**Decoding Customer Behavior- A Machine Learning Based Approach to Understanding Customers**

**Table of Contents**

[Executive Summary 6](#_Toc123859447)

[1 Introduction 8](#_Toc123859448)

[1.1 Background Study: 8](#_Toc123859449)

[1.2 Aims and objectives of the project 9](#_Toc123859450)

[1.3 Route Map: 10](#_Toc123859451)

[2 Literature Review 11](#_Toc123859452)

[2.1 Understanding Customers 11](#_Toc123859453)

[2.1.1 Who is a “Customer”? 11](#_Toc123859454)

[2.1.2 Studying Customers 11](#_Toc123859455)

[2.1.3 Customers Importance 12](#_Toc123859456)

[2.1.4 Advantages of knowing your customers: 12](#_Toc123859457)

[2.1.5 Five-Stage Model of Decision-Making by Consumers 13](#_Toc123859458)

[2.1.6 Machine Learning Algorithms: 14](#_Toc123859459)

[2.2 K-Means Clustering Algorithm 15](#_Toc123859460)

[2.2.1 Introduction: 15](#_Toc123859461)

[2.2.2 Benefits of the K-Means Clustering Algorithm: 15](#_Toc123859462)

[2.2.3 Downfalls of the K-Means Clustering Algorithm: 16](#_Toc123859463)

[2.2.4 Application of the K-Means Clustering Algorithm in understanding customers: 16](#_Toc123859464)

[2.2.5 Research Gaps in the Current Approaches 16](#_Toc123859465)

[2.2.6 Conclusion 17](#_Toc123859466)

[2.3 Recency, Frequency, and Monetary Value Analysis 17](#_Toc123859467)

[2.3.1 Introduction: 17](#_Toc123859468)

[2.3.2 Benefits of using the RFM model: 18](#_Toc123859469)

[2.3.3 Downsides of the RFM model: 19](#_Toc123859470)

[2.3.4 Recent Developments: 19](#_Toc123859471)

[2.3.5 Contradictions or Gaps: 19](#_Toc123859472)

[2.3.6 Conclusion: 19](#_Toc123859473)

[2.4 Customer Lifetime Value 20](#_Toc123859474)

[2.4.1 Introduction: 20](#_Toc123859475)

[2.4.2 The Customer Lifetime Value Model 20](#_Toc123859476)

[2.4.3 Current Literature on the CLTV Model 20](#_Toc123859477)

[2.4.4 Contradictions or Gaps in the Current Approach 21](#_Toc123859478)

[2.4.5 Insight Gained from the Literature 21](#_Toc123859479)

[2.4.6 Conclusion 22](#_Toc123859480)

[2.5 Problem Statement 22](#_Toc123859481)

[2.6 Machine Learning Algorithms with RFM analysis: 23](#_Toc123859482)

[3 Methodology 25](#_Toc123859483)

[3.1 Software’s Used: 25](#_Toc123859484)

[3.2 Dataset Description: 25](#_Toc123859485)

[3.3 Data Preparation and Exploration 28](#_Toc123859486)

[3.3.1 Feature Additions 28](#_Toc123859487)

[3.3.2 Descriptive Statistics of data 29](#_Toc123859488)

[3.3.3 Handling Missing Data 29](#_Toc123859489)

[3.3.4 Handling Duplicate values 30](#_Toc123859490)

[3.4 RFM Analysis: 33](#_Toc123859491)

[3.4.1 Calculating Recency 34](#_Toc123859492)

[3.4.2 Calculating Frequency: 36](#_Toc123859493)

[3.4.3 Calculating Monetary Value: 38](#_Toc123859494)

[3.4.4 Assigning RFM scores to customers: 42](#_Toc123859495)

[3.4.5 Calculating a total RFM score and assigning segment labels: 44](#_Toc123859496)

[3.4.6 Labels and their description: 45](#_Toc123859497)

[3.5 K-Means Clustering: 47](#_Toc123859498)

[3.5.1 Data Normalization: 47](#_Toc123859499)

[3.5.1.1 Logarithmic Method: 47](#_Toc123859500)

[3.5.1.2 Power Transformation (Yeo-Johnson) Method: 48](#_Toc123859501)

[3.5.2 Determining number of clusters(k): 50](#_Toc123859502)

[3.5.3 The Elbow Method: 50](#_Toc123859503)

[3.5.4 Producing the model: 52](#_Toc123859504)

[3.6 Customer Lifetime Value: 57](#_Toc123859505)

[3.6.1 Preparing data for CLTV prediction: 57](#_Toc123859506)

[3.6.1.1 Beta Geometric / Negative Binomial (BNB) distribution model 58](#_Toc123859507)

[3.6.1.2 Gamma Gamma Model: 59](#_Toc123859508)

[3.6.2 Interpreting the results: 60](#_Toc123859509)

[3.6.3 Recommended Marketing Strategies: 62](#_Toc123859510)

[4 Conclusion 64](#_Toc123859511)

[5 Recommendation 66](#_Toc123859512)

[6 References 68](#_Toc123859513)

**List of Tables**

Table 1: Objective of the study 9

Table 2: Route Map of the study 10

Table 3: Data Description 25

**List of Figures**

Figure 1: Five-stage decision-making process 13

Figure 2: Shape of the raw data 26

Figure 3: Count of Invoices by Countries 27

Figure 4: Description of Null values 28

Figure 5: Data Pipeline illustration 28

Figure 6: Shape of UK retail transaction data 29

Figure 7: Descriptive statistics of data 29

Figure 8: Count of missing values with missing percentages 30

Figure 9: Shape of data without missing values 30

Figure 10: Code snippet verifying missing values 31

Figure 11: bar graph with top ten most purchase products 31

Figure 12: bar graph with least ten most purchase products 32

Figure 13: bar graph with top ten most demanded products by quantity 32

Figure 14: bar graph with least ten most demanded products by quantity 33

Figure 15: RFM description chart 33

Figure 16: Customer recency, data frame image 34

Figure 17: Top 5 Customers with high recency, data frame image 35

Figure 18: Histogram representing recency distribution 35

Figure 19: Customers frequency, data frame image 36

Figure 12: Top 5 Customers with high frequency, data frame image 36

Figure 21: Bottom 5 Customers with least frequency, data frame image 37

Figure 22: Histogram representing Frequency distribution 37

Figure 23: Top 5 Customers with high Monetary value, data frame image 38

Figure 24: Bottom 5 Customers with Least Monetary value, data frame image 38

Figure 25: Histogram representing Monetary value distribution 39

Figure 26: RFM values for each customer, data frame image 40

Figure 27: Descriptive statistics of RFM values 41

Figure 28: Pie chart showing proportions of recency scores 42

Figure 29: Pie chart showing proportions of Frequency scores 43

Figure 30: Pie chart showing proportions of Monetary scores 44

Figure 31: List of labels with assigned values 45

Figure 32: Customer category with number of customers in each 46

Figure 33: Bar graph representing Customer category with number of customers in each 46

Figure 34: Distribution of logarithmically transformed RFM variables 48

Figure 35: Result of Skewness test of logarithmically transformed RFM variables 48

Figure 36: Distribution of power transformed RFM variables 49

Figure 37: Result of Skewness test of power transformed RFM variables 49

Figure 38: Result of Hopkins statistics test 50

Figure 39: Result of elbow method to find optimal clusters(k) 51

Figure 40: Result of elbow method to find optimal clusters(k) with distortion score 51

Figure 41: Result of power transformed RFM variables 52

Figure 42: Pie chart showing proportions of clusters Ids 53

Figure 43: Box plot representing five number summaries of RFM variables 53

Figure 44: RFM scores with assigned labels 54

Figure 45: Customer category with number of customers in each 54

Figure 46: Bar graph representing Customer category with number of customers in each 55

Figure 47: Scatter plot representing clusters and their respective centroids 55

Figure 48: Scatter plot representing all three clusters and their respective centroids 56

Figure 49: R F M and T values of customers 57

Figure 50: Correlation result between monetary and recency values 58

Figure 51: Probability density curve of customers recency and frequency 58

Figure 52: R F M and T values with expected number of purchases 59

Figure 53: data frame after expected average order value calculation 60

Figure 54: list of top five customers with highest predicted CLTV values 60

Figure 55: list of bottoms five customers with least predicted CLTV values 61

Executive Summary

In the modern world, customer behavior towards the brand is one of the most crucial factors in determining the success of a business. To understand customer behavior, businesses have used various tools and techniques. Machine Learning algorithms are one such tool that can be used to analyze customer data and make informed decisions. The use of Machine Learning algorithms has become increasingly popular due to their ability to process substantial amounts of data quickly and accurately. In addition to Machine Learning algorithms, RFM analysis and CLTV are also important tools for understanding customer behavior and knowing their value. RFM stands for Recency, Frequency, and Monetary value. It is used to analyze customer behavior and identify potential opportunities for upselling. CLTV stands for Customer Lifetime Value and is used to calculate the value of a customer over their lifetime.

The goal is to understand the impact of Machine Learning algorithms and other models in understanding customer behavior. The study is conducted on online retail gift shops customer data sourced from machine learning repository, which includes data from a customer management system. The data includes eight attributes which are Invoice, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, Country. An extensive literature survey is conducted to research the different methods or machine learning algorithms used to understand customers behavior. After going through a plethora of research papers and academic compilations of the scholars in this field, a few particularly important methodologies are used in this study.

Initially the data is processed through python and cleaned to get rid of the missing values, null values, duplicates, outliers, and date time irregularities. Later exploratory descriptive analysis (EDA) is performed to understand the data. Graphs are plotted to understand the variables and the relationship between them.

After getting well versed with the dataset, customer segmentation is carried out on the bases of RFM metrics. The customers are segmented on the bases of RFM (recency of purchase, frequency of purchase, and Monetary value they bring to the table). A K-Means Clustering algorithm is also performed to segment customers and later the customers lifetime value is also predicted.

The research findings suggest that K-means clustering is an effective tool for clustering data points. K-means clustering is a popular unsupervised learning technique used to group data points into distinct clusters or groups. However, when combined with RFM metrics, it is even more effective. The research found that RFM metrics can help to identify customer segments and target them more effectively, leading to increased customer loyalty and better marketing performance. Furthermore, the research found that K-means clustering is able to accurately identify customer segments and group them according to their preferences and buying behaviors.

Buying behavior is the process of studying and analyzing how consumers make their buying decisions. It is important to understand buying behavior because it can help businesses better understand their target customers and develop strategies to reach them. Knowing who your target customers are and what motivates them to make a purchase can help you create tailored marketing messages that will resonate with them. This is significant for companies in customer segmentation because it allows them to tailor their products and services to the specific needs of their target customers. Understanding buying behavior can also help companies better understand their competition and develop strategies to gain a competitive advantage.

This research provides valuable insights for marketing professionals who are looking for ways to better understand customer behavior and preferences. The findings of this research can be used to help companies identify customer segments, target them more effectively, and increase customer loyalty and marketing performance.

1 Introduction

**1.1 Background Study:**

A motivated consumer is prepared to take action. The motivated customer's view of the circumstances will affect how they choose to respond. In the same motivated state and objective circumstance, two persons may behave substantially differently because they have distinct perspectives on the event (Ghazzawi, et al., 2016). People are constantly faced with a vast array of decision-making inputs, according to (Akpoyomare, Adeosun, & Ganiyu, 2013). Consider the influence of diverse types of advertising on consumer choice; on average, consumers are exposed to about 1,500 advertisements daily. The majority of the stimuli will be screened out since it is hard for a consumer to handle all of the stimulus. The real challenge is to explain which decision-making stimuli will be screened out (Souri, 2017).

(Roper & Vecera, 2013) stated that on the other side, selective exposure necessitates more effort on the part of marketers to capture consumers' attention in the marketplace. He said that the majority of those who aren't interested in the goods won't get their message. If a message doesn't stand out from the surrounding sea of decision-making, even consumers who are in the market might not see it.

