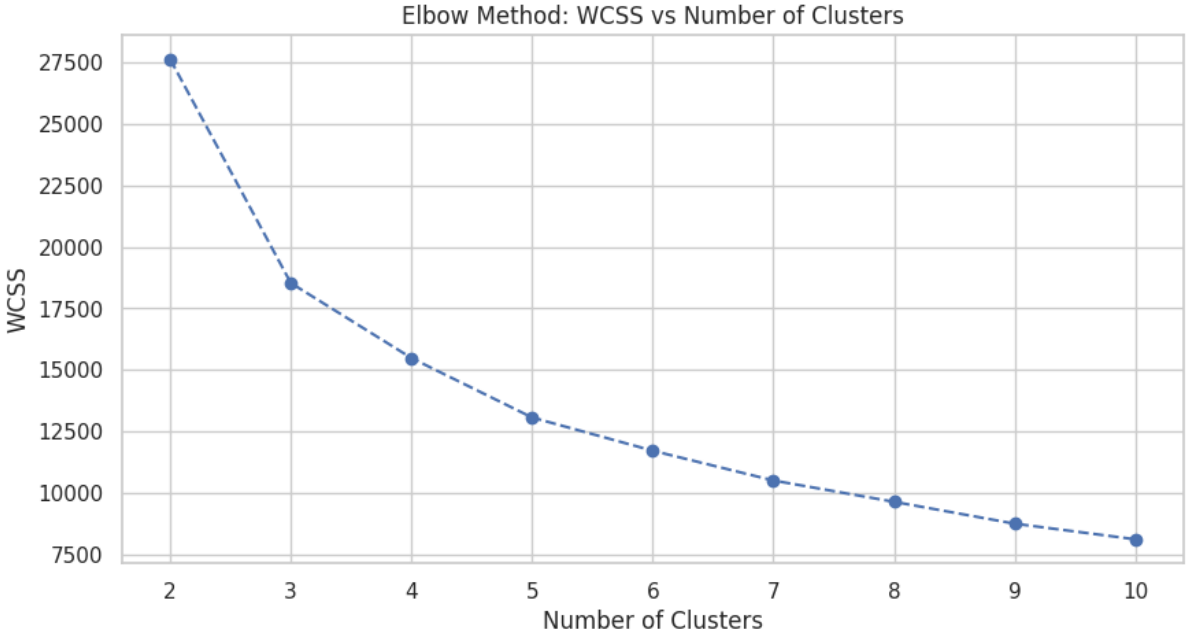
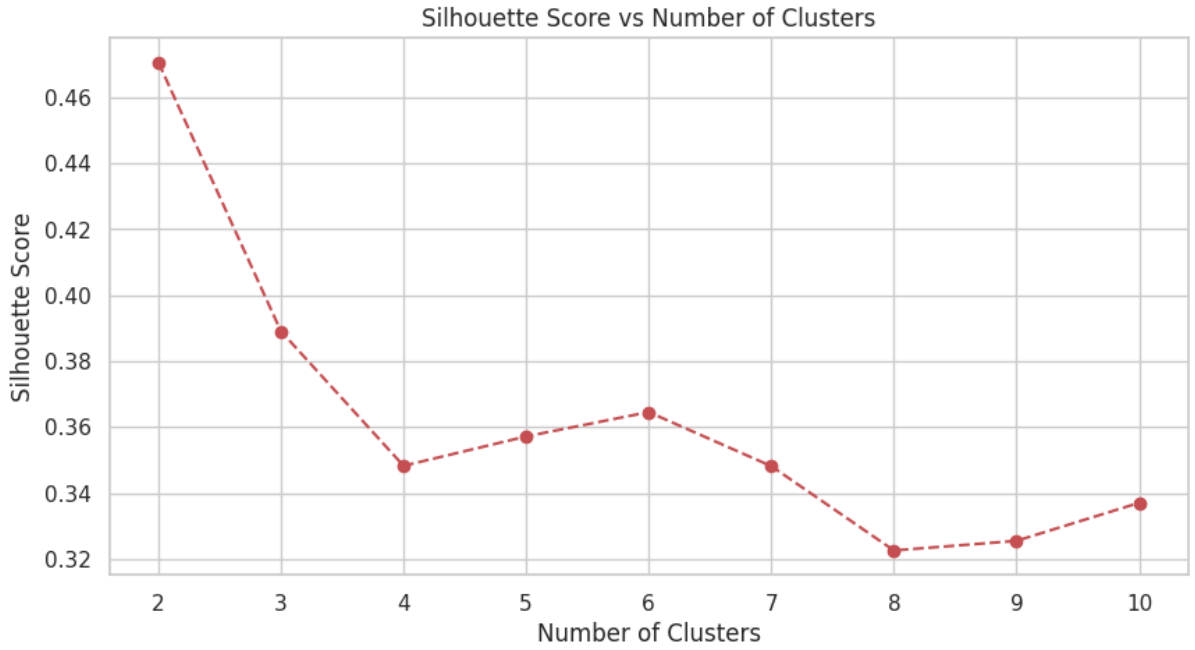


