

**CUSTOMER SEGMENTATION USING RFM ANALYSIS**

**BY**

**AHMED MUHTASIM ANJUM**

**ID: 152-15-6060**

This Report Presented in Partial Fulfillment of the Requirements for the Degree  
of Bachelor of Science in Computer Science and Engineering

Supervised By

**MD. SADEKUR RAHMAN**

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised By

**MS. FARAH SHARMIN**

Sr. Lecturer

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

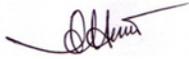
**DHAKA, BANGLADESH**

**OCTOBER 2020**

## APPROVAL

This Project titled “**Customer Segmentation using RFM Analysis**”, submitted by Ahmed Muhtasim Anjum to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 07-10-2020.

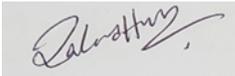
### BOARD OF EXAMINERS



---

**Dr. Syed Akhter Hossain**  
**Professor and Head**  
Department of CSE  
Faculty of Science & Information Technology  
Daffodil International University

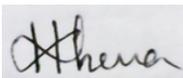
**Chairman**



---

**Md. Zahid Hasan**  
**Assistant Professor**  
Department of CSE  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



---

**Most. Hasna Hena**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



---

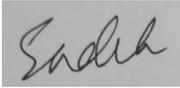
**Dr. Mohammad Shorif Uddin**  
**Professor**  
Department of Computer Science and Engineering  
Jahangirnagar University

**External Examiner**

## DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Md. Sadekur Rahman, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

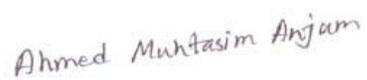
**Supervised by:**



---

**Md. Sadekur Rahman**  
**Assistant Professor**  
Department of CSE  
Daffodil International University

**Submitted by:**



---

**Ahmed Muhtasim Anjum**  
ID: -152-15-6060  
Department of CSE  
Daffodil International University

---

## ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Md. Sadekur Rahman, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Prof. Syed Akhter Hossain, Head**, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

## ABSTRACT

This report is expected as a direction for teachers and engineering students when conducting inquire about is a portion of course-work requirement. Dialog incorporates a depiction of writing look, the reason for a writing survey finding sources, and a common procedure to assist conduct a proficient and beneficial for better understanding companies that need to know the customer's information better from all perspectives. Identifying similitudes and contrasts among clients, foreseeing their behaviors, proposing superior choices, and openings to clients got to be exceptionally vital for customer-company engagement. Portioning the clients concurring to their information got to be crucial in this setting. RFM (recency, frequency, and monetary) values have been utilized for numerous a long time to distinguish which clients esteem the company, which clients require special exercises, etc. Data-mining devices and methods broadly have been utilized by organizations and people to analyze their put away information. Clustering, which one of the assignments of information mining has been utilized to gather individuals, objects, etc. We recognized that the current customer segmentation which built by fair considering customer's cost isn't adequate. Thus, models suggested in this investigation are anticipated to supply way better client understanding, well-designed techniques, and more proficient decisions. Segmentation which built by just considering customers' cost isn't adequate. Utilizing apparatuses such as this report, students can be gotten to be more pro-active around their research ventures. Instructors can utilize this report, among other instruments, to start a dialog with their understudies around desires for inquiring about assignments. Related data is rehashed within the outline segment for comfort. A clarified reference list is included for ease in finding other valuable directions.

## TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	II
Declaration	III
Acknowledgements	IV
Abstract	V
List of Tables	IX
List of Figures	X
 <b>CHAPTERS</b>	
<b>CHAPTER 1: Introduction</b>	<b>11-14</b>
1.1 Introduction	11
1.2 Motivation	12
1.3 Rationale of the Study	12
1.4 Research Questions	13
1.5 Expected Output	13
1.6 Project Management and Finance	13
1.7 Report Layout	14

<b>CHAPTER 2: Background</b>	<b>15-18</b>
2.1 Preliminaries/Terminologies	15
2.2 Related Works	16
2.3 Comparative Analysis and Summary	16
2.4 Scope of the Problem	17
2.5 Challenges	17
<b>CHAPTER 3: Research Methodology</b>	<b>19-29</b>
3.1 Research Subject and Instrumentation	19
3.2 Data Collection Procedure/Dataset Utilized	19
3.3 Statistical Analysis	20
3.4 Proposed Methodology/Applied Mechanism	24
3.5 Implementation Requirements	29
<b>Chapter 4: Experimental Results and Discussion</b>	<b>30-35</b>
4.1 Experimental Setup	30
4.2 Experimental Results & Analysis	34
4.3 Discussion	35
<b>Chapter 5: Impact on Society, Environment and Sustainability</b>	<b>36</b>
5.1 Impact on Society	36
5.2 Impact on Environment	36
5.3 Ethical Aspects	36
5.4 Sustainability Plan	36

**Chapter 6: Summary, Conclusion, Recommendation and Implication for Future Research 37**

6.1 Summary of the Study 37

6.2 Conclusions 37

6.3 Implication for Further Study 37

**REFERENCES 38**

## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 3.1: Recency Histogram	20
Figure 3.2: Frequency Histogram	21
Figure 3.3: Monetary Value Histogram	21
Figure 3.4: Recency Inertia Graph	22
Figure 3.5: Frequency Inertia Graph	22
Figure 3.6: Monetary Value Inertia Graph	23
Figure 3.7: Recency Scatter Plot	23
Figure 3.8: Frequency Scatter Plot	24
Figure 3.9: Proposed Model	28
Figure 4.1: Correlation Matrix	34

## LIST OF TABLES

<b>TABLES</b>	<b>PAGE NO</b>
Table 4.1: Recency Score Table	30
Table 4.2: Characteristics of Recency Cluster	31
Table 4.3: Characteristics of Frequency	31
Table 4.4: Characteristics of Revenue Cluster	32
Table 4.5: Characteristics of Recency Cluster	32
Table 4.6: Relation Table for Difference between Number of Days and Invoice Dates	33

# CHAPTER 1

## Introduction

### 1.1 Introduction

In today's business environment, increased importance is placed in customer equity. To increase market, share many companies are focusing on the notion of customer loyalty and profitability. It is no surprise for any marketer that each individual customer has their own variety of needs and want. In order to satisfy each customer's individual needs and individual wants companies are known to apply many segmentation applications to properly locate and develop better understanding for all customer groups. Customer segmentation is important in order to promote necessary and underselling products or services to higher valued customers. Since the customers can provide basic information for applying more targeted and personalized marketing, successful customer relationship management (CRM) is built by identifying customers' true value and loyalty. It is possible to generate important patterns and trends if a huge amount of data information is used correctly from data present in an organizations database. With the assistance of predictive analysis, insight into useful customer buying patterns is gained to predict the upcoming day of purchase for customers. But since this data is very complex, it is difficult to understand the customer necessities and increase their satisfaction which leads to their sustained retention. This is where customer relationship management is utilized which recognizes potential customers for retention. The details of future buying patterns of the customer are determined with the techniques of analyzing customer value and their past purchasing records. RFM Analysis is a new technique to know customer values in the future. The model is widely applied for analyzing value of each customer. There are three parameters of this technique which uses recency, frequency, and monetary value for getting the value of loyalty for each customer. Recency, Frequency, and Monetary directly proportional to customer's lifetime and retention.