Therefore, understanding your customer base is essential to a business success. There are a variety of historical methods businesspeople have used to try to get to know their customers better. One widespread approach is customer surveys. This entails distributing a survey to a sizable consumer base, then reviewing and analyzing the responses. This can be a helpful way to get feedback on a new product or service or to gauge customer satisfaction levels. Another common method is customer focus groups. This involves bringing a group of customers together in a room and asking them questions about their experiences and opinions. This can be a more interactive way to get feedback and can also help businesses spot trends. (Filip & Voinea, 2012)

Another way businesses can also try to understand their customers by talking to them directly. This can be done through things like one-on-one interviews or even focus groups. This can be a wonderful way to get personal feedback and to really get to know your customers. Finally, businesses have tried to understand their customers is by analyzing customer data. This can include things like purchase history, demographic information, and even social media activity. By analyzing this data, businesses can get a better sense of who their customers are and what they want. (Khan, 2012)

But when a company wants to tailor marketing strategies and promotions for customers, they need to know whether the customers are loyal to the brand and would they provide the required value to the company in a long run. To do so, companies need to check customers loyalty. But the demographic data would barely provide any factors to cross-check this thing. That is when RFM metrics comes in handy for the companies. With this data, companies can build various marketing campaigns for each segment of customers using machine learning methods. (Mitrović, Baesens, Baesens, Lemahieu, & Weerdt, 2019)

In this report we explore how using RFM metrics for customer segmentation with K-means clustering algorithm can provide better results that the usual clustering with demographic data. The research focuses on what RFM method is, how effective it is when used along with machine learning algorithm, and advantages of the RFM metric data. The research includes the related works regarding the RFM, K-means clustering and Customer Lifetime Value analysis. The research majorly focuses on how companies can tailor the best marketing plan according to their customer segments based on their recency, frequency, and monetary value of their purchase.

**1.2 Aims and objectives of the project**

The aim of the current project is:

1) to expand on the “An Analysis why customers are so important and how marketers go about in understanding their decisions” work by Benjamin Musumali (2019) by incorporating into the analysis with more detailed machine learning algorithms used currently.

2) to figure out the implications of understanding the customers and the related benefits to the business

In this study, we particularly focus on the below objectives to get to the goal.

|  |  |
| --- | --- |
| **Objectives:** |  |
| 1 | To review the approaches previously used to comprehend customers through market research. |
| 2 | To evaluate the current techniques for customer understanding and implement them on a dataset to analyze the scope. |
| 3 | To evaluate these outcomes to figure out whether the machine learning-related techniques are effective with RFM metrics. |
| 4 | To determine Customer Lifetime Value based on RFM metrics to estimates customers value to the company. |

*Table 1: Objective of the study*

The below mentioned route map is followed throughout the study the achieve the above objective.

**1.3 Route Map:**

|  |  |
| --- | --- |
|  |  |
| Literature Review | Thorough summary of the research published on a customer importance to the business. An overview of the main ideas and findings of the research organized chronologically. |
| Main Methodology | Analyzing the data to understand customer behavior using various machine learning method and then evaluating the effectiveness of that to the business growth. |
| Conclusion | Closing summary of the research project. |
| Recommendation | Recommending related items to the customer based on RFM analysis. |

*Table 2: Route Map of the study*

2 Literature Review

Here in the literature review, we first assess the benefits of knowing your consumers, the importance of understanding their needs, and the decision-making process of customers. We then review the current machine learning algorithms used for customer understanding. The existing literature on the K-means clustering algorithm. We discussed the advantages and disadvantages of the algorithm, its applications, and the research gaps in the current approaches. The well-known research articles are reviewed for the investigation of the RFM analysis literature. We will examine the meanings of RFM analysis, as well as its applications, advantages, and other relevant subjects. Finally, we review the additional analyses done using the RFM metrics, the Customer Lifetime Value Model, we examine the meanings of CLTV model, its applications, advantages, and other relevant subjects

**2.1 Understanding Customers**

**2.1.1 Who is a “Customer”?**

An individual or corporation that purchases products or services from another company is known as a customer. All these years, businesses usually have believed in the quote “the customer is always right.” This is because according to them if a customer is happy with the services, he/she is being provided with they likely stick to the services or the company. This indeed would help n business growth and more profit. To achieve this, businesses closely check their customer feedback and work on the methods to improve them. To do so, they categorize customers into few common categories. (Shanker, 2012)

(Aschettino, Birnbaum, Crocker, Grebner, & McNicholas, 2013) contends that consumers from outside the company are not involved in daily operations but do participate in the ultimate product purchasing. While internal customers take part in the business operations are strongly associated with the business.

**2.1.2 Studying Customers**

The (Examples of Marketing Strategies Used to Sell a Product, n.d.)" said that a business requires at least marketing techniques to develop product or service awareness in order to flourish, the product or service it supplies must be known in the community and have contact with its consumers readily available."

From the aforementioned, it can be seen that without marketing, Apple incorporation company's prospective clients could never be aware of its business products, and its firm might not be given the chance to develop and prosper. If the Apple incorporation firm uses marketing to advertise its goods, solutions, and business, it gives it a chance to be found by potential clients.

Businesses frequently conduct customer profiling studies to optimize their marketing strategies and customize their products to appeal to the widest possible audience. Customers are frequently categorized based on factors including age, gender, color, ethnicity, economic level, and region, which may all help firms create a profile of the "perfect client" or "customer persona." To boost traffic, businesses may use this information to strengthen their existing customer connections and seek new consumer markets.

**2.1.3 Customers Importance**

Companies are under enormous competitive pressure almost without exception. In other circumstances, such as the banking industry, this is further exacerbated by ongoing regulation change. Customers-facing staff and processes are being stretched to the breaking point by a potent confluence of micro and macroeconomic conditions, geographical transparency and increased reach provided by e-business, mergers and acquisitions, competitor activity, and localized market conditions. (Dominici & Guzzo, 2010)

Customers are crucial for the reasons listed below.

Customers are so crucial that schools and institutions provide courses on consumer behavior that focus on understanding their behavioral patterns, decisions, and quirks. They emphasize the motivations behind consumers' decisions to purchase and utilize products and services as well as how this impacts businesses and economies. Businesses can produce products and services that fulfil needs and wants, maintain consumers for repeat business, and develop effective marketing and advertising campaigns when they have a thorough understanding of their target market. (McCormick & Livett, 2012)

**2.1.4 Advantages of knowing your customers:**

There are many advantages to knowing your customers:

*First, accessibility to customers increases brand loyalty for the business.*

(Reisenwitz & Gupta, 2016) defined brand loyalty as a customer's decision to consistently buy a product made by a foe. As an instance, while some customers will always buy Tesla produced in China, others will always buy Mercedes manufactured in Germany. Better customer interactions result in client loyalty and repeat business for the firm.

*Second, by knowing your clients' requirements and expectations, you can provide better customer service, which in turn enhances the reputation of your business.*

The reputation of a corporation typically determines its success. Customer service, according to (Rutkauskas & Paulavicienė, 2015), improves a company's capacity to establish brand recognition or product recall. When a company outperforms the general public's elevated expectations, its reputation is enhanced, and customers become more trustworthy because of their larger purchasing power. As your reputation grows, your business expands and your sales increase. Participation in community activities, effective internal and external communication, and the production or promotion of high-quality goods and services to draw in both present and new customers are all ways to build your company's reputation.

*Finally, you may obtain crucial information about your clients that will enable you to better run your company, which will in turn encourage clients to support the growth of the economy.*

According to (Gordon|Kruse, 2017), the aggregate production—also known as the economy's yearly total output of goods and services—is the main indicator of the economy's health. Gross domestic product (GDP), the sum market value of all finished goods and services produced annually, is a measure of aggregate output. Consumer goods and services account for most of the consumers' disposable income (DI). He said that large import and export expenditures raise aggregate demand (AD), which supports economic expansion.

**2.1.5 Five-Stage Model of Decision-Making by Consumers**

Consumers go through a five-stage decision-making process when they purchase a product.

SOURCE: HARVARD BUSSINESS REVIEW, MAY-JUNE 2001, PP.53-54.

*Figure 1: Five-stage decision-making process*

After understanding the role of customers in businesses success and knowing their decision-making process, we will now look at the current methods and tools used to analyze customers data to understand them. The most common and popular method used all across the business in machine learning algorithms.

**2.1.6 Machine Learning Algorithms:**

Machine learning algorithms are commonly used to understand customer behavior. Businesses develop models that predict future customer behavior by analyzing customer data. This allows companies to target specific customers with personalized marketing messages and improve customer retention. Additionally, machine learning can be used to find trends in customer behavior, which can help businesses adapt their products and services to meet customer needs better.

In a research paper published by researcher (Watkins, et al., 2015), the author examined the use of machine learning algorithms to better understand customer behavior. The paper talks about some of the most used machine learning algorithms, such as support vector machines, decision trees, random forests, clustering, and neural networks, and explains how these algorithms can be used to predict customer behavior. Watkins identified that these algorithms can be used to identify customer segments, create personalized recommendations, and improve customer service. In addition, Watkins points out that there are challenges associated with using machine learning algorithms in customer understanding, such as data privacy and ethical considerations.

Machine learning algorithms are becoming increasingly popular in understanding customer behavior. By analyzing past customer behavior, these algorithms can predict future customer behavior with a high degree of accuracy. This is extremely valuable for businesses, as it can allow them to target their marketing and sales efforts more effectively. Additionally, machine learning algorithms can help businesses to find potential issues and opportunities that they may not have been aware of previously. Overall, the impact of machine learning algorithms on understanding customer behavior is extremely positive and is likely to continue to grow in the future.

K-Means clustering is one of the most popular unsupervised machine learning algorithms for customer understanding due to its ease of implementation and its ability to quickly identify clusters in large datasets (Lai, 1995). Specifically, K-Means is particularly effective when segmenting customers based on their behavior, such as purchase history, demographics, usage patterns, and other factors. Furthermore, the algorithm is computationally efficient and can be applied to large datasets, enabling organizations to quickly identify customer segments (Khalid & Herbert-Hansen, 2018). Additionally, the algorithm is simple to interpret, allowing marketers to gain insight into their customer segments (Maechler, Rousseeuw, Struyf, & Hubert, 2015). Therefore, K-Means clustering is a powerful tool for customer understanding and segmentation and is widely used by businesses for this purpose.

Hence for our research we have also used K-means Clustering Algorithm

**2.2 K-Means Clustering Algorithm**

**2.2.1 Introduction:**

Cluster analysis is a method of data analysis which aims to group data points into distinct clusters or groups. It is one of the most used techniques in data mining and has been widely used for various data-driven tasks, such as image segmentation, pattern recognition, market segmentation, etc. (Xia, Karimi, & Meng, 2017).

Hierarchical clustering and partitioned clustering are the two basic categories under which clustering methods may be divided. While partitioning clustering algorithms split the data set into a predetermined number of clusters, hierarchical clustering algorithms construct a hierarchy of clusters. K-means clustering, hierarchical clustering, and density-based clustering are common clustering techniques. (Machap & Abdullah, 2020)

K-means clustering is one of the most popular and widely used clustering algorithms. It is a type of unsupervised learning algorithm that partitions a given dataset into K distinct non-overlapping clusters based on the similarity between the data points belonging to the same cluster (Weidema, et al., 2020). The K-means algorithm has several advantages over other clustering algorithms, such as being computationally efficient and requiring minimal parameter tuning (Kho, Kasihmuddin, Mansor, & Sathasivam, 2020).

However, it is important to note that the K-means algorithm does have certain limitations. For example, it does not work well with datasets that have outliers or datasets with non-spherical clusters. In addition, it is sensitive to the first conditions and may converge to the local optima (Dalhatu & Sim, 2016).

**2.2.2 Benefits of the K-Means Clustering Algorithm:**

The K-means algorithm has several advantages over other clustering algorithms. The main advantage of the K-means algorithm is its computational efficiency. It is a fast and iterative algorithm that can be implemented easily. It is also computationally scalable and can be used for datasets with large numbers of instances (Parracho, Melo-Gonçalves, & Rocha, 2016).

In addition, the K-means algorithm requires minimal parameter tuning and is relatively easy to understand and implement. It can be used for any number of clusters and does not require any prior knowledge of the dataset. Moreover, it produces clusters of high quality and is able to find the clusters in a dataset (Dong, et al., 2018).