Figure 18 Elbow Method for optimal number of clusters

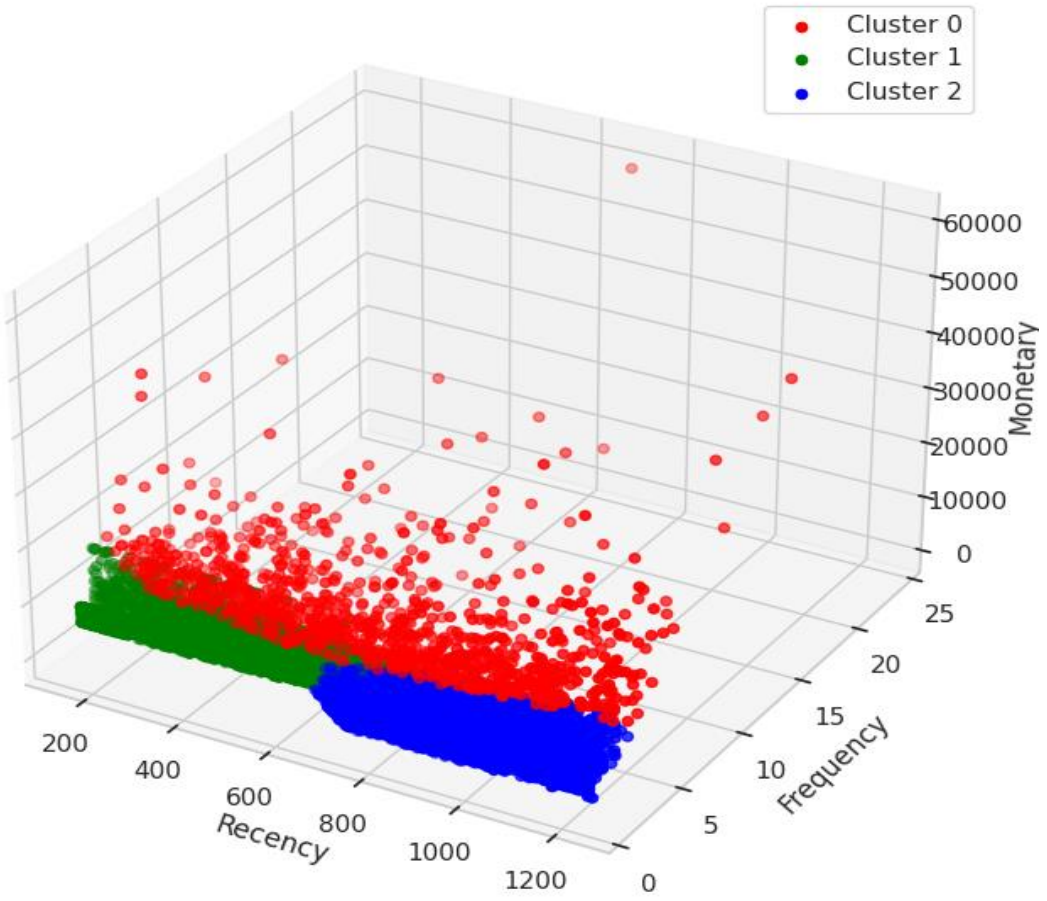


5.2.2 Recency, Frequency, and Monetary Segmentation

Following the determination of the optimal number of clusters, we performed segmentation based on Recency, Frequency, and Monetary values. Recency refers to the time elapsed since the last transaction, Frequency represents the number of transactions within a specific period, and Monetary denotes the total value of transactions. By assigning RFM scores to each customer, we categorized them into distinct segments, enabling targeted marketing strategies and personalized offerings.

Order ID	Recency	Frequency	Monetary	R	F	M	RFM Score	Cluster	RecencyScore	FrequencyScore	MonetaryValueScore	RFM_Score	segment	CLV
2854884835480	1232	1	200.0	1	1	1	111	2	3	0	0	3	Mid to High Value	216800.0
2864488448152	1227	1	8800.0	1	1	4	114	2	3	0	3	6	Mid to High Value	9539200.0
2864491954328	1227	2	3300.0	1	3	2	132	2	3	0	1	4	Mid to High Value	1788600.0
2864497066136	1227	1	3000.0	1	1	1	111	2	3	0	0	3	Mid to High Value	3252000.0
2864504438936	1227	1	3000.0	1	1	1	111	2	3	0	0	3	Mid to High Value	3252000.0

Figure 19 Recency, Frequency and Monetary Value scores and clusters



5.2.3 Customer Lifetime Value Prediction models performance and evaluation

Having segmented the customers based on RFM scores, we proceeded to leverage machine learning models for predicting Customer Lifetime Value (CLV). By integrating features derived from RFM scores along with additional demographic and behavioral data, we trained and evaluated several models including Gradient Boosting, K-Nearest Neighbors (KNN), Neural Networks, Decision Trees and Linear Regression. Here, model training and evaluation was performed using a nested loop structure. Firstly, the code iterated over each model in the models dictionary, where each model is associated with a name (model_name) and an instance of the model (model). Within this loop, another loop iterated over each metric in the metrics dictionary, where each metric is associated with a name (metric_name) and a metric function (metric_fn). Inside the