## **1.2 Motivation**

It is not possible to treat each and every customer similarly to one another with the same exposure of content, similar channels, and with similar kind of importance because they will search for alternative options where they get better understanding. Customers are familiar with different platforms which allows them to express different needs and their own unique profiles. The motivation here is proposing a better adaptation method for the customer approach depending on that. It is possible to increase the retention rate based on customer segmentation. In the past few years, a vast increase in competition has been observed among firms in their respected fields. Customer segmentation models can improve the profits of a company. retaining customers has a higher importance level than acquiring fresh ones. According to the Pareto principle (Srivastava, 2016), 20% more revenue of the company is acquired from the contribution of retaining customers than funneling in new customers. There are very standard and useful segmentation methods but the purpose of implementing RFM on our online retail dataset is to select the centroids Initially for the algorithm of K-means. It will be effective to treat the customers in a way they deserve before they expect that and act before something bad happens. That is possible with the help of Predictive analytics. It is best utilized to predict a day for customer's next purchase. Future knowledge of the next purchase day is a better sign in predicting future sales as well. That provides an opportunity to build a plan beforehand that and design new actions tactically such as whether or not to deliver a promotional offer to the customer or to push the customer with new marketing if there is no new purchase in predicted time frame.

## **1.3 Rationale of the Study**

According to marketers, there are many varieties of necessities and wants for a customer. The companies are known to use several segmentation techniques to properly select and know customer groups to offer possible underselling products or services to individual customers in order to better satisfy their different necessities. Also, the selection of segments based on its competitive advantages, there is importance for segmentation for companies to develop new fruitful segments reacting to non-profitable groups. However, it is difficult for many marketers to identify the right customers to formulate new and better campaigns for marketing. Thus, resulting in failed promotions and programs over loyalty with a waste in resources to fuel marketing.

#### **1.4 Research Questions**

1. Who are the best customers?
2. Who has the potential to become valuable customers?
3. Which of the customers can be retained?
4. Which of the customers are most likely to respond to engagement campaigns?

#### **1.5 Expected Output**

The outcome of the presented work is to help marketing people in meaningful customer segmentation. The rest of the study focuses on analyzing all the clustering approaches regarding iterations, cluster compactness, execution time, and various other factors. Clustering customers into different groups helps decision-makers identify market segments more clearly and thus develop more effective marketing and sale strategies for customer retention. We clustered customers into segments according to RFM and Extended RFM parameters using K-means Algorithm. The clustering of customers is required to know the distinction between those groups of customer segments. In order to approach the appropriate customers with appropriate promotions and offers, targeted customers are found with a in-depth analysis on the clusters.

#### **1.6 Project Management and Finance**

The management of the project is based over the study of research papers obtained from google scholar and using open source online platform for studying Business model, Machine Learning models and knowledge about Python. Using Python version 3.8.5 as the prime programming language for its enriched library for data management and manipulation. Using one of the popular integrated development environments, Jupyter Notebook for writing the programming codes and for cleaning & transforming data, numerical simulation, statistical modeling and machine learning applications. The project is self-funded and the research is from home and data is collected from an open source data library online.

## **1.6 Report Layout**

The layout of the project report is simple which starts by introducing the concept of the business side of the research and the motivation behind it, with rational acknowledgment of the study. It declares the questions that are promised to be answered in this paper with proper acknowledgment of expected outcome and management both financially and physically. Then background research is discussed by explaining terminologies, related works, analysis, summary, future scope and challenges. Moving forward with the research methodology after explanation of the subject matter and introducing the instruments to be utilized with data utilization, statistical analysis, applied mechanism, and required implementation. Afterwards a complete analysis of the acquired results with a discussion and explanation for understanding. Concluding it with the social and environmental impact of the process and discussing its sustainability. In the end, there is a summary of the entire process with future implications and then a conclusion.

## **CHAPTER 2**

### **Background**

#### **2.1 Preliminaries/Terminologies**

This part introduces the terminologies which are referred to in this paper. In particular, it discusses the concept of CLV, RFM and weighted RFM models, Data mining methods, and Customer segmentation. First, we have customer relationship management (CRM), CRM is an enterprise approach to understanding and influencing customer behavior through meaningful communication to improve customer acquisition, customer retention, customer loyalty, and customer profitability. The goal of CRM is to develop closer and deeper relationships with customers and to maximize the lifetime value of a customer with an organization. The concept of Customer Lifetime Value (CLV) in CRM is utilized to evaluate and predict all future profits generated from a customer. RFM stands for Recency, Frequency, and Monetary Value. Recency value is the time interval of days of a customer's buying linking two of their purchases. It indicates that the customer repeatedly explores market of a specific company in a short period if there is smaller count of recency. Similarly, greater count indicates that the customers visit the company's market place less. Frequency is the number of purchases in a specific period made by a customer. The loyal the customers of the company have the higher the value of frequency. Monetary is the amount of money the customer spends in a certain duration or period. The company has more revenue from a customer who spends a higher amount of money. A weighted RFM is when There are dedicated weights to R, F, and M depending on the characteristics of the industry. Different weights should be assigned to RFM parameters which can also be known as RFM scores ranging from 5 to 1 for each recency, frequency, and monetary value parameters. In this paper, the main focus is on utilizing machine learning techniques. Machine learning is a field of computer science that teaches computers how to learn, so they can utilize their higher function in order to provide assistance to humans in solving difficult tasks. Typically, there are a number of machine learning techniques but here an unsupervised machine learning is applied to identify different groups which is known as clustering. An unsupervised learning is a type of machine learning that has minimum human supervision for previously undetected patterns in a data set with no pre-existing

labels. The process of grouping similar objects is declared as Clustering in machine learning. In order to create groups of customers with equivalent patterns of buying than those who have different patterns from the others is the main motive of clustering here. There are many methods for clustering like partitioning, hierarchical, density, and grid, etc. Here K-means clustering algorithm is used to assign customers a recency score.

## **2.2 Related Works**

A good number of papers had written about various methods for separating the customers in groups or segments. The RFM is a commonly tried model for analyzing the value of a customer. Many scholars have applied it to achieve customer segmentation. For past twenty years, countless prediction and classification models have been developed by several researchers considering RFM models. For example, Etzion et al. [1] grouped profitable customers and calculated the value for their lifetime with the company. Cui et al. [2] designed a model, for predicting response using variables from RFM. Cheng & Chen [3] discovers how to predict the loyalty of a customer using a model of data-mining. Additional literature includes RFM models integrated with clustering algorithms, actually related to the model which will use in this paper. He and Li [4] applied a new way to better improve lifetime of a customer (CLV) with contentment, and customer actions. Cho and Moon [5] applied a weighted frequent pattern mining in a customized recommendation system. Zahrotun [6] selected customer relationship management (CRM) using which found the best customer, from online data of the customers. Sheshasaayee and Logeshwari [7] developed a fresh approach of segmenting with the RFM and finding life time value of customers and integrating those methods. Ning Lu; Hua Lin; Jie Lu; Guangquan Zhang [8] predicted the churn of customers for a company.

## **2.3 Comparative Analysis and Summary**

The Comparative Analysis denoted in the related work section and their summary is discussed in this section. Investing more necessary resources on customers that add their investments to the company is the main focus of this approach. The suggested three-dimensional approach of He and Li [4] declares that consumers and their needs are different from one another. Cho and Moon [5] applied a weighted frequent pattern mining in a customized recommendation system. Here, the RFM model was used for customer

profiling to find potential customers. The application of CRM in buying goods from online, helps in detecting potential customers to increase profit for the company by segmenting. The K-means algorithm and K-medoids algorithm are both partitional approaches. The segmentation approach where RFM and customer with their life time is given a value, these two methods have two phases, where one phase is for statistical and another phase is clustering. A new distinct prediction model is created where regression of logistical values can be accomplished by isolating the transactional data and that is churn prediction.