**2.2.3 Downfalls of the K-Means Clustering Algorithm:**

Despite its advantages, the K-means algorithm has some limitations. One of the main disadvantages of the algorithm is that it is sensitive to the first conditions. If the initial conditions are not properly set, the algorithm may converge to the local optima and do not find the global optima (Kindhi, Sardjono, Purnomo, & Verkerke, 2019).

In addition, the K-means algorithm does not work well with datasets that have outliers or datasets with non-spherical clusters. It assumes that the clusters are spherical and so it is not suitable for datasets with non-spherical clusters (Kindhi, Sardjono, Purnomo, & Verkerke, 2019).

Furthermore, the K-means algorithm is not suitable for datasets with a large number of dimensions as it may suffer from the curse of dimensionality. This is because the algorithm requires a large amount of computation for each iteration and so it becomes inefficient for datasets with many dimensions (Fernández-Martínez & Fernández-Muñiz, 2020)

**2.2.4 Application of the K-Means Clustering Algorithm in understanding customers:**

K-means clustering is a powerful tool for customer segmentation and understanding. The algorithm is based on the assumption that the data points in a dataset can be divided into distinct clusters, or subgroups, based on their features. By dividing customers into clusters based on their features, marketers can gain insight into customer preferences, behaviors, and needs. This can be used to develop targeted marketing campaigns and tailor products and services to better meet customer needs. Additionally, k-means clustering can be used to identify potential customer segments that may have been previously unknown or overlooked. (Chatterjee & Yuen, 2019)

**2.2.5 Research Gaps in the Current Approaches**

Despite its popularity, the K-means algorithm has some shortcomings and there are some research gaps in the current approaches. One of the main research gaps is the lack of robustness of the algorithm to outliers and non-spherical clusters. The algorithm assumes that the clusters are spherical and so it is not suitable for datasets with non-spherical clusters (Suresh & Paul, 2010)

In addition, the algorithm is sensitive to the initial conditions and may converge to the local optima. This limits the accuracy of the algorithm and its ability to find the global optima (Kindhi, Sardjono, Purnomo, & Verkerke, 2019).

Furthermore, the algorithm is not suitable for datasets with a large number of dimensions as it may suffer from the curse of dimensionality. This is because the algorithm requires a large amount of computation for each iteration and so it becomes inefficient for datasets with a large number of dimensions (Fernández-Martínez & Fernández-Muñiz, 2020).

**2.2.6 Conclusion**

In this paper, we reviewed the existing literature on the K-means clustering algorithm. We discussed the advantages and disadvantages of the algorithm, its applications, and the research gaps in the current approaches. We also provided our insights into the algorithm and its potential to be used in various data analysis tasks.

Overall, the K-means algorithm is a fast and iterative algorithm that can be implemented easily and requires minimal parameter tuning. It can be used for any number of clusters and does not require any prior knowledge of the dataset. However, the algorithm has some limitations and there are some research gaps in the current approaches.

Even though machine learning algorithms work best for customer segmentation, there are few other models that can help in bettering the results. These models leverage machine learning algorithm to identify customer segments, analyze customer behavior, and predict customer churn. They can also be used to create personalized marketing campaigns, understand customer lifetime value, and identify key customer segments. By providing insights into customer behavior, these models can help businesses develop more effective marketing strategies and optimize their customer experience. Therefore, we have considered the two most important ones based on their popularity and their effectiveness.

**2.3 Recency, Frequency, and Monetary Value Analysis**

**2.3.1 Introduction:**

These days, businesses largely rely on their long-term, devoted clientele. Companies must keep and sustain a positive connection with their clients to advance this. It forces them to comprehend the trends in and behavior of their clients to flourish as a firm. They have been employing a variety of techniques, with RFM analysis being one of the more well-liked ones. Recency, Frequency, and Monetary (RFM) analysis is a technique for segmenting customers those entails looking at how recently the consumer made purchases, how often the client made purchases, and how much money the customer spent (Anitha & Patil, 2019).

The purpose of RFM analysis is to figure out which consumers are most likely to make future purchases and which ones are least likely to do so. Using this technique, businesses may focus their marketing efforts on the clients who are most likely to make future purchases from them. Recency is widely considered to be the most significant RFM metric. However, earlier data write down that RFM values are more likely to be firm-specific and depending on the characteristics of the products (Wu, Lin, & Liu, 2014).

According to (Hu & Yeh, 2014), customers with lower recency and a high frequency tend to have low purchasing potential in comparison to other customers. And there are significant differences between groups across recency and frequency.

There has been a lot of research conducted on the effectiveness of RFM analysis. In a study conducted by (Asllani & Halstead, 2015), the authors investigated the effectiveness of RFM analysis in customer segmentation and retention. The authors found that RFM analysis is an effective tool for customer segmentation and retention. They concluded that it is a useful tool for businesses to find customers who are most likely to purchase again, as well as customers who have the potential to become long term, high value customers.

In another study conducted by (Chan, Hwang, & Wu, 2016), the authors analyzed the effectiveness of RFM analysis in a retail setting. The authors found that RFM analysis is effective in predicting customer behavior and can be used to find customers who have the potential to become long term, high value customers. The authors concluded that the use of RFM analysis can help businesses to effectively name and analyze customer behavior and develop strategies to increase customer retention.

In a study conducted by (Aggelis & Christodoulakis, 2005), the authors analyzed the effects of RFM analysis on customer loyalty. The authors found that customers who had higher recency and frequency scores were more likely to be loyal than those with lower scores. The authors also found that customers who had higher monetary scores were more likely to be loyal, but this was not as strong as the effect of recency and frequency. The authors concluded that RFM analysis can be used to name customers who are more likely to be loyal and recommend strategies to increase customer loyalty.

**2.3.2 Benefits of using the RFM model:**

There are several reasons why the RFM model is popular in understanding customers and segmentation:

1. RFM is affordable for obtaining crucial consumer behavior research and makes it simple to quantify customer behavior

2. RFM is quite helpful in predicting responses and can increase a company's earnings in the near future

3. The purchase behavior can be condensed into a relatively small number of variables, modelling using RFM variables is particularly efficient.

4. RFM variables are not acquired from the aggregate level data in the demographic databases but rather from an internal database that has customer-specific information about transaction history. RFM is hence more useful for client targeting.

5. RFM is a well-known way to assess the quality of a customer connection since it is good at naming valued clients

**2.3.3 Downsides of the RFM model:**

1. One of the main drawbacks is that the method is heavily reliant on data. If the data is incomplete or inaccurate, the results of the analysis may be skewed.

2. The method does not account for other factors that may affect customer behavior, such as social media interactions or promotional activities.

3. RFM analysis is that it does not provide a comprehensive view of customer behavior. For example, the method does not take into account customer preferences or the types of products they are purchasing.

As a result, it may not be able to accurately predict customer loyalty or spending habits.

**2.3.4 Recent Developments:**

Recently, there have been some advancements in RFM analysis. For example, researchers have developed methods to incorporate other data sources into RFM analysis, such as social media data or promotional activities. This allows companies to get a more comprehensive view of customer behavior.

Additionally, researchers have developed methods to combine RFM analysis with other techniques, such as machine learning or artificial intelligence. This allows companies to gain deeper insights into customer behavior and develop more accurate segmentation models.

**2.3.5 Contradictions or Gaps:**

Despite the success of RFM analysis in predicting customer loyalty and profitability, there are some contradictions or gaps in the current approach. For example, some studies have found that the recency component of RFM analysis is not a reliable predictor of customer loyalty (Song, Zhao, Haihong, & Ou, 2016). This suggests that businesses should be cautious when using RFM analysis to target customers, as the recency component may not be an accurate predictor of customer loyalty.

In addition, some studies have found that RFM analysis is not always effective in predicting customer profitability (Miglautsch, 2002). This suggests that businesses should not rely solely on RFM analysis when targeting customers, as it may not always be effective in predicting customer profitability.

**2.3.6 Conclusion:**

The literature on RFM analysis provides several insights into the effectiveness of the approach. It is clear that RFM analysis can be used to identify and target high-value customers, and that it can be used to inform the customer segmentation process. Additionally, there are some contradictions and gaps in the current approach, including a lack of research on the impact of RFM analysis on customer loyalty.

Overall, RFM analysis is a valuable analytical tool that can be used to identify and target high-value customers. By leveraging the insights provided by the literature, companies can use RFM analysis to maximize their profits and improve their customer loyalty.

**2.4 Customer Lifetime Value**

**2.4.1 Introduction:**

The Customer Lifetime Value (CLTV) is a measure of a customer’s total expected value to a business throughout their entire lifespan as a customer. It is a metric used to figure out the profitability of a customer, as it quantifies the value that each customer brings to the business. It is important for companies to understand the CLTV of their customers, as it allows them to better understand the long-term value of their customers, and it also helps them to distribute their marketing and customer retention efforts in a more effective manner. In this paper, we will review the current approach to the CLTV method and discuss any contradictions or gaps in the current approach. We will also discuss any insights that can be gained from the literature on the CLTV method. (Singh & Jain, 2013)

**2.4.2 The Customer Lifetime Value Model**

The CLTV model is used to measure the lifetime value of a customer to a business. It is calculated by taking the customer’s average purchase value and multiplying it by the customer’s estimated lifespan as a customer. The CLTV model can be used to figure out the customer’s total expected value to a business over a given period of time. The customer’s estimated lifespan is based on the customer’s past purchase behavior and other factors such as customer loyalty and customer satisfaction.

The CLTV model is a useful tool for businesses as it helps them to better understand the value of their customers, and it also helps them to make more informed decisions about marketing and customer retention efforts. The CLTV model can be used to figure out the profitability of a customer, as it takes into account the customer’s expected lifetime value to the business. (Benoit & Poel, 2009)

**2.4.3 Current Literature on the CLTV Model**

There is a considerable amount of literature available on the CLTV model. Some of the authors who have written extensively on the CLTV model.

(Ekinci, Ülengin, Uray, & Ülengin, 2014) this paper discusses the importance of customer lifetime value as a measure of customer profitability. The authors argue that the CLTV model is a valuable tool for businesses, as it allows them to better understand the long-term value of their customers. He also states that the CLTV model can be used to make more informed decisions regarding marketing and customer retention efforts.

(Khajvand, Zolfaghar, Ashoori, & Alizadeh, 2011) this paper discusses the benefits of using the CLTV model as a tool for strategic planning. Authors states that the CLTV model can be used to measure the profitability of a customer, and it also allows businesses to make more informed decisions regarding marketing and customer retention efforts. Authors argues that the CLTV model can be used to identify high-value customers, and it can also be used to measure the success of customer retention efforts.

(Cheng, Chiu, Cheng, & Wu, 2012) this paper provides a comprehensive review of the CLTV model. Authors discusses the benefits of the CLTV model and its applications in a variety of industries. Authors also says that the CLTV model can be used to measure the lifetime value of a customer to a business, and it can also be used to make more informed decisions about marketing and customer retention efforts.

**2.4.4 Contradictions or Gaps in the Current Approach**

Despite the extensive amount of literature on the CLTV model, there are still some contradictions or gaps in the current approach. For example, (Ekinci, Ülengin, Uray, & Ülengin, 2014) argues that the CLTV model is a valuable tool for businesses, as it allows them to better understand the long-term value of their customers. However, (Januszewski, 2011) argues that the CLTV model can be used to measure the profitability of a customer but does not address how it can be used to understand the long-term value of a customer.

Another contradiction or gap in the current approach is that while both (Ekinci, Ülengin, Uray, & Ülengin, 2014) and (Januszewski, 2011) agree that the CLTV model can be used to identify high-value customers, does not address how the CLTV model can be used to do this. This suggests that there is a gap in the literature about the use of the CLTV model to identify high-value customers. (Singh & Jain, 2013)

**2.4.5 Insight Gained from the Literature**

The literature on the CLTV model provides several insights into the use of the model. Firstly, it is clear that the CLTV model is a valuable tool for businesses, as it allows them to better understand the long-term value of their customers. Secondly, the CLTV model can be used to measure the profitability of a customer, and it can also be used to make more informed decisions regarding marketing and customer retention efforts. Thirdly, the CLTV model can be used to name high-value customers, and it can also be used to measure the success of customer retention efforts.