inner loop, the model was trained on the training data (X_train, y_train) using the fit() method. Then, the model's performance was evaluated using cross-validation with 5 folds (cv=5) and a specific evaluation metric, in this case, negative mean squared error (scoring='neg_mean_squared_error'). The negative mean squared error was converted to positive to maintain consistency. Additionally, the model's performance was evaluated on a separate test set (X_test, y_test) using the specified metric function. Finally, the results of each model's evaluation, including the evaluation metric scores, were stored in the results dictionary, where each model's name is associated with its corresponding evaluation scores. It's worth mentioning that the dataset was split into training and testing sets in an 80:20 ratio, with 80% of the data used for training and 20% for testing. This process enables the comparison of different models based on their performance metrics on the test set. Below are the results:

Table 2 Models performance metrics

Model	MAE	MSE	RMSE	R2
Gradient Boosting	1.101617e+05	3.802703e+10	1.950220e+05	0.994361
KNN	5.068723e+05	1.061554e+12	1.030318e+06	0.842619
Neutral Networks	1.648331e+06	5.209136e+12	2.266125e+06	0.226061
Decision Trees	2.476279e+03	1.392117e+09	3.731109e+04	0.979361
Linear Regression	5.976193e+05	5.557691e+12	7.454992e+04	0.917604

The Gradient Boosting model exhibits a relatively low MAE, MSE, and RMSE, indicating that it produces accurate predictions with minimal errors. Additionally, the high R-squared value of 0.9943 suggests that approximately 99.43% of the variance in the CLV can be explained by the model. This model performs exceptionally well and is highly suitable for CLV prediction.

The KNN model exhibits higher MAE, MSE, and RMSE compared to Gradient Boosting, indicating less accuracy and higher prediction errors. The R-squared value of 0.8426 suggests that approximately 84.26% of the variance in the CLV is explained by the model. While KNN may provide reasonably accurate predictions, its performance is inferior to Gradient Boosting in this context.