## **2.4 Scope of the Problem**

The developed perceptions that were introduced in business and marketing due to the researchers named and discussed in the related work sections are explained in this section. Jiang and Tuzhilin [9] proposed in order to increase performance in marketing, both segmentation of customers and targeting of buyers are necessary. Multiple methods for segmentation of buyers has been utilized by a number of authors after this. In order to find the demand and expectations of customers for providing a good service, the assistance of segmentation is required. Cho and Moon [5] applied a weighted frequent pattern mining in a customized recommendation system. To increase the profit of the company best suggestions for the customer is provided by using an RFM model. In an appropriate marketing strategy, the customer relationship management method is useful for offering customers new and improved facilities in multiple categories specific to their requirements with customer data from online. The newly discovered algorithm used for clustering which is similar to K-means and medoids algorithms has less execute time than the standard methods with the increase of clusters numbers. There are plans to use a neural network for enhancing the segmentation with implemented two-phase model before K-means. For churn prediction it is observed that utilizing individual marketing strategies the customers which has maximum churn score value is possible to retain if identified properly.

## **2.5 Challenges**

The challenges faced along the way of developing the models designed to help business companies progress in previous sections are discussed here. Jiang and Tuzhilin [9] proposed for better understanding marketing performances, both segmentation and targeting based on customers and buyers are necessary respectively. With a step-by-step

©Daffodil International University

approach, but the problem was in terms of optimization. The author proposed K-Classifiers Segmentation algorithm to solve the problem. Cho and Moon [5] applied a weighted frequent pattern mining in a customized recommendation system. The challenge was to apply unique weights to each value of transaction in order to generate association rules in weighted format with the help of mining. In the customer relationship management method fuzzy c-means clustering is applied for performing segmentation based on customers but there is a delay in time due to multiple iterations. The proposed algorithm by Shah and Singh [10] have reduces the cluster error criterion but fails to provide any optimal solution for any instance. The method designed by Sheshasaayee and Logeshwari [7] is also not optimized enough. The churn prediction is an experimental implementation with no practical approach.

## **CHAPTER 3**

### **Research Methodology**

#### **3.1 Research Subject and Instrumentation**

The subject matter of the project is based on the study of customer segmentation using RFM analysis with the implementation of an Unsupervised Machine Learning model which is later going to be utilized in predicting the next purchase date of a customer. The Instruments that were utilized in accomplishing the outcome of this research are listed below,

##### **Hardware**

1. A Personal Computer
2. Intel Core i3 Processor
3. 8 GB RAM
4. Motherboard
5. 500GB HDD
6. 21-inch Monitor
7. Mouse
8. Keyboard

##### **Software**

1. Jupyter Notebook
2. Python 3.8.5

#### **3.2 Data Collection Procedure/Dataset Utilized**

The data set is collected and utilized from open-source database integrated in Kaggle.com, from the link listed below

<https://www.kaggle.com/vijayuv/onlineretail>

The dataset consists eight-months of detailed purchases for a specific online retail company stationed in the UK. The data is imported within the CSV file, in order to make the date field workable it is converted from string to Date Time. Also, it has been filtered properly of all the other countries other than the UK. In order to build a model, the acquired dataset is to be split into two parts. A behavioral data of six-months is used firstly to train in order

to predict the date of customer's very first buying in the upcoming period of three months. The prediction will calculate if there is no purchase too.

### 3.3 Statistical Analysis

The statistical analysis is represented using the graphical figures.

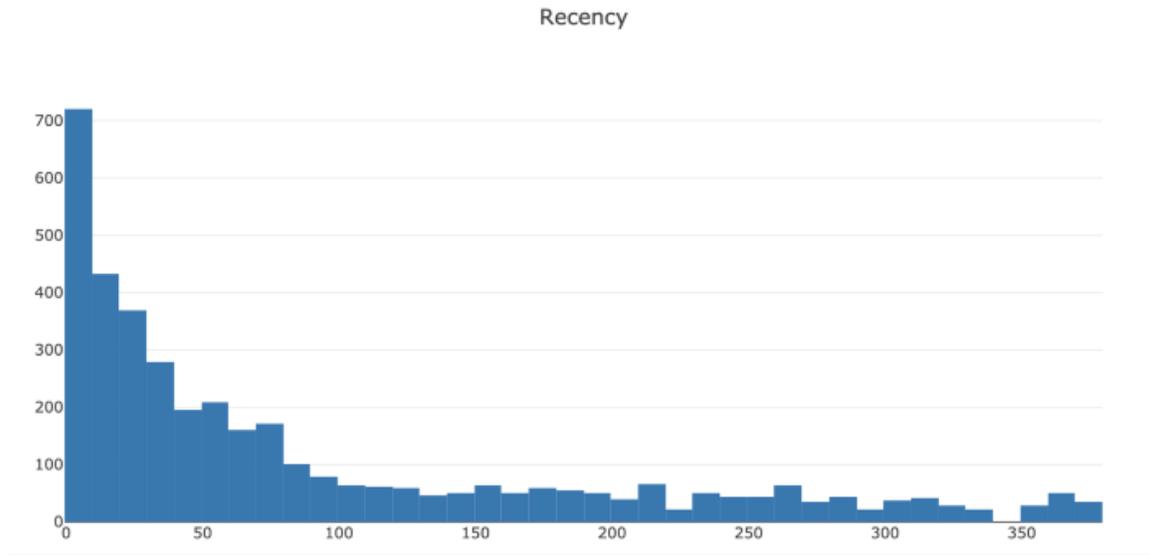


Fig 3.1: Recency Histogram

In this figure recency of customers is displayed with a plotted histogram, where it is charted by Last purchase day vs Number of inactive days prior to last purchase.

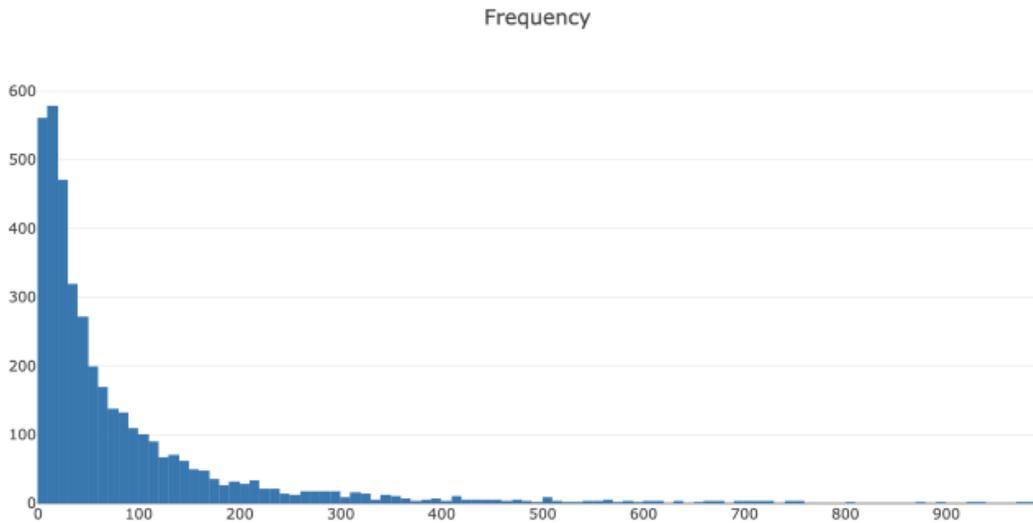


Fig 3.2: Frequency Histogram

In this figure Frequency of customers is displayed with a plotted histogram, where it is charted by Number of orders vs Time interval between each order.