**2.4.6 Conclusion**

In conclusion, the literature on the CLTV model provides an overview of the current approach to the CLTV method. It is clear that the CLTV model is a valuable tool for businesses, as it allows them to better understand the long-term value of their customers. Additionally, the CLTV model can be used to measure the profitability of a customer, and it can also be used to make more informed decisions regarding marketing and customer retention efforts. While there are some contradictions or gaps in the current literature, the literature provides several insights into the use of the CLTV model. (Kaul, 2017)

**2.5 Problem Statement**

With all the above detailed literature review presented, we can see some research gaps that needs to be addressed. First and foremost, K means clustering algorithm, is not suitable for datasets with a large number of dimensions as it may suffer from the curse of dimensionality. (Fernández-Martínez & Fernández-Muñiz, 2020). This is because the algorithm requires a large amount of computation for each iteration and so it becomes inefficient for datasets with a large number of dimensions. This is the major problem when working with the demographic data is inaccurate demographic information, as many people may not provide completely accurate information.

In a paper published by, (The Disadvantages of Target Marketing, n.d.), One of the main issues highlighted by the authors was demographic data can be biased and contain errors due to the way it is collected and reported. For instance, the authors noted that if data is collected through self-reporting, the respondent may not accurately report their own information, leading to inaccuracies in the data. Additionally, the authors noted that demographic data can be misused or misinterpreted due to the way it is presented and analyzed. For example, a researcher may take a small sample size of data and draw conclusions about a larger population, which may not be accurate.

Considering all the limitations and drawbacks of demographic data, Recency, Frequency and Monetary value metric data can be quite useful. Any simple transactional data can be converted into RFM metric data and the analysis and be processed. According to the research conducted by (Sarvari, Ustundag, & Takci, 2016). RFM metric data has several advantages over traditional demographic data when it comes to segmenting online customers. Specifically, RFM data is much more granular and dynamic in nature, allowing for a greater level of segmentation and more accurate predictions of customer behavior.

Furthermore, RFM data is more easily accessible and less expensive to obtain than demographic data, making it a more cost-effective solution for segmenting customers and predicting their future behavior. Additionally, RFM data is timelier and can be used to immediately identify and react to changes in customer behavior. All these advantages make RFM data a more effective and efficient tool for segmenting customers and predicting their future behavior than traditional demographic data. (Christy, Umamakeswari, Priyatharsini, & Neyaa, 2018)

Therefore, looking at all the shortcoming and gaps in the research, this paper focuses on effectiveness of methods and machine learning algorithm in understanding customers. The paper evaluates how customer segmentation can be made more effective when used along with RFM metric data.

**2.6 Machine Learning Algorithms with RFM analysis:**

Machine learning algorithms can be used with RFM (Recency, Frequency, and Monetary Value) analysis to gain insights into customers and their behavior. This combination of techniques can provide a powerful tool for understanding customers.

Machine learning algorithms can be used to identify patterns in customer behavior over time. By analyzing the data from RFM analysis, machine learning algorithms can identify customer segments with similar behavior and preferences. This can help businesses better understand the needs and wants of their customers, allowing them to tailor their marketing strategies to target specific segments. (Zeng, Wang, Li, & Jiang, 2015)

For example, a business may use machine learning algorithms to identify customers who are likely to be loyal customers or those who are likely to churn. This can help the business target marketing messages and promotions to the right customers at the right time.

Machine learning algorithms can also be used to analyze customer data to uncover patterns in customer behavior that may not be clear with RFM analysis alone. For example, an algorithm may be able to identify customers who are likely to purchase certain products or services or those who are likely to engage with a certain type of content. This can help businesses identify opportunities to increase sales or improve customer engagement.

In summary, machine learning algorithms can be used in conjunction with RFM analysis to gain a better understanding of customers and their behavior. This can help businesses better target their marketing messages and promotions, as well as identify opportunities to increase sales or improve customer engagement. (Namvar, Khakabimamaghani, & Gholamian, 2011)

Another particularly important impact that the RFM metric brings is the easy for calculating CLTV of the customers. The CLTV has almost the same metrics as the RFM with just an extra variable indicating the customers age from the time of their last purchase. By understanding CLTV, companies can better understand the value of their customers, predict customer behavior, and optimize their marketing and customer service strategies to maximize profits. CLTV can also help companies make more informed decisions about pricing, product development, and customer acquisition. By understanding the customer’s lifetime value, companies can make better decisions on which customers to target, retain, and reward. Additionally, CLTV can help companies better measure the success of their marketing efforts, allowing them to refine their strategies and optimize their spend.

3 Methodology

**3.1 Software’s Used:**

Python is the primary application used to carry out the data cleaning, data preparation, and modelling processes. Data analysis works particularly well with Python. It comes with several built-in data structures that may be used to store data, including lists and dictionaries. Additionally, it offers several potent libraries that may be used to handle data, including NumPy and pandas. While Power BI was used for most of the visualization. The software is perfect for combining data from many sources since it enables us to connect to a wide variety of data sources. Additionally, Power BI provides a range of visualization choices, making it simpler to identify trends and patterns. The fact that Power BI has integrated statistics and machine learning tools also.

**3.2 Dataset Description:**

The dataset utilized in this work was obtained for exploratory data analysis from the UCI Machine Learning Repository. This multinational data collection includes every transaction made by a UK-based, registered online retailer between December 1, 2010, and December 9, 2011. The firm primarily offers one-of-a-kind presents for every occasion. The firm has a large number of wholesalers as clients. Each column lists the properties of each client, and each row represents a single customer.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Attribute Type | Description |
| InvoiceNo | Nominal | A six-digit integral number that is assigned to every transaction specifically. This code denotes a cancellation if it begins with the letter "C." |
| StockCode(Product (item) code) | Nominal | A 5-digit integral number uniquely assigned to each distinct product. |
| Description( Product (item) name) | Nominal | Name of the product(gift item name to be precise) |
| Quantity | Numeric | The quantities of each product (item) per transaction |
| InvoiceDate | Numeric | The day and time when a transaction was generated. |
| UnitPrice | Numeric | Product price per unit in sterling (Â£). |
| CustomerID | Nominal | A 5-digit integral number uniquely assigned to each customer. |
| Country | Nominal | The name of the country where a customer lives. |

*Table 3: Data Description*

*Understanding Dataset:*

Text

Description automatically generated

*Figure 2: Shape of the raw data*

The raw dataset has a total 541909 entries or rows and eight columns mentioned above. The data has null/missing values and some duplicate values. The entire data was cleaned and modelled accordingly to go ahead with the analysis. Also, the data has transnational transaction data of almost thirty-eight countries as shown below.

*Figure 3: Count of Invoices by Countries*

The country column has a total of thirty-eight unique country values and there are no null values present.

A picture containing table

Description automatically generated

*Figure 4: Description of Null values*

Since the online store is United Kingdom based, most of the orders are from the same. Due to the presence of very less entries for other countries, the study only includes entries from the United Kingdom for further analysis.

*Diagram

Description automatically generatedData Pipeline for the Methodology:*

*Figure 5: Data Pipeline illustration*

**3.3 Data Preparation and Exploration**

**3.3.1 Feature Additions**

A new column ‘total\_price’ is added to the dataset. The column is derived by multiplying quantity with the unitcost column. The ‘total\_price’ stands for the total value of the order of each invoice. This column would help us in figuring out the monetary value of the order by the customers during the RFM analysis.

**3.3.2 Descriptive Statistics of data**

Text, letter

Description automatically generated

*Figure 6: Shape of UK retail transaction data*

The raw UK dataset has a total 490300 entries or rows and nine columns mentioned above.

Graphical user interface, application

Description automatically generated

*Figure 7: Descriptive statistics of data*

The description table presents count, mean, standard deviation, minimum, quartiles and maximum of the quantitative columns. The largest amount of price spent on a particular order is 168469 and 0 the least. And on an average around nine units of items are ordered on every order. The gift shop sells items with the unit price ranging from 0 to 38970. The entries in "Quantity” and “UnitPrice” that have negative values begin with "C," as can be seen above. The term "Cancelled" seems to be abbreviated with the letter "C."

**3.3.3 Handling Missing Data**

One of the most frequent problems when it comes to getting your data ready for analysis is missing values. Missing values may result from a variety of reasons, including human error, interruptions in the flow of data, privacy concerns, and others. For whatever reason, missing values have an effect on how well machine learning model’s function.

Table

Description automatically generated

*Figure 8: Count of missing values with missing percentages*

In our dataset only CustomerID and Description columns have missing values. Both these columns are categorical in nature and imputing the values using the mode of the respective columns would not make any sense or is not making any good to the analysis. Here the missing values are missing completely at random, there is not even the slightest connection between the missing data and any other observed or missing values in the dataset. In other words, the data points that are missing represent a random sample of the total data. Nothing systematic exists that increases the likelihood of certain data being absent compared to other data. There is no pattern that could lead to the cause of the missing data. Hence, the limit is set to be ninety for the elimination of the null values. After removing the null values, the shape of the data is given below.

Text

Description automatically generated with low confidence

*Figure 9: Shape of data without missing values*

**3.3.4 Handling Duplicate values**

Duplicate values in the dataset can affect our analysis in several ways. For example, if we are trying to calculate the average of a set of values, the presence of duplicate values can skew our results. For example, if there are duplicate values for a certain demographic, such as age, then the results of the analysis may be skewed to reflect a higher or lower age than is present in the dataset. In addition, duplicate values can also affect the accuracy of the results, as well as the validity of the conclusions that are drawn from the data. Hence, we need to get rid of these. All the duplicate rows are removed from the data.

Graphical user interface, text, application

Description automatically generated

*Figure 10: Code snippet verifying missing values*

*Data Visualization:*

*Most and least purchased products by the customers:*

A picture containing background pattern

Description automatically generated

*Figure 11: bar graph with top ten most purchase products*

Let us have a look at the top ten most popular products/items among the customers. The above is a horizontal bar graph standing for the same. Among all the items produced by the gift shop, White Hanging Heart T-Lighter Holder is the most popular one with almost two thousand orders placed by the customers for the product. While the below bar graph shows the least popular or brought product ordered by the customer. Among which all the below are the least popular ones that are only ordered once.

Background pattern

Description automatically generated

*Figure 12: bar graph with least ten most purchase products*

Most and least ordered product with respect to the quantity:

Chart, bar chart

Description automatically generated

*Figure 13: bar graph with top ten most demanded products by quantity*

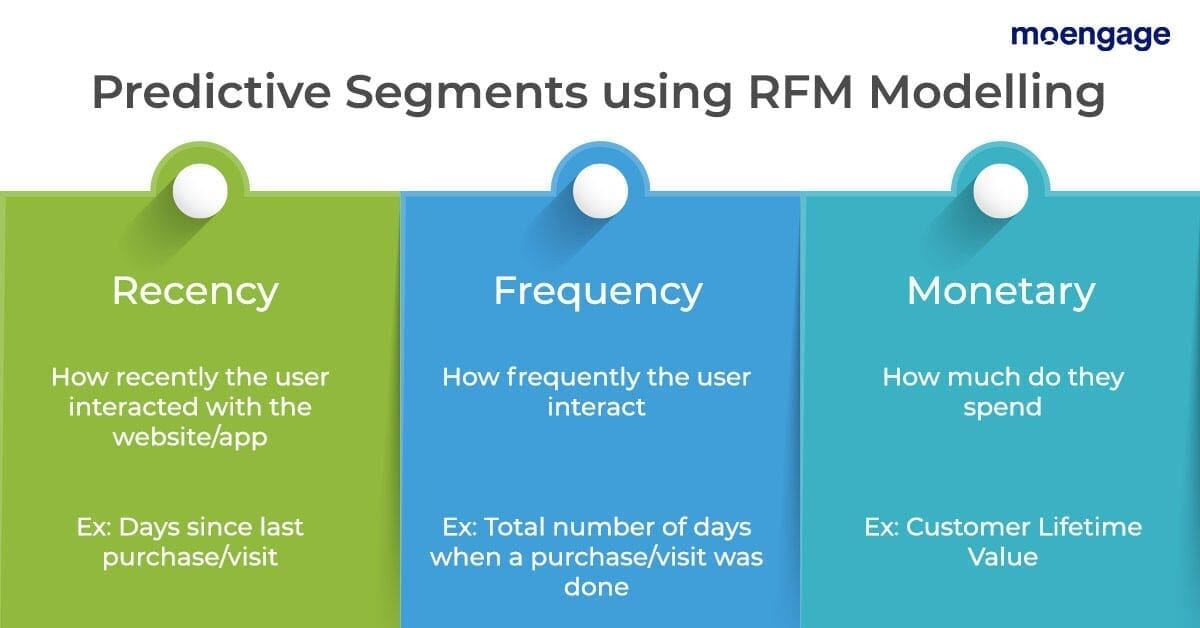
The above horizontal bar graph stands for the top ten products that were ordered in maximum quantity by the customers. ‘World war two gliders asstd designs’ was the product that was ordered in the highest quantity and then was the ‘jumbo bag red retrospot.’ While the graph below is the least ordered products quantity-wise. Among which ‘blue padded soft mobile’ is the least ordered one with quantity of almost zero. The negative number in the quantity of ‘rotating silver angels t-light hldr’ writes down the cancelled or abandoned orders by the customers.