The Neural Networks model exhibits significantly higher MAE, MSE, and RMSE compared to both Gradient Boosting and KNN, indicating poor predictive performance and substantial errors. The low R-squared value of 0.2260 suggests that only 22.60% of the variance in the CLV is explained by the model. Neural Networks may not be suitable for CLV prediction in this context due to its poor performance.

The Decision Trees model exhibits the lowest MAE, MSE, and RMSE among all models, indicating high accuracy and minimal prediction errors. The exceptionally high R-squared value of 0.9794 suggests that approximately 97.94% of the variance in the CLV is explained by the model. Decision Trees perform exceptionally well and are highly suitable for CLV prediction in the ecommerce sector.

The Linear Regression model explains roughly 91.8% of the variance in CLV, as denoted by its R-squared (R2) value of 0.918. These metrics collectively suggest that the model effectively captures the relationship between predictor variables and CLV.

In conclusion, Gradient Boosting and Decision Trees emerge as the top-performing models for CLV prediction, exhibiting high accuracy and minimal prediction errors. These models provided valuable insights in understanding and predicting customer lifetime value, facilitating strategic decision-making, personalized marketing strategies, and enhanced customer relationship management. On the other hand, KNN and Neural Networks show comparatively inferior performance and may not be suitable for CLV prediction in this context. We select Gradient Boosting model as the best performing model for CLV prediction and below is its performance of Actual vs predicted CLV.