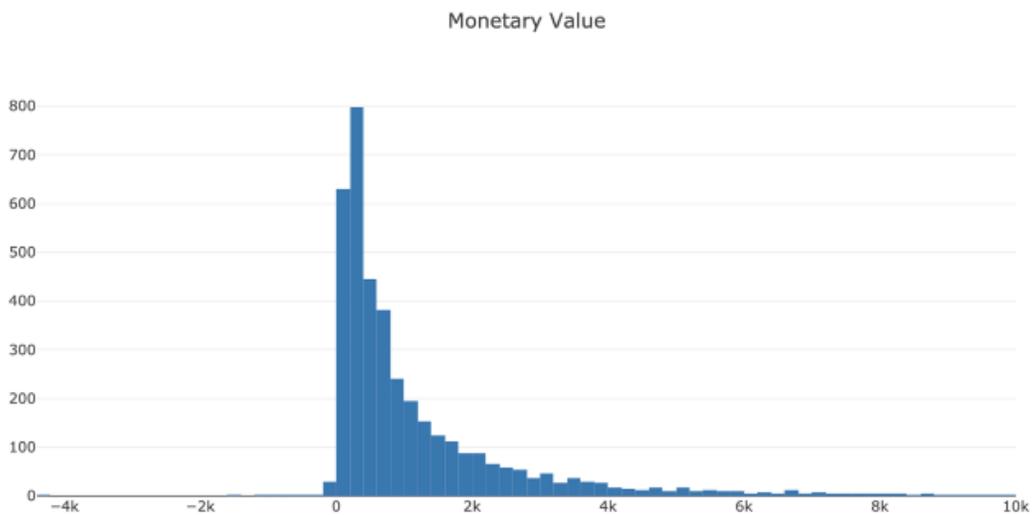


Fig 3.3: Monetary Value Histogram

In this figure Monetary value of customers is displayed with a plotted histogram, where it is charted by Number of customers vs Total amount spent.

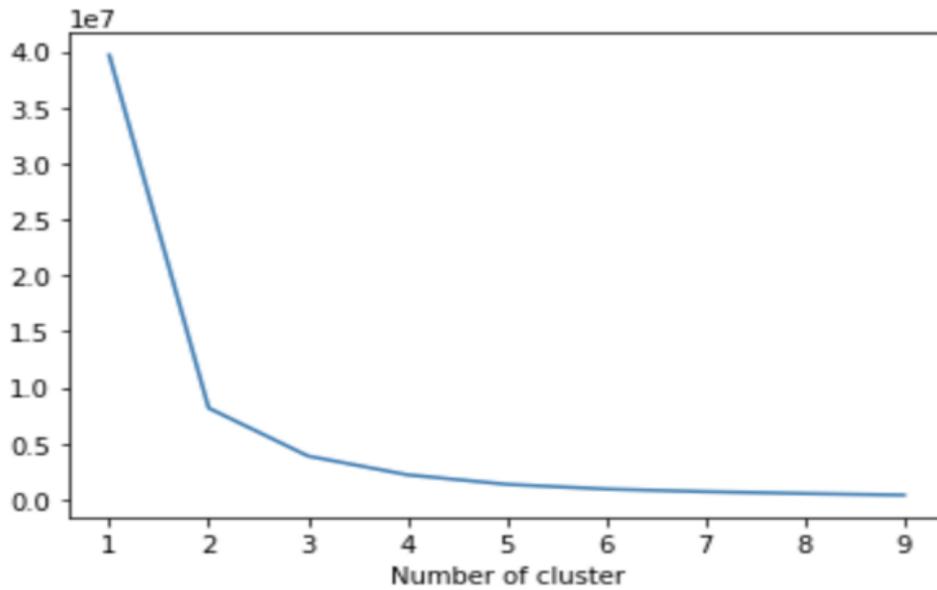


Fig 3.4: Recency Inertia Graph

In this figure a curve is shown for Recency of customers and is charted by Last purchase day vs Number of inactive days prior to last purchase.

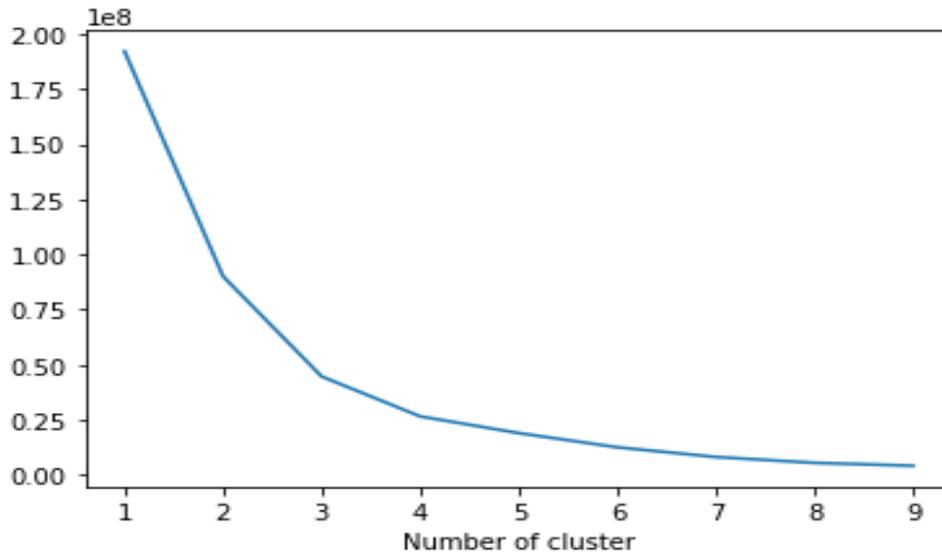


Fig 3.5: Frequency Inertia Graph

In this figure a curve is shown for Frequency of customers and is charted by Number of orders vs Time interval between each order.

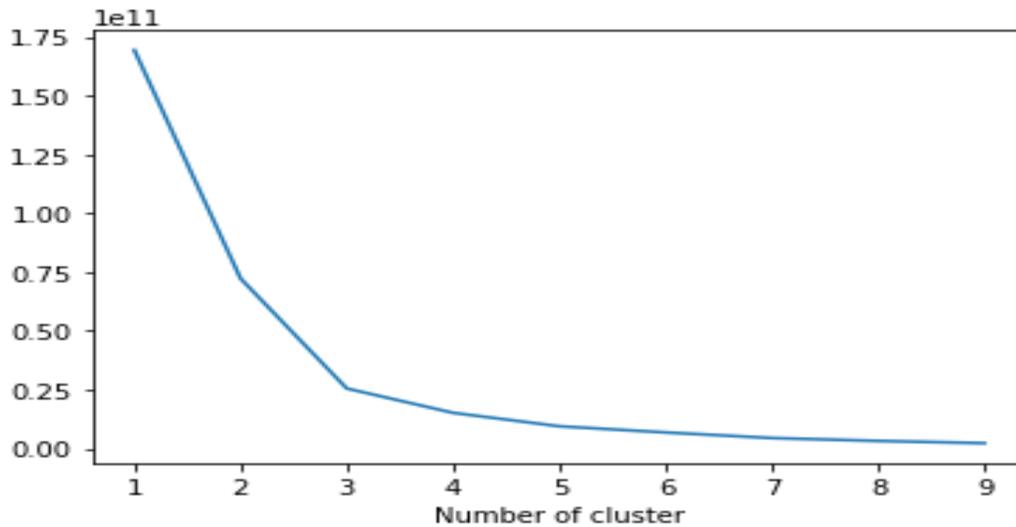


Fig 3.6: Monetary value Inertia Graph

In this figure a curve is shown for Monetary value of customers and is charted by Number of customers vs Total amount spent.