Background pattern

Description automatically generated

*Figure 14: bar graph with least ten most demanded products by quantity*

**3.4 RFM Analysis:**



*Figure 15: RFM description chart*

The goal of the RFM analysis is listed below-

* Recognize customers who are most likely to make future purchases
* Increase customer retention by churn analyzing
* Learn more about the trends and behaviors of our customers
* Build marketing and sales strategies around the analysis

**3.4.1 Calculating Recency**

In RFM analysis, R stands for Recency.

Recency (When Did Your Customer Most Recently Purchase a Good or Service?)

Recency is figured out by counting the days since the last transaction or interaction.

For calculating the recency,we have chosen the most recent date available as a point of reference to evaluate how many days ago was the customer's last purchase.

Now the last purchase date is calculated using the reference date and the invoice date after converting them into datetime datatype.

Finally, the day since the last purchase is calculated by subtracting the last purchase date column by reference date column and converted to numeric data for further calculations.

*The new data-frame is presented below:*

Table

Description automatically generated

*Figure 16: Customer recency, data frame image*

*The Recency data frame is presented below:*

Customers with least and highest recency score

Table

Description automatically generated

*Figure 17: Top 5 Customers with high recency, data frame image*

*The histogram of recency data is presented below:*

Chart, histogram

Description automatically generated

*Figure 18: Histogram representing recency distribution*

The above histogram describes the distribution of recency of the customers purchase. From the graph it is noticeably clear that the recency is maximum around 15-25 days. So, it is good news for the company as the customers have brought products very recently from the company. The above graph stands for a right skewed distribution with the right tail of distribution being longer.

**3.4.2 Calculating Frequency:**

F stands for Frequency in RFM analysis. Frequency (How Many Times Did the Customer Buy During a Fixed Time Period or Year)

A consumer who often buys your brand is more likely to be a loyal fan of it. Businesses must examine the total number of purchases made by clients during a specific time frame in order to compute frequency.

**Frequency: Number of purchases** To calculate how many times a customer bought something; we need to count how many invoices each customer has. The frequency is counted for each customer ID and presented as below:

Table

Description automatically generated

*Figure 19: Customers frequency, data frame image*

*Customers with high frequency score:*

Table

Description automatically generated

*Figure 12: Top 5 Customers with high frequency, data frame image*

The high- frequency score writes down that the customer buys the company’s product most often and can be considered as the loyal customer. Company can provide the top five frequent buying customers with some sort of discounts, offers or rewards for their loyalty.

*Customers with low frequency score:*

Table

Description automatically generated

*Figure 21: Bottom 5 Customers with least frequency, data frame image*

The low- frequency score writes down that the customer buys the company’s product least often and can be considered as the least loyal customer. Company can analyze the issue in-order to increase the customers loyalty towards them by giving them some sort of discounts, offers, or rewards.

Shape

Description automatically generated

*Figure 22: Histogram representing Frequency distribution*

The above histogram describes the distribution of frequency of the customers buy. From the graph it is noticeably clear that the frequency is maximum around 0- 10 times. This type of histogram resembles a right-skewed distribution where the data set has values that are heavily concentrated in the lower end, with few values at the higher end. This type of graph is common when dealing with data that is not normally distributed, such as income levels. The company and device some strategies to increase the frequency of purchase.

**3.4.3 Calculating Monetary Value:**

The M in RFM analysis stands for Monetary Value. Monetary Value  (How Much Has the Consumer Spent on Your Brand So Far?)

A high financial value score writes down that a consumer has made the most purchases from your brand. The costs of the consumers' items are added up to figure out the total value.

*Top five customers with high monetary value:*

Table

Description automatically generated

*Figure 23: Top 5 Customers with high Monetary value, data frame image*

The table shows the customer IDs of the customers who have spent the most amount on item purchases. These customers have been helping the company with large order quantity and amount. These customers should be valued high by the company.

*Least five customers with high monetary value:*

Table

Description automatically generated

*Figure 24: Bottom 5 Customers with Least Monetary value, data frame image*

The least amount of purchase made by bottom five customers are listed above.

*Distribution of Monetary Value:*

Chart, histogram

Description automatically generated

*Figure 25: Histogram representing Monetary value distribution*

The above histogram describes the distribution of monetary value of the customers buy. From the graph it is noticeably clear that the monetary value is maximum around 10000- 20000. This type of histogram resembles a right-skewed distribution where the data set has values that are heavily concentrated in the lower end, with few values at the higher end. This type of graph is common when dealing with data that is not normally distributed, such as income levels. The company and device some strategies to increase the monetary value of purchase.

Strategies companies can use to increase recency, frequency, and monetary value of the customers purchases:

**1. Offer rewards and loyalty programs:** Implementing a rewards and loyalty program is one of the best ways to keep customers engaged and motivated to make frequent purchases.

**2. Personalize the shopping experience:** Personalizing the shopping experience can make customers feel special and increase the likelihood of them returning.

**3. Provide incentives for frequent purchases:** Offering incentives for frequent purchases, such as discounts or free shipping, can be a wonderful way to encourage customers to buy more.

**4. Offer exclusive deals and discounts:** Offering exclusive deals and discounts to loyal customers can be an effective way to increase purchase frequency.

**5. Focus on customer service:** Providing excellent customer service and addressing customer issues quickly can make customers feel valued and more likely to make repeat purchases.

**6. Launch a referral program:** Referral programs can be a fantastic way to increase purchase frequency as they encourage customers to share their experiences and recommend products to their friends and family.

**7. Offer multi-channel shopping experiences:** Companies can implement multi-channel shopping experiences that allow customers to purchase products from their website, mobile app, or physical store. This will help ensure customers have access to the products they need and will increase the likelihood of a purchase.

***Creating RFM table:***

We create a RFM table combining all the Recency, Frequency and Monetary values of each customer for further analysis. The first five rows are shown below:

Table

Description automatically generated

*Figure 26: RFM values for each customer, data frame image*

***Descriptive Statistics of the RFM table:***

Table

Description automatically generated

*Figure 27: Descriptive statistics of RFM values*

The descriptive statistics of the Recency, Frequency and Monetary value table provide insights into the average, maximum, minimum, quartiles, and standard deviation of the three variables. The Recency variable refers to the amount of time since the last purchase and is measured in days. The Frequency variable refers to the number of purchases made by customers and is measured in purchases. The Monetary Value variable refers to the amount of money spent by customers and is measured in dollars.

The average recency value is 99 days. This means that on an average each of the 3920 customers purchase an item every 99 days later. The largest recency value is 380 days. The smallest recency value is 7 days.

The average frequency value is four purchases. This means that on an average each of the 3920 customers has bought at least four times. The maximum frequency value is 209 purchases. The minimum frequency value is one purchase. The standard deviation for frequency is 7.2 purchases.

Similarly for the average monetary value is 1858.42. This means that on an average each of the 3920 customer has a purchase order worth 1858.42. The maximum monetary value is 259657.30. The minimum monetary value is 3.75.

**3.4.4 Assigning RFM scores to customers:**

1. For Recency:

The scoring is based on the quartiles as displayed above. The customers in the first quartile are assigned a score of 4, standing for highest recency score. Second quartile with three, third with two and the last with one, standing for least recent customers.

1. For Frequency:

The frequency score as completely converse to the recency score, last quartile is assigned a score of 4, standing for most frequent customer and goes to one for first quartile, standing for least frequent customer.

1. For Monetary Value:

This score is same as frequency, last quartile is assigned four, standing for most monetary valued customer and goes to one for first quartile, standing for customers who have spent the least amount.

**Recency label counts and proportions:**

Table

Description automatically generated with medium confidence Chart, pie chart

Description automatically generated

*Figure 28: Pie chart showing proportions of recency scores*

The four pies on the chart represent the four different scores, and each slice would stand for a different percentage of the score. The size of the slices are proportional to their value, with all the four slices almost similar to each other with slight difference. There are 1003 customers with highest recency score, 969 customers for both label 2 and 3, and 979 customers who are least recency score.

**Frequency label counts and proportions:**

Text

Description automatically generated **Chart, pie chart

Description automatically generated**

*Figure 29: Pie chart showing proportions of Frequency scores*

The frequency scores have all the four pies or slices of varied sizes. The most frequent customers are 1350 in numbers, followed by 1563 with a little less frequent and only 355 customers who are least frequent. Company should focus on these customers more, to increase their frequency of purchase.

**Monetary label counts and proportions:**

Table

Description automatically generated **Chart, pie chart

Description automatically generated**

*Figure 30: Pie chart showing proportions of Monetary scores*

The four pies on the chart represent the four different scores, and each slice would stand for a different percentage of the score. The size of the slices are proportional to their value, with all the four slices almost similar to each other with slight difference. There are 980 customers with highest monetary score, 980 and 982 customers for label 3 and 2 respectively, and 978 customers who are least monetary score.

**3.4.5 Calculating a total RFM score and assigning segment labels:**

The total RFM scores are calculated by combining all the recency, frequency, and monetary scores of customers. The labels of each customers recency, frequency and monetary scores are added to make a customer RFM score.

The below given segment labels are assigned for the customers based on the customer RFM scores. For example, A customer is labeled as Champion if he/she has a recency score of 2 or 3 or 4, a frequency score of 4 and a monetary score of 4. And a customer is labeled as Top lost customer if he/she has a recency score of 1, a frequency score of 1/2/3/4 and a monetary score of 1/2. The assignment for all the labels is given in the below table.

Table

Description automatically generated

*Figure 31: List of labels with assigned values*

**3.4.6 Labels and their description:**

1. **Champions:** Dream consumer. Purchased recently, regularly, and in huge volumes. Encourage people in their networks to spread favorable word of the brand. Because they bring in a disproportionate amount of money for the company, they need to be managed carefully.
2. **Top Loyal Customer:** High-spending customer who wants to feel acknowledged in order to promote the company. With the potential to offer much more, a significant contributor to business revenue.
3. **Loyal Customer:** Every company's core market, where customers have a clear and favorable impression of the brand, is this one. These customers are content with the goods and services they receive and are not inclined to go elsewhere. It is crucial to make these clients feel appreciated.
4. **Top Recent Customer:** Top recent customers are those who have made a purchase in the most recent time period. Because they are more likely to repurchase goods and services, these clients are often more valuable to businesses.
5. **Recent Customers:** Have just begun consuming the brand; in order to develop into long-term brand lovers, the brand must foster them. Important segment that must be regularly watched to prevent dropouts following a few recent purchases.
6. **Top Customers Needed Attention:** A valuable customer. As the name of the sector implies, it needs constant attention since these clients are loyal brand customers. These people might switch to other brands, so they need to be made feel appreciated in order to stick with it.
7. **Customers Needed Attention:** This group of users includes those who, for a variety of reasons, are thinking about discontinuing using the brand. Examining their actions and further targeting communications is vital to ensure a rise in brand trust and loyalty in order to avert this.
8. **Top Lost Customer:** Existing consumers that have not purchased  in a while and are in jeopardy of doing so. The company must make yet another effort to assure client retention and must develop the most important and relevant messaging possible to carry out so.
9. **Lost Customer:** These clients have completely discontinued using the brand and have chosen to use competitors in its place. It is quite challenging to re-engage this audience since they already have a bad opinion of the brand or another, more reliable possibility.