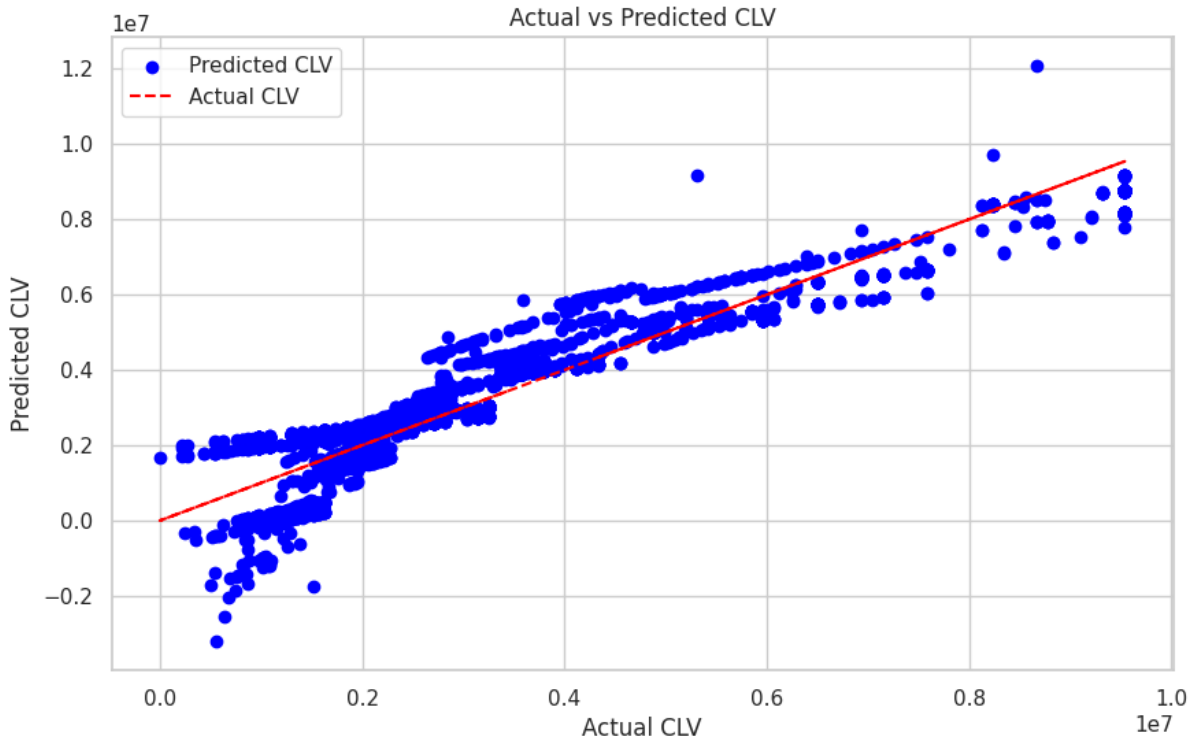


Figure 20 Actual vs Predicted CLV

6. DISCUSSION OF FINDINGS

6.1 Customer segmentation findings

In the context of customer segmentation using RFM (Recency, Frequency, Monetary) analysis, the provided RFM Cluster to Segment Mapping categorizes customers into different clusters based on their RFM scores. Here's a thorough interpretation along with marketing strategies that can be assigned to each cluster:

1. Cluster 0: High Value

Customers in this cluster exhibit high recency, frequency, and monetary value. They are the most valuable customers for the business, as they have made recent purchases, make purchases frequently, and spend a significant amount of money.

Marketing Strategies:

VIP Treatment: Provide exclusive offers, early access to new products, and personalized services to make them feel valued and appreciated.

Loyalty Programs: Offer loyalty rewards, points, or tiers to incentivize repeat purchases and foster long-term loyalty.

Upselling and Cross-selling: Recommend premium products or complementary items to increase their average order value.

Feedback and Engagement: Solicit feedback and engage with them through surveys, reviews, and social media to understand their preferences better and enhance their shopping experience.

2. Cluster 2: Mid to High Value

Customers in this cluster have moderate to high recency, frequency, and monetary value. While they may not be as active or valuable as those in Cluster 2, they still contribute significantly to the business's revenue.

Marketing Strategies:

Segmented Promotions: Offer targeted promotions and discounts based on their purchase history and preferences to encourage repeat purchases.

Product Recommendations: Use personalized product recommendations based on their past purchases to increase cross-selling opportunities.

Reactivation Campaigns: Implement campaigns to re-engage customers who have shown a decrease in activity to prevent churn and encourage them to make new purchases.

Customer Feedback: Encourage feedback and reviews to understand their evolving needs and preferences and tailor marketing efforts accordingly.

3. Cluster 1: Mid to High Value

Similar to Cluster 1, customers in this cluster also demonstrate moderate to high recency, frequency, and monetary value. They represent another segment of valuable customers for the business.

Marketing Strategies:

Segment-Specific Campaigns: Design targeted marketing campaigns based on segment-specific preferences and behaviors to maximize engagement and conversion rates.

Exclusive Offers: Provide exclusive discounts or promotions to encourage repeat purchases and foster loyalty among this segment.

Referral Programs: Incentivize referrals by offering rewards or discounts for bringing in new customers, leveraging the existing customers' networks to expand the customer base.

4. Cluster 3: Mid to Low Value

Customers in this cluster exhibit moderate recency, frequency, and monetary value, but comparatively lower than those in Clusters 1 and 2. They contribute to the business's revenue but may require additional efforts to increase their engagement and spending.

Marketing Strategies:

Reactivation Campaigns: Implement targeted campaigns to re-engage customers who have shown a decline in activity or have become inactive to encourage them to make new purchases.

Win-back Offers: Offer special incentives or discounts to win back customers who have lapsed or made infrequent purchases.

Educational Content: Provide educational content or resources related to products or services to encourage them to explore more offerings and increase their spending.

6.2 Customer lifetime value findings

Customer Lifetime Value (CLV) serves as a cornerstone metric for businesses, encapsulating the total revenue anticipated from a customer over their relationship with the company. Through the meticulous analysis of RFM clusters and the nuanced understanding of customer segments, invaluable insights into CLV emerge, empowering businesses to formulate strategies aimed at optimizing this pivotal metric.

Segmenting customers based on RFM analysis reveals distinct cohorts with varying CLV potentials. Among these segments, "High-Value Customers" stand out as the most lucrative, characterized by their frequent transactions, substantial spending, and active engagement with the brand. Recognizing the significance of these customers, strategies emphasizing enhanced customer experience, loyalty cultivation, and personalized services become imperative. Investments in VIP treatment, exclusive loyalty programs, and tailored marketing initiatives can nurture these relationships, fostering repeat purchases and propelling long-term profitability.