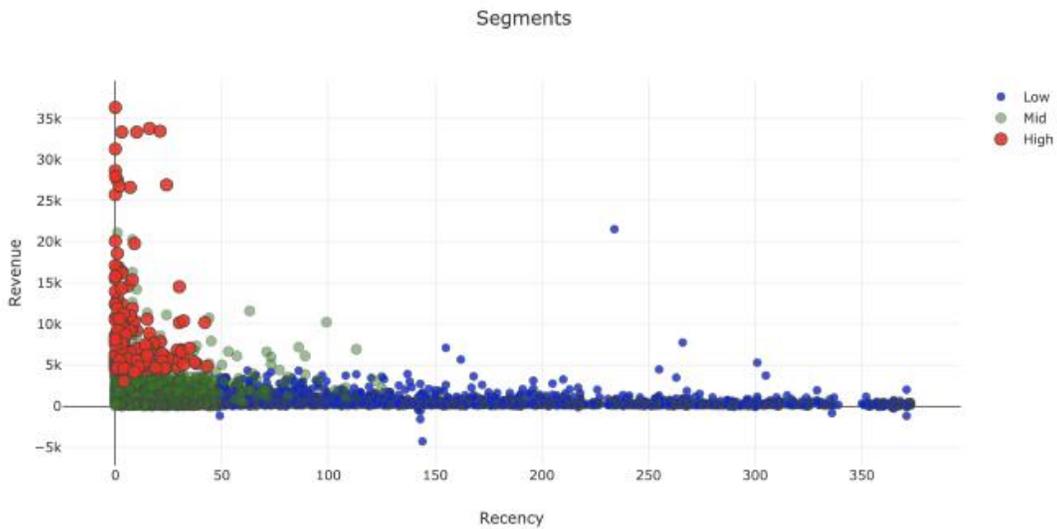


Fig 3.7: Recency Scatter Plot

In this figure cluster of high to low value customers for Recency are shown and is charted by Last purchase day vs Number of inactive days prior to last purchase.

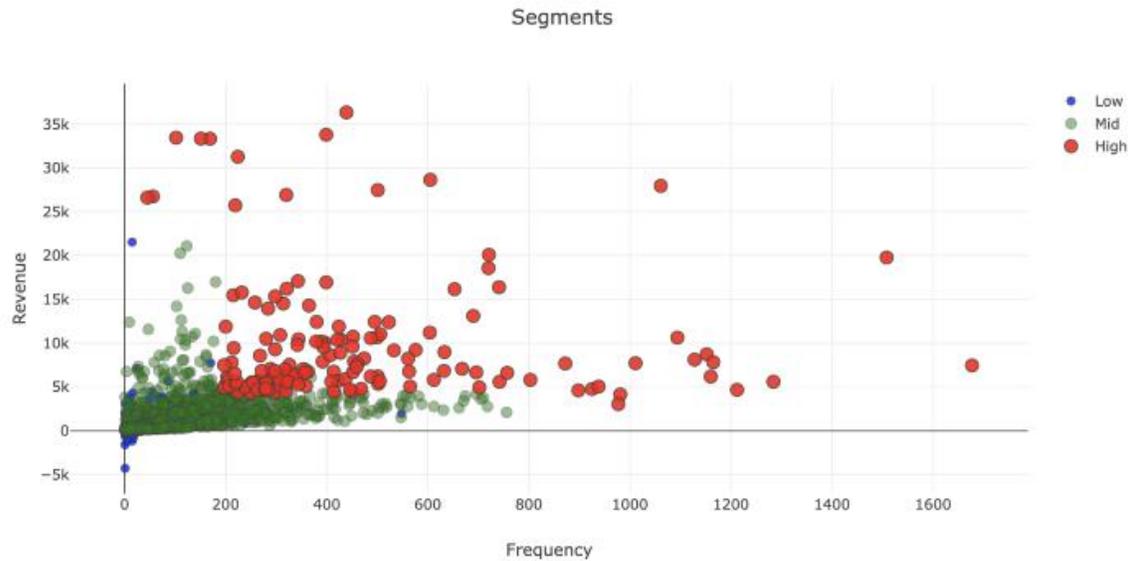


Fig 3.8: Frequency Scatter Plot

In this figure cluster of high to low value customers for Frequency are shown and is charted by Number of orders vs Time interval between each order.

### 3.4 Proposed Methodology/Applied Mechanism

Before focusing on customer segmentation, a question might arise as, why it is necessary to segment the customers? Because it is not possible to treat every customer the same, the customers will shift to a better option. Customers with experience on being available over various platforms will develop new tastes and that might grow into their personality to adapt. It is wise to adapt different initiatives in order to counter such. But there are very standard and effective segmentation procedures available. Here, implementation is done using one of them to a business. RFM,

- Low Value: fewer active customers, very less frequent buyer/visitor the others and add very small or zero or sometimes even negative revenue.
- Mid Value: These are the grey area customers. They invest not too much and not too high and are frequent and add decent revenue.
- High Value: They produce high revenue in a company are frequent and hardly unavailable. The kind of not worth loosing.

The proposed methodology can be broadly divided into 4 steps. Step-1 is calculating Recency, Frequency, and Monetary Value, Step-2 is applying a machine learning algorithm

that is unsupervised for detecting unique groups or clusters for each class segment, Step-3 is selecting a Machine Learning Model, and Step-4 is building and running the machine learning model. The analysis done in RFM (recency, frequency, and monetary values) model divides customers that are important on the basis of consumption of service by customers, the frequency of their visit and the amount of investment they introduce to the company. These act as the three variables which are separated from a huge chunk of database. Thus, the loyalty value of customers is obtained from recency, frequency and monetary value of customers. Hence considering Recency, Frequency, and Monetary value as the three criteria for getting the loyalty value of customers is the RFM model. The elaborated definitions would be,

Recency of last purchase (R): The interval between the latest buying and the present time of a customer is represented by it. The recency value increases if the interval value decreases.

Frequency of purchase (F): The time period of buying for a customer within a particular interval like twice in a week, once in a year is represented by Frequency. The value of F increase if the number of transactions within a specific interval increase too.

Monetary value of the purchase (M): The monetary value is the purchases value of a service or a product in a specific time period. The value of M increases with the value of monetary.

At first, to calculate recency it is necessary to figure out each customer's date of purchase that was made recently also counting their days of inactivity. For creating the variable of recency, it is decided to use a reference date of the day before the previous transaction date. The process will be initiated using the dataset of the online retail industry. After recent transactions are check, with the procedure frequency, and total amount of expense is calculated for customers. Here K-means clustering is applied to assign recency, frequency, and monetary value scores. It is the standard algorithm for clustering which takes the parameters and the number of clusters as inputs and segments the information into the characterized number of clusters such that the intra-cluster likeness is maximum. It is required to determine; how many clusters are needed in the K-means algorithm. To discover it out, elbow method is connected. Elbow method essentially tells the ideal cluster number for ideal inactivity. The K-means algorithm for recency frequency, and the amount spent calculation,

Input:

Dataset of customers which includes 'n' entries

k: No. of clusters

Output:

Divided data of customers to k clusters

Algorithm:

1. At first, depending on the esteem of k, k arbitrary focuses are chosen as beginning centroids.
2. The separations of each information point from the centroids chosen prior are assessed utilizing the Euclidian removal.
3. The separate values are compared and the information point is allotted to the centroid which has the most limited Euclidian separate esteem.
4. The past steps are rehashed. The method is halted in case the clusters gotten are the same as that of the past step.