Table

Description automatically generated

*Figure 32: Customer category with number of customers in each*

Chart, bar chart

Description automatically generated

*Figure 33: Bar graph representing Customer category with number of customers in each*

The table above lists the categories in which customers are segmented and the number of customers in each segment in a descending order. The horizontal bar graph is just the graphical representation of the table with count of customers in each segment. The above bar graph shows a noticeably clear segments of customers with their count. The company has around eight hundred lost customers, who have completely discontinued using the brand and their products, which is the highest number in all the segments. And when we see at the champion customers, there are around 350 in number. These people make up almost 10% of the total customers which is an incredibly good thing for the company. And the top four important customers for the business make around 45% of the customer base. This indicates that the company has been in good terms with their customers, and this would last long producing a good profit for the company.

**3.5 K-Means Clustering:**

K-Means Clustering is an unsupervised machine learning algorithm used for data clustering. It is a type of clustering algorithm that attempts to find the most efficient division of a given dataset into a specified number of clusters. It is based on the idea that data points that are close together in a given space belong to the same cluster. The goal of K-Means Clustering is to minimize the within-cluster variation or the sum of squared distances between data points and their cluster centroid.

**3.5.1 Data Normalization:**

Data normalization is a statistical process where data is rescaled to have a mean of zero and a standard deviation of one. This is done to ensure that each feature (or variable) is measured on the same scale, which allows for a more effective and fair comparison between features when clustering. Normalization is required in k-means clustering because it ensures that the distances between data points are calculated accurately, which is essential for the clustering to work properly.

For our dataset we have utilized both logarithmic and power transformation and among these the one that gives the best results is taken.

**3.5.1.1 Logarithmic Method:**

Logarithmic data normalization is a method of data normalization that scales the data using the logarithm of the data values. It is used to adjust the values of highly skewed data sets so that they can be more easily compared to each other. The logarithmic transformation can help to reduce the skewness of the data and make the data more symmetric.

For this we have considered the skew threshold-limit of 0.75 to evaluate skewness. Because below abs(1) seems acceptable for the linear models.

Initially the data was plotted on the histogram to check the distribution and to check whether the distribution is symmetric or asymmetric.

The following results were found after logarithmic transformation:

Chart, histogram

Description automatically generated

*Figure 34: Distribution of logarithmically transformed RFM variables*

A screenshot of a computer

Description automatically generated with medium confidence

*Figure 35: Result of Skewness test of logarithmically transformed RFM variables*

Out of the three variables only two of them have a symmetric distribution, writing down the transformation has had some impact on the data but not fully.

**3.5.1.2 Power Transformation (Yeo-Johnson) Method:**

Power transformation data normalization is a method of transforming numerical data to a more normal or Gaussian distribution. The Yeo-Johnson method is a variant of the power transformation data normalization which allows for the use of both positive and negative values. This is required because many machine learning algorithms assume that the data follows a normal distribution, and if the data does not meet this assumption, it can lead to inaccurate results. Power transformation normalization helps to make the data more Gaussian-like, allowing the algorithms to produce more accurate results.

For this we have considered the skew threshold-limit of 0.75 to evaluate skewness. Because below abs(1) seems acceptable for the linear models.

Initially the data was plotted on the histogram to check the distribution and to check whether the distribution is symmetric or asymmetric.

Chart, histogram

Description automatically generated

*Figure 36: Distribution of power transformed RFM variables*

A number and text on a white background

Description automatically generated

*Figure 37: Result of Skewness test of power transformed RFM variables*

From the above graphs and results we can clearly see that for dealing with skewness power transformer(‘yeo-johnson’ method) was more rewarding than the logarithmic method. Hence, we use the data which is power transformed with ‘yeo-johnson’ method.

**3.5.2 Determining number of clusters(k):**

*Conducting Hopkins statistics:*

Hopkins Statistics is a statistical technique used to assess the clustering tendency of a dataset. It is conducted before clustering to evaluate the likelihood of the dataset forming meaningful clusters.

The Hopkins Statistics test uses a measure of uniformity and randomness called the Hopkins Statistic (HS) to figure out the likelihood of meaningful clusters forming in the dataset. The test works by sampling the dataset and measuring the distances between the sampled points. The HS statistic is calculated by taking the ratio of the distances within a cluster to the distances between clusters. The HS is given as a number between 0 and 1, where values between 0 and 0.5 indicate that the dataset is more likely to form meaningful clusters than not. Values between 0.5 and 1 indicate that the dataset is more likely to form random clusters than meaningful clusters. The Hopkins Statistics test can be used to figure out whether the dataset is suitable for clustering before going ahead to the clustering stage.

Graphical user interface, text, application

Description automatically generated

*Figure 38: Result of Hopkins statistics test*

Our dataset produces a value as given above. The result is almost near to zero, indicating the dataset is clusterable.

**3.5.3 The Elbow Method:**

The elbow method is a heuristic method used to determine the optimal number of clusters in a data set. It is based on the idea that the optimal number of clusters is where the inertia of the clusters begins to decrease more slowly, or where the rate of decrease in inertia begins to slow. The inertia of a cluster is a measure of how much of the variability in the data is within the cluster itself. The significance of the elbow method is that it allows for a visual representation of the data and can help identify the optimal number of clusters for a given data set.

Chart, line chart

Description automatically generated

*Figure 39: Result of elbow method to find optimal clusters(k)*

Chart, line chart

Description automatically generated

*Figure 40: Result of elbow method to find optimal clusters(k) with distortion score*

The elbow method suggests that the optimal number of clusters should be three, it means that the data can be best divided into three clusters. Each cluster will have its own distinct characteristics, which can be used to identify and classify the data points in that cluster. For example, as we are clustering for customer types, each cluster might stand for a different customer type, such as “loyal customers,” “occasional customers,” and “new customers.” The characteristics of each cluster can then be used to identify and classify customers into one of these three categories.

**3.5.4 Producing the model:**

We now build a model with three clusters and assign the cluster ids to the customer ids. After the model is built, the data with cluster ids is as follows:

Table

Description automatically generated

*Figure 41: Result of power transformed RFM variables*

Text

Description automatically generatedChart, pie chart

Description automatically generated

*Figure 42: Pie chart showing proportions of clusters Ids*

The K-means algorithm is implemented on the data and the proportion of the data being divided by cluster ids is given above. We can see that cluster zero and cluster two are almost of the same proportion around 36-37% of the data while the cluster one is a bit less with 26% of the total data.

Chart, box and whisker chart

Description automatically generated

*Figure 43: Box plot representing five number summaries of RFM variables*

The above box plot is plotted with all the three variables. The leftmost is for Recency, followed by Frequency and then monetary value. The box plot consists of a rectangular box with vertical lines extending from the top and bottom of the box. The box is divided into quartiles, with the top and bottom of the box representing the upper and lower quartiles, respectively. The vertical lines, or whiskers, extending from the box represent the range of the data set, with the maximum and minimum values plotted. The median of the data set is located inside the box. Box plots are used to compare distributions of data between distinct groups, and to identify outliers. We see that Recency values are quite wide having bigger range and only cluster id one has some outliers. While the cluster two of frequency values is almost nil with minimum range and extreme outliers. Finally, the monetary values for all the three clusters, the range is quite small with some outliers in each of the clusters.

**Assigning Labels to the data/ cluster ID’s:**

Table

Description automatically generated

*Figure 44: RFM scores with assigned labels*

Graphical user interface, text

Description automatically generated with medium confidence

*Figure 45: Customer category with number of customers in each*

Chart, bar chart

Description automatically generated

*Figure 46: Bar graph representing Customer category with number of customers in each*

Chart

Description automatically generated

*Figure 47: Scatter plot representing clusters and their respective centroids*

Chart, scatter chart

Description automatically generated

*Figure 48: Scatter plot representing all three clusters and their respective centroids*

Cluster 0 : The first cluster is more related to the "Lost/Needed Attention Customers" Existing consumers that have not purchased  in a while and are in jeopardy of doing so. The company must make yet another effort to assure client retention and must develop the most important and relevant messaging possible to carry out so.

Cluster 1 : The second cluster belongs to the "Current Customer" who have a clear and favorable impression of the brand, is this one. These customers are content with the goods and services they receive and are not inclined to go elsewhere. It is crucial to make these clients feel appreciated.

Cluster 2 : The third cluster can be interpreted as "Top/Best Customers" who have made a purchase in the most recent time period. Because they are more likely to repurchase goods and services, these clients are often more valuable to businesses. In order to develop into long-term brand lovers, the brand must foster them. Important segment that must be regularly watched to prevent dropouts following a few recent purchases.

**3.6 Customer Lifetime Value:**

CLTV stands for Customer Lifetime Value, which is a metric used to measure the total value a customer will bring to` a business over their lifetime. It is typically used to evaluate customer loyalty and predict future sales. CLTV is calculated by multiplying the average purchase value by the average number of repeats purchases a customer will make over a specified period of time.

*CTLV = Number of Transactions \* Average Purchase Value*

**3.6.1 Preparing data for CLTV prediction:**

The same processed data used for RFM analysis is used here with just addition of one more column “t.” The variables and their description are given as follows:

**Frequency:** Representing Number of purchases and transactions

**T:** The customer's age from the time of their first purchase to the present.

**Recency:** The gap in days of customers first and last purchase

**Monetary:** The typical customer total sales.

The data prepared for analysis is presented in the data frame below:

Table

Description automatically generated

*Figure 49: R F M and T values of customers*

Now let us estimate expected number of transactions and Average purchase value.

For number of transactions, we use Beta Geometric / Negative Binomial (BNB) distribution model.

**3.6.1.1 Beta Geometric / Negative Binomial (BNB) distribution model**

The Beta Geometric / Negative Binomial (BNB) distribution model is a type of stochastic process used to predict the number of purchases that a customer is likely to make over a given period of time. It is commonly used in marketing and customer analytics to forecast customer purchase behavior.

The BNB model is based on the assumption that purchases are independent, identically distributed (i.i.d), and follow a geometric distribution. This means that each purchase is independent of the previous one and follows the same probability distribution. The absence of multicollinearity is proved by the correlation matrix.

Text

Description automatically generated

*Figure 50: Correlation result between monetary and recency values*

The model assumes that the probability of a purchase decreases with each purchase, which is known as a negative binomial. This is also true as the probability is seen decreasing with the increase in the number of purchases i.e., frequency.

A picture containing text, electronics, display, screenshot

Description automatically generated

*Figure 51: Probability density curve of customers recency and frequency*

The BNB model estimates the expected number of purchases by calculating a parameter called the purchase rate. This rate is calculated as the ratio of the number of purchases to the total number of customers. The purchase rate is then used to calculate the probability of a customer making a purchase given their prior purchase history. The expected number of purchases is then calculated using the probability of a purchase for each customer.

The data after the expected number of purchases, looks as below:

Table

Description automatically generated

*Figure 52: R F M and T values with expected number of purchases*

**3.6.1.2 Gamma Gamma Model:**

For average purchase value, we Gamma Gamma model.

The Gamma-Gamma Model is a marketing model used to estimate the average purchase value (APV) of a customer. It is based on the idea that the purchase volume of a customer is related to their potential loyalty to a brand. The model uses the purchase history and recency of purchases to calculate the APV of a customer.

The model works by taking the total amount spent by a customer over a period of time and dividing it by the number of purchases made. This gives the average purchase value. The model also takes into account the recency of the purchases. This means that customers who have recently made a purchase are given a higher APV than those who have not made one in a while.

The Gamma-Gamma Model is often used in customer segmentation and loyalty programs. By understanding the APV of a customer, businesses can target their marketing efforts. This can help to increase customer loyalty and lifetime value.

The data after the expected number of purchases, looks as below:

Table

Description automatically generated

*Figure 53: data frame after expected average order value calculation*

The Final CLTV calculation for the next one year is estimated and presented below:

Table

Description automatically generated with medium confidence

*Figure 54: list of top five customers with highest predicted CLTV values*

The above table represents the list of top 5 United Kingdom customers who have the highest lifetime value for the next year. This indicates that these top five customers are the ones who are going to spend the most on you brand during their lifetime.