Further analysis uncovers segments labeled as "Mid to High-Value," representing customers with moderate to high CLV potential. While not as affluent as the "High-Value" cohort, these customers significantly contribute to the company's revenue stream. Tailored marketing efforts, segmented promotions, and personalized recommendations serve as effective levers to

stimulate purchasing behavior and amplify CLV within these segments. Establishing robust connections through effective communication and value-added services bolsters customer retention, thereby elevating lifetime value over time.

Conversely, the "Mid to Low-Value" segment comprises customers with comparatively lower CLV potential. While individually less lucrative, these customers collectively form a valuable customer base. Implementing reactivation campaigns, enticing win-back offers, and providing educational content serve as strategic initiatives to reignite engagement and escalate spending among these customers. By tailoring offerings and communications to resonate with their preferences, businesses can effectively augment CLV and drive sustainable profitability.

In conclusion, RFM clustering offers profound insights into customer lifetime value, enabling businesses to orchestrate targeted strategies aimed at optimizing revenue generation, fortifying customer relationships, and fostering long-term profitability. Strategic management of CLV stands as a linchpin for sustainable growth and competitive advantage in the dynamic landscape of modern commerce.

6.3 Strengths and limitations of the study

6.3.1 Strengths

The effectiveness of this study is demonstrated through the adept utilization of clustering algorithms and evaluation metrics to delineate unique customer segments, thereby yielding significant insights into customer behavior and preferences. Employing a multitude of algorithms and metrics bolstered the integrity and dependability of the results. Moreover, scrutinizing feature importance and delving into cluster characteristics furnished an intricate comprehension of the data's underlying framework, empowering businesses to make well-informed decisions leveraging the study's insights.

The study also adopts a comprehensive approach to examining customer lifetime value (CLV) by analyzing various factors including revenues and repeat purchase probability etc. This holistic perspective offers a nuanced understanding of the implications associated with CLV. Moreover, the research provides practical tools, such as clustering techniques, which empower managers to measure CLV accurately and make well-informed marketing decisions, bridging the gap between theoretical concepts and practical applications. By integrating insights from marketing and data science, the study offers a multidisciplinary perspective on CLV, enriching the analysis and facilitating a deeper understanding of customer relationships. Additionally, the study identifies potential avenues for future research, such as exploring the interaction between CLV and brand loyalty, thereby contributing to the advancement of knowledge in this field. The practical implications of the findings are significant for businesses operating in the e-commerce sector, as a thorough understanding of CLV and its determinants enables organizations to develop targeted marketing strategies, optimize resource allocation, and enhance customer relationships, thereby fostering long-term profitability and growth.

6.3.2 Limitations

Although providing valuable insights into customer lifetime value (CLV), the study may encounter scope constraints due to its concentration on particular factors like revenues and repeat purchase behavior, potentially overlooking other influential elements such as customer

satisfaction, market competition, and broader economic trends. Data Availability Challenges may pose obstacles to the analysis, with limitations in data accessibility and quality potentially compromising the accuracy and applicability of findings. Additionally, the employment of mathematical models and clustering techniques might lead to Model Simplification, oversimplifying the complexities of real-world customer relationships and failing to fully capture nuances in consumer dynamics, such as non-linear relationships and heterogeneous behavior. Moreover, the study's findings may suffer from Limited Generalizability beyond the e-commerce sector, as unique factors specific to online retailing, such as digital marketing strategies and online shopping behaviors, may not align with traditional brick-and-mortar settings. External Validity Concerns also arise, with external factors like market fluctuations, technological advancements, and regulatory changes potentially impacting the relevance and applicability of research outcomes to real-world business scenarios, complicating the translation of findings into actionable insights.