Every customer gets three unique scores depending on the variable's recency, frequency, and monetary. On a range of 5 to 1 scoring is determined. The top customer is the one with a score of 5, and the rest are 4, 3, 2, and 1. Each score has its own unique characteristics. Within the RFM analysis client division is done by, to begin with sorting clients based on their Recency value that's the foremost later will be at the best, at that point with the Recurrence esteem with the foremost visit at the best, and at long last the money related esteem with the most noteworthy financial esteem at the best. The clients are subsequently part into five quintiles with the best 20% having a score of 5 another 20% having a score 4 and so on. The method is rehashed for all three criteria and at last, the values are consolidated to induce each person client rank. The clients are relegated to diverse quintiles and how their values can be blended to urge client values. Before jumping into the selection machine learning the model, there are two necessary actions required. First, identifying the classes in the label. For both statistics and business needs, it is necessary to decide the boundaries first. In terms of the first one, it should make sense and be simple to require activity and communicate. Considering these two, there will be three classes: 0–20: Clients that will buy in 0–20 days — Lesson title: 2 21–49: Clients that will buy in 21–49 days — Lesson title: 1  $\geq$  50: Clients that will buy in more than 50 days — Lesson title: 0 The final step is to see the relationship between our highlights and name and showed in

statistical analysis. The model which gives the highest accuracy should be the required model for this particular problem. Now applying a principal concept in Machine Learning, which is Cross-Validation. The stability of different machine learning models across different datasets is determined. The including models to be tested are Logistic Regression, Random Forest Classifier, Gaussian Naive Bayes, Decision Tree Classifier, Extreme Gradient Boosting Classifier (XGBoost), and K Neighbors Classifier. Let's part prepare and test tests and degree the exactness of diverse models. It gives the score of each show by selecting diverse test sets. In case the deviation is moo, it implies the demonstrate is steady. In this case, the deviations between scores are satisfactory but for the Decision Tree Classifier. The proposed model for the project would be seen in the next page.

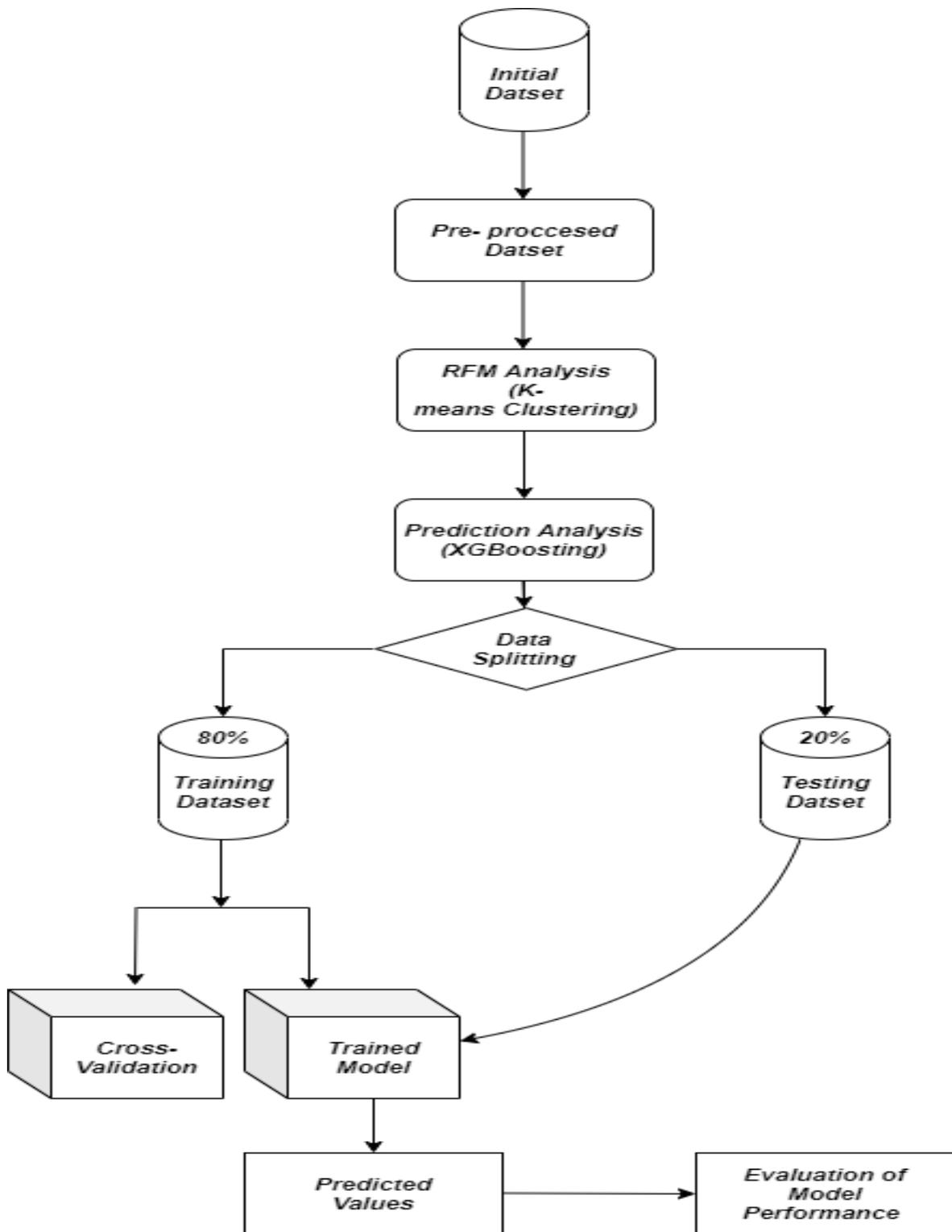


Fig 3.9: Proposed Model

### **3.5 Implementation Requirements**

For the implementation process it is important to import necessary libraries. The required libraries for such procedure are,

Imported libraries for data

- Datetime (Date and Time Manipulation Library)
- Pandas (Data Manipulation and Analysis Library)
- Matplotlib (Mathematic Plotting Library)
- Numpy (Numerical Python Library)
- Seaborn (Drawing Visualization Library)
- Division (Future Statement Library)
- KMeans (Clustering Library)
- Plotly (Visualization Library)

## CHAPTER 4

### Experimental Results and Discussion

#### 4.1 Experimental Setup

The proposed strategy is assessed by performing it on the value-based information set of online retail store clients for one year is gotten from the University of California Irwin (UCI) store. Customer division is displayed in this segment in a step-by-step preparation. The dataset comprises of eight properties counting the client ID, item code, item title, the cost of the item, date and time of buy, etc. The first information set comprises of 18,267 occurrences with eight traits. The dataset contains the buy of data from 1-12-2010 to 09-12-2011 of the clients. The occurrences with lost values in vital traits, unit cost, and amount less than 0, and the date surpassing the current date are all expelled amid information pre-processing. The significant occurrences such as receipt date and time, the amount of item per exchange, item cost per unit concerning recency, money related, and recurrence are sifted, and as it were those records have been inputted into the benchmark calculations. The altered dataset contains 772 occasions with three extra qualities recency, recurrence, and money related determined from RFM calculation. The process is carried out as usual, to calculate recency, it is watched to take note of the foremost later buy date of each client with how numerous days they are dormant. After having no. of inert days for each client, K-means clustering is connected to allot clients a recency score.

Table 4.1: Recency Score Table

	<b>CustomerID</b>	<b>Recency</b>
<b>0</b>	17850.0	301
<b>1</b>	13047.0	31
<b>2</b>	13748.0	95
<b>3</b>	15100.0	329
<b>4</b>	15291.0	25

It is observed that the median is 49, even though the average is 90-days recency. From the inertia graph it is observed that 3 is the optimal one. Since it is possible to proceed with

higher or lower number of clusters based on business requirements, lets determine the possible number of clusters to be 4.

Table 4.2: Characteristics of Recency Cluster

	count	mean	std	min	25%	50%	75%	max
<b>RecencyCluster</b>								
<b>0</b>	568.0	184.625000	31.753602	132.0	156.75	184.0	211.25	244.0
<b>1</b>	1950.0	17.488205	13.237058	0.0	6.00	16.0	28.00	47.0
<b>2</b>	478.0	304.393305	41.183489	245.0	266.25	300.0	336.00	373.0
<b>3</b>	954.0	77.679245	22.850898	48.0	59.00	72.5	93.00	131.0

K-means relegate clusters as numbers but not in a requested way. it isn't sensible to announce cluster is the most noticeably awful and cluster 4 is the most excellent. Applying the same prepare for Frequency and Revenue. To calculate frequency, it is determined with the entire number of orders for each client. To form frequency clusters, to begin with, calculate the frequency, and notice how that appears within the client database.