**3.6.2 Interpreting the results:**

The Pareto principle, also known as the 80/20 rule, states that 80% of outcomes come from 20% of the inputs. In terms of customer loyalty, the Pareto principle suggests that 80% of a company’s customer lifetime value (CLTV) comes from 20% of its customers.

Real-world examples of this principle in action can involve a company that has 10,000 customers. According to the Pareto principle, only 2,000 of those customers (the top 20%) are responsible for 80% of the company’s CLTV. This means that the remaining 8,000 customers only account for the remaining 20% of the CLTV.

In order to maximize their CLTV, companies should identify and focus their efforts on the 2,000 customers that generate 80% of their CLTV. This could include providing them with additional incentives or reward programs, or simply offering them better customer service. Companies should also strive to identify and acquire new customers that have the potential to generate the same level of CLTV as their best customers.

By applying the Pareto principle to their customer loyalty efforts, companies can ensure that they are focusing their efforts on the customers that will have the greatest impact on their CLTV.

**Customers with least CLTV values:**

A picture containing table

Description automatically generated

*Figure 55: list of bottoms five customers with least predicted CLTV values*

The above table represents the list of least 5 United Kingdom customers who have the least lifetime value for the next year. This indicates that these bottom five customers are the ones who are going to spend the least on you brand during their lifetime. Focusing on these customers would be still important for the companies as they can try to turn the tables around to make these customers to trust their brand and increase their loyalty.

In order to increase their CLTV value, companies should focus on understanding their customers better and developing strategies to engage and retain them. Companies should identify and track key customer information, such as their purchase history, preferences, engagement levels, and demographic information. Once they have this data, they can use it to develop personalized marketing campaigns and loyalty programs that will help them increase customer loyalty and drive more sales.

For example, a company might develop a loyalty program that rewards customers with discounts or exclusive products. By understanding the needs and preferences of their customers, they can be sure to offer rewards that will be attractive to them and encourage them to remain loyal to the company. Additionally, the company can use its customer data to create segmented marketing campaigns that are tailored to specific customer groups. By targeting campaigns to specific groups, the company can more effectively engage customers and increase their CLTV.

**3.6.3 Recommended Marketing Strategies:**

Below are the proven strategies used to improve CLV by successful companies world-wide.

1. Increase Customer Retention

Retaining existing customers is one of the most effective ways to increase Customer Lifetime Value (CLV). Companies should focus on creating a customer experience that encourages loyalty and drives repeat purchases. Strategies to increase customer retention include offering rewards and loyalty programs, creating personalized offers and experiences, providing exceptional customer service, and collecting and analyzing customer data to better understand customer behavior.

2. Increase Average Order Value

Increasing the average order value (AOV) is another effective way to improve CLV. Companies can do this by offering discounts on multiple items, upselling and cross-selling related products, and providing exclusive offers and bundles. Additionally, companies should segment their customers and tailor their offers to different customer types.

3. Increase Frequency of Purchases

Increasing the frequency of purchases is another way to improve CLV. Companies can do this by offering incentives and frequent discounts, sending targeted emails and ads, and providing personalized experiences. Companies should also use customer data to understand customer behavior and tailor their offers to individual customers.

4. Increase Customer Referrals

Encouraging customers to refer their friends is a fantastic way to increase CLV. Companies can do this by offering referral rewards and incentives, creating referral campaigns and contests, and providing an effortless way for customers to share their purchases with their friends.

5. Develop a Unique Value Proposition

Developing a unique value proposition that sets your company apart from the competition is essential for increasing CLV. Companies should focus on creating an experience and offering products and services that customers cannot find anywhere else. This could include a unique product, a personalized service, or a loyalty program.

6. Increase Customer Engagement

Engaging customers is key to increasing CLV. Companies can do this by sending personalized emails, offering exclusive discounts and rewards, hosting events and webinars, and creating engaging content and experiences. Companies should also focus on providing a seamless and convenient customer experience across all channels.

7. Invest in Customer Support

Investing in customer support is critical for improving CLV. Companies should focus on providing exceptional customer service, responding to customer inquiries quickly, and making it easy for customers to contact the company with their questions or concerns. Additionally, companies should invest in automated customer support solutions to improve the customer experience.

4 Conclusion

K-means clustering is an unsupervised learning algorithm that is used for clustering data into distinct groups. In order to use the k-means clustering algorithm, the following data requirements must be met:

1. The data must be numeric: K-means clustering requires that the data be numeric in nature. This means that the data must be numerical values such as integers, floats, and doubles; it cannot be categorical data such as strings or Boolean values.

2. The data must be scaled: The data must be scaled so that all of the features are on the same scale. This is necessary because the algorithm works by calculating the Euclidean distance between points in the feature space. If the features are not scaled, then the algorithm will produce inaccurate results.

3. The data must be labeled: Labeling the data is not a requirement but it is highly recommended. Labeling the data helps to interpret the results of the clustering algorithm.

4. The data must be normalized: Normalization is not a requirement, but it is recommended for datasets that do not follow a normal distribution. Normalization helps to make sure that the results of the clustering algorithm are not biased by extreme values.

So, to meet all the data requirements can be quite challenging sometimes. Now-a-days companies perform clustering on customers data which includes their age, salary, geographical locations etc. But researchers have observed that RFM (Recency, Frequency, and Monetary) data is generally more powerful than demographic data in understanding customer behavior. This is because RFM data provides detailed insights into customer behavior, including when they last purchased, how often they purchase, and how much they spend. Demographic data, on the other hand, is not as dependable in predicting customer behavior, as it does not take into account any changes in customers’ behavior over time. For example, a customer’s age, gender, or income level may not accurately reflect their current buying habits. Additionally, RFM data can be used to identify customer segments and target them with tailored marketing messages, while demographic data cannot.

The same can be seen in this study. With the help of customers data which was not suitable for the K-means clustering was segmented using RFM analysis. The RFM analysis initially performed aided in the implementation of the K-Means clustering algorithm and making it more effective than before. Performing customer segmentation is an excellent way to understand them but using RFM analysis before applying K-Means clustering can help businesses better understand their customers and target them with relevant marketing campaigns. By combining RFM analysis with K-Means clustering, businesses can better identify customer segments and target them with the right message and offer. This will help them to maximize their marketing ROI and increase customer loyalty.

5 Recommendation

Based on the evidence presented, it is clear that machine learning algorithms can be used effectively to understand customer behavior. However, there are still some areas where further research and development are needed. It is recommended that businesses continue to explore the potential of machine learning algorithms for understanding customer behavior. Additionally, businesses should consider investing in advanced machine learning algorithms, such as deep learning algorithms, to further improve their understanding of customer behavior.

Furthermore, businesses should explore the potential of combining machine learning algorithms with other technologies, such as natural language processing (NLP) and artificial intelligence (AI), to improve their understanding of customer behavior. Additionally, businesses should continue to explore the potential of using machine learning algorithms to personalize customer experiences and improve customer service. Finally, businesses should explore the potential of using machine learning algorithms to identify customer segments and target them with personalized marketing campaigns.

However, there are several things that a company can do to understand customers beyond analyzing their data using machine learning algorithms. Here are a few recommendations:

1. Conduct market research: Market research is a systematic and scientific process of collecting, analyzing, and interpreting data about a specific market or industry. It can be conducted through various methods such as surveys, focus groups, interviews, and online analytics. Market research helps companies understand their customers' needs, preferences, and behavior, as well as the competitive landscape.

2. Use customer feedback: Customer feedback is a powerful tool for understanding customers. It provides valuable insights into what customers like and dislike about a company's products and services, as well as areas for improvement. Companies can use various channels to collect customer feedback, such as surveys, social media, email, and phone calls. It's important to act on the feedback received and communicate the changes made to customers.

3. Personalize the customer experience: Personalization can help companies better understand and connect with their customers. It involves tailoring products, services, and communication to the individual needs and preferences of each customer. Personalization can be achieved through various means, such as personalized recommendations, targeted marketing, and personalized communication.

4. Foster customer relationships: Building strong, long-lasting relationships with customers is essential for understanding them. Companies can foster customer relationships through various means, such as providing excellent customer service, offering personalized experiences, and engaging with customers on social media. Companies can also establish customer loyalty programs to reward and retain their most valuable customers.

5. Empower employees to understand customers: Companies can empower their employees to understand customers by providing them with the necessary resources and training. This includes giving them access to customer data, encouraging them to listen to and act on customer feedback, and providing them with customer service skills training. Companies can also establish customer-centric cultures by making customer satisfaction a top priority and recognizing and rewarding employees for going above and beyond for customers.

6. Collaborate with customers: Companies can also collaborate with customers to better understand their needs and preferences. This can be done through co-creation, in which companies collaborate with customers to design and develop new products and services. Companies can also collaborate with customers through co-innovation, in which customers and companies work together to find solutions to specific problems or challenges.

7. Observe and interact with customers: Companies can also gain a deeper understanding of their customers by observing and interacting with them directly. This can be done through in-person interactions such as focus groups or through online interactions such as chat or video sessions. By observing and interacting with customers, companies can get a first-hand understanding of their needs, preferences, and behavior.

8. Analyze customer behavior: Companies can analyze customer behavior through various means such as tracking website or app usage, analyzing social media interactions, or tracking customer purchasing habits. This can provide valuable insights into how customers interact with the company and its products and services, as well as areas for improvement.

By implementing these recommendations, companies can gain a deeper understanding of their customers and better meet their needs and wants. This can lead to increased customer satisfaction and loyalty, which can drive business growth and success.

6 References

Aggelis, V., & Christodoulakis, D. (2005). *Customer clustering using RFM analysis*. Retrieved 1 5, 2023, from https://dl.acm.org/citation.cfm?id=1369601

Akpoyomare, O. B., Adeosun, L. P., & Ganiyu, R. A. (2013). The Influence of Product Attributes on Consumer Purchase Decision in the Nigerian Food and Beverages Industry: A Study of Lagos Metropolis. *American Journal of Business and Management, 1*(4), 196-201. Retrieved 1 5, 2023, from http://worldscholars.org/index.php/ajbm/article/view/ajbm1237

Anitha, P., & Patil, M. M. (2019). RFM model for Customer Purchase Behavior using K-Means Algorithm. *Journal of King Saud University - Computer and Information Sciences*. Retrieved 1 5, 2023, from https://sciencedirect.com/science/article/pii/s1319157819309802

Aschettino, L., Birnbaum, C. L., Crocker, J., Grebner, L. A., & McNicholas, F. C. (2013). How Can I Help You? Top 10 Customer Service Tips for HIM Professionals. *Journal of AHIMA, 84*(5), 32-36. Retrieved 1 5, 2023, from https://library.ahima.org/doc?oid=106374

Asllani, A., & Halstead, D. (2015). A Multi-Objective Optimization Approach Using the RFM Model in Direct Marketing. *Academy of Marketing Studies Journal, 19*(2), 65. Retrieved 1 5, 2023, from http://alliedacademies.org/articles/a-multiobjective-optimization-approach-using-the-rfm-model-in-direct-marketing.pdf

Benoit, D. F., & Poel, D. V. (2009). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. *Expert Systems With Applications, 36*(7), 10475-10484. Retrieved 1 5, 2023, from http://dblp.uni-trier.de/db/journals/eswa/eswa36.html

Chan, C. C., Hwang, Y.-R., & Wu, H.-C. (2016). Marketing segmentation using the particle swarm optimization algorithm: a case study. *Journal of Ambient Intelligence and Humanized Computing, 7*(6), 855-863. Retrieved 1 5, 2023, from https://link.springer.com/article/10.1007/s12652-016-0389-9

Chatterjee, A., & Yuen, P. (2019). Endmember Learning with K-Means through SCD Model in Hyperspectral Scene Reconstructions. *Journal of Imaging, 5*(11), 85. Retrieved 1 5, 2023, from https://dspace.lib.cranfield.ac.uk/handle/1826/14755

Cheng, C.-J., Chiu, S., Cheng, C.-B., & Wu, J.-Y. (2012). Customer lifetime value prediction by a Markov chain based data mining model: Application to an auto repair and maintenance company in Taiwan. *Scientia Iranica, 19*(3), 849-855. Retrieved 1 5, 2023, from https://sciencedirect.com/science/article/pii/s1026309812000107