7. RECOMMENDATIONS, CONCLUSION AND FUTURE WORK

The exploration of customer lifetime value (CLV) economics provides a fundamental shift in perspective, emphasizing the importance of cultivating ongoing relationships with customers rather than focusing solely on individual transactions. While optimization techniques often prioritize short-term marketing strategies, this paper underscores the need for a more comprehensive managerial analysis that considers long-term perspectives.

Our research contributes to this paradigm shift by providing practical tools for managers to measure CLV and make informed marketing decisions. However, the cases addressed in this paper represent only a subset of potential scenarios, and further investigation into more complex situations is warranted. Modeling CLV in hybrid cases, such as those involving variations in repeat purchase behavior or the impact of inflation, remains a challenge for future research.

Our study also highlights the interconnectedness between CLV and brand loyalty, emphasizing the importance of understanding both constructs in tandem. While previous research has distinguished between repeat purchase behavior and brand loyalty, our work suggests that a holistic understanding of customer relationships requires integrating these concepts.

To advance the field, future research should explore additional factors underlying CLV, such as customer satisfaction and competitive dynamics. By expanding our understanding of CLV and its implications for marketing strategy, we can better equip businesses to cultivate lasting customer relationships and drive sustainable growth.

REFERENCES

1. Agrawal, P. &. (2018). Digital supply chain management: An Overview. . In IOPConference Series: Materials Science and Engineering (Vol. 455, No. 1, p. 012074). Allahabad, India: IOP Publishing.
2. Alshamsi, A. (2022). Customer Churn prediction in ECommerce Sector. Rochester Institute of Technology.
3. Benfang Yang, J. L. (2022). Precise Marketing Strategy Optimization of E-Commerce Platform Based on KNN Clustering. Journal of Mathematics, vol. 2022.
4. Christy, A. J. (2021). RFM ranking–An effective approach to customer segmentation. . Journal of King Saud University-Computer and Information Sciences, 33(10), , 1251-1257.
5. Cvijović Jelenaa, K. S. (2017). Customer relationship management in banking industry: Modern approach. Industrija,Vol.45, No.3,.
6. Estrella-Ramón, A. M. (2014). A model of customer lifetime value and ex poste value-based segmentation. - International Campus of Excellence in Agrifood.
7. Fader, P. a. (2018). How to Project Customer Retention' Revisited: The Role of Duration Dependence. SSRN.
8. Hyunseok Hwang, T. J. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. Research Gate.
9. Jan Valendin, T. R. (2022). Customer base analysis with recurrent neural networks. International Journal of Research in Marketing, 988-1018.
10. Jasek, P. V. (2018). Modeling and Application of Customer Lifetime Value in Online Retail. Informatics, 5(1), 2. MDPI AG.
11. K. Bhade, V. G. (2018). A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization. 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), (pp. 1-6). Bengaluru, India.
12. Kanyinda, K. (2021). Application of DMAIC to improve energy consumption in a commercial building .
13. Koul, S. a. (2021). Customer segmentation techniques on e-commerce. 2021 International Conference on Advance Computing and Innovative Technologies in, 135-138.
14. Lee, Z.-J., Lee, C.-Y., Chang, L.-Y., & Sano, N. (2021). Clustering and Classification Based on Distributed Automatic Feature Engineering for Customer Segmentation. Symmetry 2021,13,1557.
15. Lemmens, A. C. (2007). Consumer confidence in Europe: United in diversity. International Journal of Research in Marketing, 113-127.
16. Lu, M. Y. (2018). Research on e-commerce customer repeat purchase behavior and purchase stickiness. . Nankai Business Review International, 9(3), 331-347.

17. Marisa, F. W. (2023). Potential Customer Analysis Using K-Means with Elbow Method. . JIKO (Jurnal Informatika dan Komputer), 7(2), , 307-312.

18. Mazzoni, C. A. (2005). Multidimensional segmentation of the cell phone market:preliminary results of an empirical survey . In Atti del IV congresso internazionale marketing trends-(). EAP: ESCP.

19. Mehrbakhsh Nilashi, B. M.-B. (2021). An analytical approach for big social data analysis for customer decision-making in eco-friendly hotels. Expert Systems with Applications.