Table 4.3: Characteristics of Frequency

	count	mean	std	min	25%	50%	75%	max
<b>FrequencyCluster</b>								
<b>0</b>	3496.0	49.525744	44.954212	1.0	15.0	33.0	73.0	190.0
<b>1</b>	429.0	331.221445	133.856510	191.0	228.0	287.0	399.0	803.0
<b>2</b>	22.0	1313.136364	505.934524	872.0	988.5	1140.0	1452.0	2782.0
<b>3</b>	3.0	5917.666667	1805.062418	4642.0	4885.0	5128.0	6555.5	7983.0

As the same documentation as recency clusters, the high-frequency number shows way better clients. Let's watch how the client database looks like when it is clustered based on income. Revenue for each client is calculated, applying the same clustering strategy.

Table 4.4: Characteristics of Revenue Cluster

	count	mean	std	min	25%	50%	75%	max
<b>RevenueCluster</b>								
0	3688.0	908.182672	923.507907	-4287.63	263.3325	572.685	1258.675	4330.67
1	233.0	7775.420687	3638.011093	4345.50	5178.9600	6568.720	9167.820	21535.90
2	27.0	43070.445185	15939.249588	25748.35	28865.4900	36351.420	53489.790	88125.38
3	2.0	221960.330000	48759.481478	187482.17	204721.2500	221960.330	239199.410	256438.49

From the acquired scores i.e. cluster numbers an overall score is generated for recency, frequency & revenue.

Table 4.5: RFM Score Table

	Recency	Frequency	Revenue
<b>OverallScore</b>			
0	304.584388	21.995781	303.339705
1	185.362989	32.596085	498.087546
2	78.972856	47.060803	871.842586
3	20.662252	68.374172	1089.271213
4	14.892617	271.755034	3607.097114
5	9.662162	373.290541	9136.946014
6	7.740741	876.037037	22777.914815
7	1.857143	1272.714286	103954.025714
8	1.333333	5917.666667	42177.930000

The scoring over clearly appears that clients with score 8 is the finest clients while is the worst. To keep things straightforward, way better to title these scores:

Low Value: 0 to 2

Mid Value: 3 to 4

High Value: 5+

Now all this can be identified as a feature and it is possible to include new features here as such determining the days of purchase between last 3 purchases, and difference between

purchase days with mean and standard deviation. After that, a unused column is made which incorporates the dates of the final 3 buys and to watch how the information outline looks like.

Table 4.6: Relation Table for Difference between Number of Days and Invoice Dates

	CustomerID	InvoiceDate	InvoiceDay	PrevInvoiceDate	T2InvoiceDate	T3InvoiceDate	DayDiff	DayDiff2	DayDiff3
<b>649</b>	12747.0	2011-03-01 14:53:00	2011-03-01	NaN	NaN	NaN	NaN	NaN	NaN
<b>65091</b>	12747.0	2011-05-05 15:31:00	2011-05-05	2011-03-01	NaN	NaN	65.0	NaN	NaN
<b>90473</b>	12747.0	2011-05-25 09:57:00	2011-05-25	2011-05-05	2011-03-01	NaN	20.0	85.0	NaN
<b>124699</b>	12747.0	2011-06-28 10:06:00	2011-06-28	2011-05-25	2011-05-05	2011-03-01	34.0	54.0	119.0
<b>184410</b>	12747.0	2011-08-22 10:38:00	2011-08-22	2011-06-28	2011-05-25	2011-05-05	55.0	89.0	109.0
<b>7326</b>	12748.0	2011-03-08 12:30:00	2011-03-08	NaN	NaN	NaN	NaN	NaN	NaN
<b>10606</b>	12748.0	2011-03-11 11:37:00	2011-03-11	2011-03-08	NaN	NaN	3.0	NaN	NaN
<b>17545</b>	12748.0	2011-03-18 13:08:00	2011-03-18	2011-03-11	2011-03-08	NaN	7.0	10.0	NaN
<b>20123</b>	12748.0	2011-03-21 15:40:00	2011-03-21	2011-03-18	2011-03-11	2011-03-08	3.0	10.0	13.0
<b>24764</b>	12748.0	2011-03-24 13:37:00	2011-03-24	2011-03-21	2011-03-18	2011-03-11	3.0	6.0	13.0

Presently in this arrangement, there's a prerequisite for making an extreme choice. The calculation over is very valuable for clients who have numerous buys. But the same can't be decided for the ones with 1–2 buys. For occurrence, it is as well early to tag a client as visit who has as it were 2 buys but back to back. The final step is to see the relationship between our highlights and label. The correlation network is one of the cleanest ways to display this.