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2018). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*. Retrieved 1 5, 2023, from https://sciencedirect.com/science/article/pii/s1319157818304178

Dalhatu, K., & Sim, A. T. (2016). Density base k-Mean's Cluster Centroid Initialization Algorithm. *International Journal of Computer Applications, 137*(11), 48-51. Retrieved 1 5, 2023, from https://ijcaonline.org/archives/volume137/number11/24323-2016908923

Dominici, G., & Guzzo, R. (2010). Customer Satisfaction in the Hotel Industry: A Case Study from Sicily. *International Journal of Marketing Studies, 2*(2), 3. Retrieved 1 5, 2023, from http://ccsenet.org/journal/index.php/ijms/article/view/8103

Dong, L., He, L., Mao, M., Kong, G., Wu, X., Zhang, Q., . . . Izquierdo, E. (2018). CUNet: A Compact Unsupervised Network For Image Classification. *IEEE Transactions on Multimedia, 20*(8), 2012-2021. Retrieved 1 5, 2023, from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8241834

Ekinci, Y., Ülengin, F., Uray, N., & Ülengin, B. (2014). Analysis of customer lifetime value and marketing expenditure decisions through a Markovian-based model. *European Journal of Operational Research, 237*(1), 278-288. Retrieved 1 5, 2023, from https://sciencedirect.com/science/article/pii/s0377221714000162

*Examples of Marketing Strategies Used to Sell a Product*. (n.d.). Retrieved 1 5, 2023, from http://smallbusiness.chron.com/examples-marketing-strategies-used-sell-product-10850.html

Fernández-Martínez, J. L., & Fernández-Muñiz, Z. (2020). The curse of dimensionality in inverse problems. *Journal of Computational and Applied Mathematics, 369*, 112571. Retrieved 1 5, 2023, from https://sciencedirect.com/science/article/pii/s037704271930576x

Filip, A., & Voinea, L. (2012). Understanding the Processes of Customer Acquisition, Customer Retention and Customer Relationship Development. *International Journal of Economic Practices and Theories, 2*(2), 62-67. Retrieved 1 5, 2023, from http://ijept.org/index.php/ijept/article/view/understanding\_the\_processes\_of\_customer\_acquisition,\_customer\_retention\_and\_customer\_relationship\_development

Ghazzawi, K., Nemar, S. E., Sankari, A., Tout, S., Dennaoui, H., & Shoghari, R. e. (2016). The Impact of CSR on Buying Behavior: Building Customer Relationships. *Management Science, 6*(4), 103-112. Retrieved 1 5, 2023, from http://article.sapub.org/10.5923.j.mm.20160604.02.html

Gordon|Kruse, L. M. (2017). Little Boxes: Micro-apartments have become trendy in planning circles, but their austerity is just another limit on the aspirations of the poor. *The Nation, 306*(280, no. 2). Retrieved 1 5, 2023, from http://thenation.com

Hu, Y.-H., & Yeh, T.-W. (2014). Discovering valuable frequent patterns based on RFM analysis without customer identification information. *Knowledge Based Systems, 61*(1), 76-88. Retrieved 1 5, 2023, from https://sciencedirect.com/science/article/pii/s0950705114000586

Januszewski, F. (2011). *POSSIBLE APPLICATIONS OF INSTRUMENTS OF MEASUREMENT OF THE CUSTOMER VALUE IN THE OPERATIONS OF LOGISTICS COMPANIES*. Retrieved 1 5, 2023, from http://logforum.net/pdf/7\_4\_2\_11.pdf

Kaul, D. (2017). Customer Relationship Management (CRM), Customer Satisfaction and Customer Lifetime Value in Retail. *Review of Professional Management- A Journal of New Delhi Institute of Management, 15*(2), 55-60. Retrieved 1 5, 2023, from http://i-scholar.in/index.php/rpmndim/article/view/163914

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science, 3*, 57-63. Retrieved 1 5, 2023, from https://sciencedirect.com/science/article/pii/s1877050910003868

Khalid, W., & Herbert-Hansen, Z. N. (2018). *Using k-means clustering in international location decision*. Retrieved 1 5, 2023, from https://emerald.com/insight/content/doi/10.1108/jgoss-11-2017-0056/full/html

Khan, I. (2012). Impact of customer satisfaction and retention on customer loyalty. *International Journal of Scientific & Technology Research, 1*(2), 106-110. Retrieved 1 5, 2023, from http://ijstr.org/final-print/march2012/impact-of-customers-satisfaction-and-customers-retention-on-customer-loyalty.pdf

Kho, L. C., Kasihmuddin, M. S., Mansor, M. A., & Sathasivam, S. (2020). Logic mining in football matches. *Indonesian Journal of Electrical Engineering and Computer Science, 17*(2), 1074-1083. Retrieved 1 5, 2023, from http://ijeecs.iaescore.com/index.php/ijeecs/article/view/18825

Kindhi, B. A., Sardjono, T. A., Purnomo, M. H., & Verkerke, G. J. (2019). Hybrid K-means, fuzzy C-means, and hierarchical clustering for DNA hepatitis C virus trend mutation analysis. *Expert Systems With Applications, 121*, 373-381. Retrieved 1 5, 2023, from https://sciencedirect.com/science/article/pii/s0957417418307875

Lai, A. W. (1995). Consumer Values, Product Benefits and Customer Value: a Consumption Behavior Approach. *ACR North American Advances*. Retrieved 1 5, 2023, from http://acrwebsite.org/search/view-conference-proceedings.aspx?id=7772

Machap, L., & Abdullah, A. (2020). Functional analysis of cancer gene subtype from co-clustering and classification. *Indonesian Journal of Electrical Engineering and Computer Science, 18*(1), 343-350. Retrieved 1 5, 2023, from http://ijeecs.iaescore.com/index.php/ijeecs/article/view/21118

Maechler, M., Rousseeuw, P. J., Struyf, A., & Hubert, M. (2015). *"Finding Groups in Data": Cluster Analysis Extended Rousseeuw etal.* Retrieved 1 5, 2023, from https://cran.r-project.org/web/packages/cluster/index.html

McCormick, H., & Livett, C. (2012). Analysing the influence of the presentation of fashion garments on young consumers’ online behaviour. *Journal of Fashion Marketing and Management, 16*(1), 21-41. Retrieved 1 5, 2023, from https://emerald.com/insight/content/doi/10.1108/13612021211203014/full/html

Miglautsch, J. R. (2002). Application of RFM principles: What to do with 1-1-1 customers? *The Journal of Database Marketing & Customer Strategy Management, 9*(4), 319-324. Retrieved 1 5, 2023, from https://link.springer.com/content/pdf/10.1057/palgrave.jdm.3240080.pdf

Mitrović, S., Baesens, B., Baesens, B., Lemahieu, W., & Weerdt, J. D. (2019). tcc2vec: RFM-informed representation learning on call graphs for churn prediction. *Information Sciences*. Retrieved 1 5, 2023, from https://sciencedirect.com/science/article/pii/s0020025519301537

Namvar, M., Khakabimamaghani, S., & Gholamian, M. R. (2011). An approach to optimised customer segmentation and profiling using RFM, LTV, and demographic features. *International Journal of Electronic Customer Relationship Management, 5*, 220-235. Retrieved 1 5, 2023, from https://espace.library.uq.edu.au/view/uq:e6d0da6

Parracho, A., Melo-Gonçalves, P., & Rocha, A. (2016). Regionalisation of precipitation for the Iberian Peninsula and climate change. *Physics and Chemistry of The Earth, 94*, 146-154. Retrieved 1 5, 2023, from https://sciencedirect.com/science/article/pii/s1474706515000777

Reisenwitz, T. H., & Gupta, S. (2016). Brand Loyalty and Store Loyalty for Consumers Purchasing a Product Warranty in a Health Care Setting: An Investigation of the Differences across Gender, Age, and Income Groups. *Journal of Business Strategies, 33*(1), 1. Retrieved 1 5, 2023, from https://questia.com/library/journal/1g1-459287688/brand-loyalty-and-store-loyalty-for-consumers-purchasing

Roper, Z. J., & Vecera, S. P. (2013). Response terminated displays unload selective attention. *Frontiers in Psychology, 4*, 967-967. Retrieved 1 5, 2023, from https://ir.uiowa.edu/psychology\_pubs/18

Rutkauskas, J., & Paulavicienė, E. (2015). Concept of Productivity in Service Sector. *The Engineering Economics, 43*(3), 35-41. Retrieved 1 5, 2023, from http://inzeko.ktu.lt/index.php/ee/article/view/11284

Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. *Kybernetes, 45*(7), 1129-1157. Retrieved 1 5, 2023, from https://emerald.com/insight/content/doi/10.1108/k-07-2015-0180/full/html

Shanker, A. (2012). A Customer Value Creation Framework for Businesses That Generate Revenue with Open Source Software. *Technology Innovation Management Review, 2*(3), 18-22. Retrieved 1 5, 2023, from https://timreview.ca/article/534

Singh, S. S., & Jain, D. C. (2013). *Measuring Customer Lifetime Value: Models and Analysis*. Retrieved 1 5, 2023, from https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=2214860

Song, M., Zhao, X., Haihong, E., & Ou, Z. (2016). *Statistic-based CRM approach via time series segmenting RFM on large scale data*. Retrieved 1 5, 2023, from https://dl.acm.org/ft\_gateway.cfm?id=3007873&ftid=1821105&dwn=1

Souri, F. (2017). The Effect Of Internal Stimulus On Posthaste Buying Behavior Of The Chain Store Customers (Case Study: Hyperstar). *International Journal of Scientific & Technology Research, 6*(6), 216-224. Retrieved 1 5, 2023, from http://ijstr.org/final-print/june2017/the-effect-of-internal-stimulus-on-posthaste-buying-behavior-of-the-chain-store-customers-case-study-hyperstar.pdf

Suresh, K., & Paul, S. (2010). A Simple Approach to Clustering in Excel. *International Journal of Computer Applications, 11*(7), 24-28. Retrieved 1 5, 2023, from https://ijcaonline.org/archives/volume11/number7/1595-2144

*The Disadvantages of Target Marketing*. (n.d.). Retrieved 1 5, 2023, from http://smallbusiness.chron.com/disadvantages-target-marketing-36131.html

Watkins, L., Aitken, R., Hinder, C., Lawson, R., Mather, D., Paul, A., . . . Williams, J. (2015). The New Zealand consumer lifestyle segments. *New Zealand sociology, 30*(1), 111. Retrieved 1 5, 2023, from https://questia.com/library/journal/1p3-3776211181/the-new-zealand-consumer-lifestyle-segments

Weidema, M., Geer, E. v., Koelsche, C., Desar, I. M., Kemmeren, P., Hillebrandt-Roeffen, M. H., . . . Flucke, U. (2020). DNA methylation profiling identifies distinct clusters in angiosarcomas. *Clinical Cancer Research, 26*(1), 93-100. Retrieved 1 5, 2023, from https://clincancerres.aacrjournals.org/content/early/2019/09/27/1078-0432.ccr-19-2180

Wu, H.-H., Lin, S.-Y., & Liu, C. W. (2014). Analyzing Patients’ Values by Applying Cluster Analysis and LRFM Model in a Pediatric Dental Clinic in Taiwan. *The Scientific World Journal, 2014*, 685495-685495. Retrieved 1 5, 2023, from https://ncbi.nlm.nih.gov/pmc/articles/pmc4090562

Xia, H., Karimi, H. A., & Meng, L. (2017). Parallel implementation of Kaufman’s initialization for clustering large remote sensing images on clouds. *Computers, Environment and Urban Systems, 61*, 153-162. Retrieved 1 5, 2023, from https://sciencedirect.com/science/article/pii/s0198971514000659

Zeng, X., Wang, Q., Li, Q., & Jiang, J. (2015). *A Multi-indicator Customer Segmentation Method Based on Consuming Behaviors Analysis*. Retrieved 1 5, 2023, from https://computer.org/csdl/proceedings/icnisc/2015/1843/00/1843a289-abs.html