20. Ming, Y. a. (2017). Customer segmentation based on RFM purchase tree. Rehabilitation Medicine, 306.

21. Muningsih, E. (2018). Komparasi metode clustering k-means dan k-medoids dengan model fuzzy RFM untuk pengelompokan pelanggan. Jurnal Sains dan Manajemen.

22. Nie, D. S. (2022). From data acquisition to validation: a complete workflow for predicting individual customer lifetime value. Journal of Marketing Analytics.

23. Oliver Dzobo, K. A. (2014). Multi-dimensional customer segmentation model for power system reliability-worth analysis. Research Gate.

24. Ozgen, P. (2017). A New Model for Customer Equity. Journal of Business Research - Turk.

25. P. P. Pramono, I. S. (2019). Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method. 16th International Conference on Service Systems and Service Management (ICSSSM), (pp. 1-5). Shenzhen, China.

26. Philip E Pfeifer, M. E. (2004). Customer Lifetime Value, Customer Profitability and the Treatment of Acquisition Spending. Journal of Managerial Issues 17(1):25.

27. Prasetyo, S. S. (2020). Penerapan fuzzy c-means kluster untuk segmentasi pelanggan e-commerce dengan metode recency frequency monetary (RFM). . Jurnal Gaussian, 9(4), , 421-433.

28. Radit Rahmadhan, M. W. (2022). Segmentation using Customers Lifetime Value: Hybrid Kmeans Clustering and Analytic Hierarchy Process. Journal of Information Systems Engineering and Business Intelligence.

29. Ridloah, S. (2016). A Qualitative Analysis Into The Strategic Priorities of the Indonesian Bank Industry. Jurnal Dinamika Manajemen.

30. Saghir, M. B. (2019). Churn prediction using neural network-based individual and ensemble models. 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), (pp. 634-639).

31. Shao, D. (2016). Analysis and prediction of insurance company's customer loss based on BP neural network.

32. Shi, Z. ,.-C. (2019). Spatiotemporal Data Clustering: A Survey of Methods. ISPRS International Journal of Geo-Information. 8. 112. 10.3390/ijgi8030112. .

33. Sien Chen, Y. H.-L. (2020). A two-stage machine learning approach for modeling customer lifetime value in the Chinese Airline Industry. 2020 AMA Summer Academic Conference. San Jose, USA.

34. Sommella, E. S. (2023). Digital Customer Relationship Management (e-CRM) in the Fashion Industry. In M. D. Brandstrup, The Garment Economy- Understanding History, Developing Business Models, and Leveraging Digital Technologies (pp. 287-205). Springer Texts in Business and Economics.

35. Su-yeon Kim, T. J. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert Systems with Applications.

36. T. Li, G. K. (2022). An Integrated Cluster Detection, Optimization, and Interpretation Approach for Financial Data. IEEE Transactions on Cybernetics, vol. 52, no. 12, 13848-13861.

37. Takhun Kim, D. K. (2022). Instant customer base analysis in the financial services sector. Expert Systems with Applications.

38. VanderPlas, J. (2019). Python Data Science Handbook. In J. VanderPlas, Python Data Science Handbook (pp. 433-479). Sebastopol, CA, USA: O'Reilly Media, Inc.

39. Wu, S. Y. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. IEEE Access.

40. Wu, Z. J. (2022). A PCA-AdaBoost model for E-commerce customer churn prediction. Ann Oper Res.

41. Xiahou, X. a. (2022). B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. Journal of Theoretical and Applied Electronic Commerce Research 17, no. 2, 458-475.

42. Yogesh Jadhav, D. P. (2020). Customer Segmentation and Buyer Targeting Approach. International Journal of Recent Technology and Engineering (IJRTE).

43. Yoseph, F. e. (2020). The Impact of Big Data Market Segmentation Using Data Mining and Clustering Techniques. Journal of Intelligent & Fuzzy Systems, 6159 – 6173.

44. Yue Li, X. C. (2021). Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm.

45. Zhang, D. (2015). Establishment and application of customer churn prediction model. . Beijing Institute of Technology.

APPENDICES

Appendix A: Turnitin Report