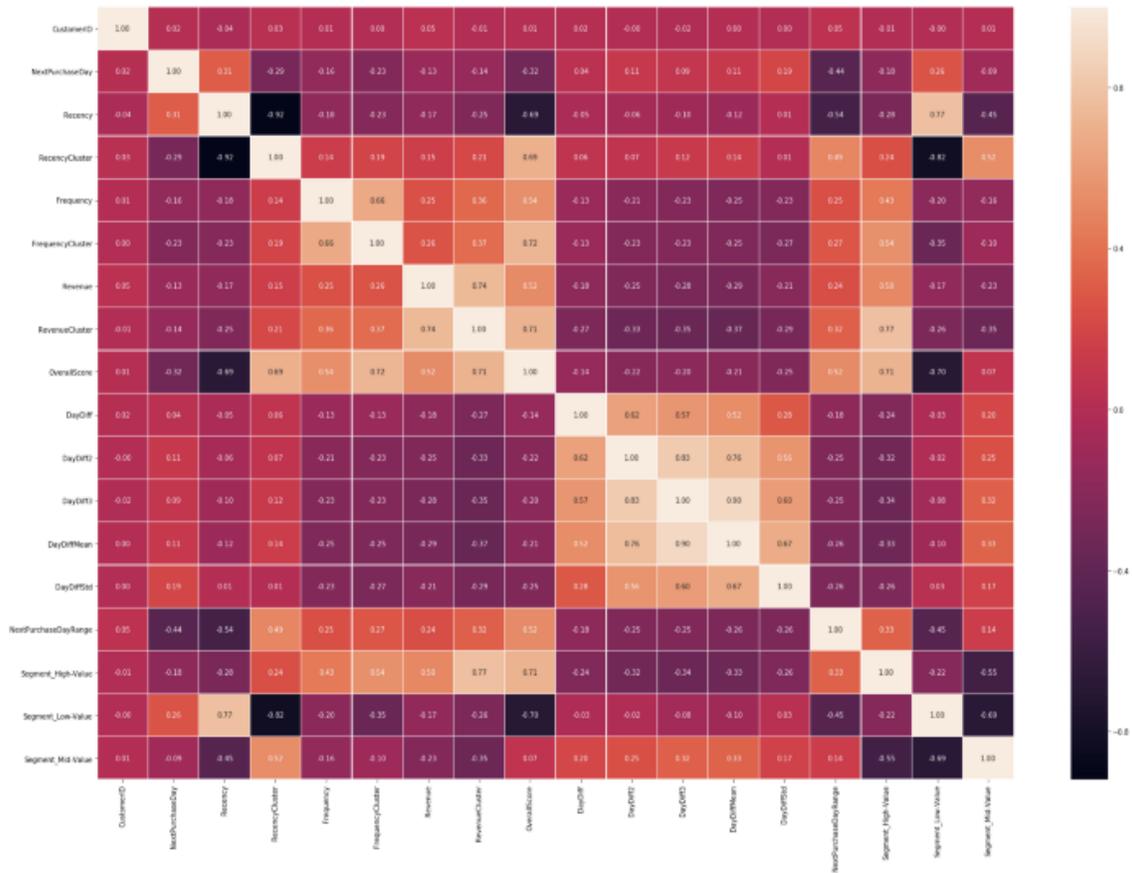


Fig 4.1: Correlation Matrix

## 4.2 Experimental Results & Analysis

The Process is carried out as usual, by comparing the exactness of comes about from diverse classification strategies like Naive Bayes, Decision tree, Neural Network, etc on the same set of information. It was decided to use Extreme Gradient Boosting for improving it further, with the application of Hyperparameter Tuning which resulted in increased scores in the beginning the accuracy test of the extreme gradient boosting was set at 58%. But later the score expanded from 58% to 62%, which is a very big change. Also, it is possible to use a distinctive number of classes on the yield to see on the off chance that the precision changes by how much sum in case of clustering by utilizing K-means. Predictive analytics helps us to provide many opportunities such as predicting the next purchase day of the customer. Knowing the next buying day could be a great pointer for anticipating deals as well. Also, it

becomes easier to determine the next course of action with the help of customer segmentation. The best procedures are very clear:

High Esteem: Move forward Retention

Mid Esteem: Move forward Maintenance + Increment Frequency

Low Esteem: Increment Frequency

### **4.3 Discussion**

The study is directed towards, the RFM segmentation model for customer and is suggested to a company that functions in the online retail industry in UK. The company has already created a grouping of customers according to customer's expense. The delivered segmentation of customers is by using recency, frequency, and monetary as identifiers while clustering the customers was implemented with the k-means clustering method. The suggested model meets the requirement for the clusters which are completely unique from the available clusters. The company can declare the customers as optimum customers due to obtaining close to average scores in RFM scoring. Otherwise, the company can choose for not sending any new offers to any of these customers, as they have only bought from the company once.

## **CHAPTER 5**

### **Impact on Society, Environment and Sustainability**

#### **5.1 Impact on Society**

The study is based on the customers acquired by a company of an online retail store. Every customer there is a member of the existing society. It is possible to impact the customers social life style based on the study because the prediction of behavior is common in variety field of interests. But the final outcome is completed based on an economical growth. Application of such studies might lead to better trade opportunities in upcoming future and thus might be effective to bring about a data driven growth in the current economical setup.

#### **5.2 Impact on Environment**

This study does not impact the physical environment rather than the social environment. There are possibilities to develop better social understanding and grouping people in day to day life based on this study.

#### **5.3 Ethical Aspects**

This study is ethically safe due to the fact that this study does not violate any sorts of privacy matter. Also, the dataset is collected from an open source to avoid privacy laws.

#### **5.4 Sustainability Plan**

Further study is required to develop a sustainability plan.

## **CHAPTER 6**

### **Summary, Conclusion, Recommendation and Implication for Future Research**

#### **6.1 Summary of the Study**

The study in this paper presents incorporating RFM analysis into customer segmentation techniques to provide market intelligence. The aim is to bring the attention of data miners and marketers to the importance and advantages of using RFM analysis for customer segmentation in order to evaluate the proposed model and empirically demonstrate the benefits of using this model in direct marketing.

#### **6.2 Conclusions**

There is a vital role in customer segmentation for retail companies. Better segmenting of customers is pivotal in reaching a company's sales target. Companies get a better understanding of the target market if the customers that have equivalent requirements, necessities and behavior are grouped together. Thus, companies could reevaluate the current course of action and develop a new method for better sales, such as; update marketing, price management, promotions, building extra customer touchpoints, etc.

#### **6.3 Implication for Further Study**

A case study was carried out using the datasets collected within two years period by a sports store in Turkey through its e-commerce website. According to experimental study results, the proposed approach provides better product recommendations than simple recommendations, by considering several parameters together: the customer's segment, the current RFM values of the customer, potential future customer behavior, and products frequently purchased together.

## REFERENCES

- [1] Etzion, O., Fisher, A., & Wasserkrug, S. (2004, Walk). e-CLV: a modeling approach for client lifetime assessment in e-commerce spaces, with an application and case consider for online barbers. In innovation, e-Commerce, and e-Service, 2004. IEEE'04. 2004 IEEE Worldwide Conference on (pp. 149-156). IEEE.
- [2] Cui, G., Wong, M. L. & Lui, H. K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, 52(4), 597-612.
- [3] Chen, Y. S., Cheng, C. H., Lai, C. J., Hsu, C. Y., Syu & H. J. (2012). Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment. *Computers in Biology and Medicine*, 42(2), 213-221.
- [4] He X., Li, C., 2016. The research and application of customer segmentation on e-commerce websites. In: 2016 6th International Conference on Digital Home (ICDH), Guangzhou, pp. 203–208. doi :10.1109/ICDH.2016.050.
- [5] Cho, Young, Moon, S.C., 2013. Weighted mining frequent pattern-based customer's RFM score for personalized u-commerce recommendation system. *J. Converg.* 4, 36–40.
- [6] Zahrotun, L., 2017. Implementation of data mining technique for customer relationship management (CRM) on online shop tokodipers.com with fuzzy c-means clustering. In: 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, pp. 299–303.
- [7] Sheshasaayee, A., Logeshwari, L., 2017. Efficiency analysis of the TPA clustering methods for intelligent customer segmentation. In: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, pp. 784–788.
- [8] Lu, H., Lin, J.Lu., Zhang, G., May 2014. A customer churns prediction model in the telecom industry using boosting. *IEEE Trans. Ind. Inf.* 10 (2), 1659–1665. [https://doi.org/ 10.1109/TII.2012.2224355](https://doi.org/10.1109/TII.2012.2224355).
- [9] Jiang, T., Tuzhilin, A., March 2009. Improving personalization solutions through optimal segmentation of customer bases. *IEEE Trans. Knowledge Data Eng.* 21 (3), 305–320. [https:// doi.org/10.1109/ TKDE.2008.163N](https://doi.org/10.1109/TKDE.2008.163N).
- [10] Shah, S., Singh, M., 2012. Comparison of a Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid Algorithm. In: 2012 International Conference on Communication Systems and Network Technologies, Rajkot, pp. 435–437.

# Plagiarism Report

## Customer Segmentation using RFM Analysis

### ORIGINALITY REPORT

<b>17%</b>	<b>11%</b>	<b>14%</b>	<b>8%</b>
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

### PRIMARY SOURCES

<b>1</b>	<b>A. Joy Christy, A. Umamakeswari, L. Priyatharsini, A. Neyaa. "RFM Ranking – An Effective Approach to Customer Segmentation", Journal of King Saud University - Computer and Information Sciences, 2018</b>	<b>5%</b>
	Publication	
<b>2</b>	<b>s3.amazonaws.com</b>	<b>3%</b>
	Internet Source	
<b>3</b>	<b>home.sl.on.ca</b>	<b>3%</b>
	Internet Source	
<b>4</b>	<b>cdn.intechweb.org</b>	<b>2%</b>
	Internet Source	
<b>5</b>	<b>www.ijceas.com</b>	<b>1%</b>
	Internet Source	
<b>6</b>	<b>Submitted to TechKnowledge</b>	<b>1%</b>
	Student Paper	
<b>7</b>	<b>Sajjad Shokouhyar, Sina Shokoohyar, Sepehr Safari. "Research on the influence of after-sales service quality factors on customer satisfaction",</b>	<b>&lt;1%</b>