S	
E1	
E2	
For Office Use Only	

UCSC

Masters Project Final Report

(MCS)

2022

Project Title	Customer Segmentation Using Machine Learning
Student Name	S D Jayaratne
Registration No. & Index No.	2018/MCS/039 18440393
Supervisor's Name	Mr. K.P.M.K. Silva

For Office Use ONLY

Customer Segmentation Using Machine Learning

A dissertation submitted for the Degree of Master of Computer Science

S. D. Jayaratne

University of Colombo School of Computing

2022

DECLARATION

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: S.D. Jayaratne

Registration Number: 2018/MCS/039

Index Number: 18440393

Signature: Date: 17/11/2022

This is to certify that this thesis is based on the work of

~~Mr.~~/Ms. S.D. Jayaratne

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Mr. K.P.M.K. Silva

Signature: Date: 18/11/2022

I would like to dedicate this thesis to

my beloved family,

for their love, support, and encouragement

&

academic and non-academic staff of

University of Colombo School of Computing

for their support and guidance given

to make this thesis a success.

ACKNOWLEDGEMENT

I would like to take this opportunity to acknowledge the limitless support & guidance given by the project supervisor, Mr. K.P.M.K. Silva who guided my efforts without any bounds to make this thesis a success.

Also, I would like to thank the module coordinators of module MCS3204 for the continuous guidance given throughout the project duration. My special thanks to the University of Colombo School of computing for giving me the opportunity to carry out this research project.

Last but not least, I would like to thank all those who like to remain anonymous though, the support provided to me was valuable and much appreciated.

ABSTRACT

In the face of huge competition among business organizations and with the modern economy, organizations have a huge number of customers hence, it is required to mine customer resources in order to achieve targeted measures for different types of customers and provide them with the services they want. Since it is a complex task to treat each and every customer separately, this can be achieved through customer segmentation by grouping the customer base into refined customer groups based on their similar needs and behaviors. This is very crucial to the development of the organization as it enables to improve their customer satisfaction further the implementation of customer segmentation leads to gaining new customers and this will be beneficial to extract a higher value from the existing customers through maintaining a better customer relationship. When the segmentation system is efficiently designed, customers of one segment have similar interests and behaviors, and they will most probably respond similarly to the situations where the elements of the marketing mix for example pricing, promotions, and for sales channels. This will be very significant for financial organizations to improve profit-driving opportunities targeting each unique customer group.

This research project focuses on a banking sector dataset and this study explores multiple machine learning models for segmenting customers and for identifying the most valuable customer group according to the customer payment behaviors. I have used a hybrid approach utilizing both supervised learning model and unsupervised machine learning model in this study. The banking dataset was analyzed and processed in order to train the machine learning models. The customer base was segmented into four customer segments and each customer group was analyzed to recognize the most valuable customer group. And the output of the trained clustering model was used to develop the customer segmentation prediction system using supervised machine learning models to predict the customer group of the user input customer. The customer dataset was trained using six different unsupervised machine learning algorithms and the obtained customer segments from the best-performance machine learning model were used for training the prediction model supervised machine learning algorithms were trained on the clustered dataset and the best-performing model was selected to build the prediction model. The prediction model guarantees an accuracy of 0.97 along with the other performance metrics.

Keywords: machine learning, customer segmentation, clustering, classification.

TABLE OF CONTENTS

DECLARATION i
ACKNOWLEDGEMENT iii
ABSTRACT iv
LIST OF FIGURES vii
LIST OF TABLES x
ABBREVATIONS xi
CHAPTER 1: INTRODUCTION 1
1.1 Statement of the Problem 2
1.2 Motivation 3
1.3 Research Aims and Objectives 3
1.3.1 Aims 3
1.3.2 Objectives 4
1.4 Scope 5
1.5 Structure of the Thesis 5
CHAPTER 2: LITERATURE REVIEW 6
2.1 A Literature Review 6
2.2 Research Gap 18
2.3 Presentation of Scientific Material 20
CHAPTER 3: METHODOLOGY 26
3.1 Research understanding phase 27
3.2 Data understanding phase 27
3.3 Data Preprocessing 38
3.3.1 Handling Missing Values 39
3.3.2 Outlier Detection 41
3.3.3 Normalization and Standardization 42
3.3.4 Feature Selection 43
3.3.5 Dimensionality Reduction 43
3.3.5 Encoding 43
3.4 Modeling 44
3.4.1 Customer Segmentation Model Building 44
3.4.2 Customer Segment Prediction Model Building 53
CHAPTER 4: EVALUATION AND RESULTS 56
4.1 Evaluation of the Results 56

4.1.1 Correlation Matrix with Heat Map 56
4.1.2 PCA Analysis 58
4.1.3 Clustering Model Analysis 61
4.1.4 Cluster Results Evaluation and Interpretation 76
4.1.5 Prediction Model Evaluation 88
4.2 Customer Segmentation Prediction System 94
CHAPTER 5: CONCLUSION AND FUTURE WORK 98
5.1 Limitations in the Research Study 100
5.2 Future Work 100
APPENDICES I
REFERENCES: II

LIST OF FIGURES

Figure 1: The process of marketing strategy for banks and other financial institutions. Source: (Kotler and Armstrong, 2010) ... 8
Figure 2: The process of segmentation for financial institutions ... 9
Figure 3: Customer Value Matrix, Source: (Marcus, C., 1998) ... 11
Figure 4: The clustering process ... 14
Figure 5: The CRISP-DM Process, Source: (Chapman et al. ,2000) ... 26
Figure 6: Information of the data columns of customer dataset ... 28
Figure 7: Balance Distribution ... 30
Figure 8: Balance Frequency Distribution ... 30
Figure 9: Purchases Distribution ... 31
Figure 10: One-off Purchases Distribution ... 31
Figure 11: Installments Purchases Distribution ... 32
Figure 12: Cash Advance Distribution ... 32
Figure 13: Purchases Frequency Distribution ... 33
Figure 14: One-off Purchases Distribution ... 33
Figure 15: Purchases Installments Distribution ... 34
Figure 16: Cash Advance Frequency Distribution ... 34
Figure 17: Cash Advance Trx Distribution ... 35
Figure 18: Purchases Trx Distribution ... 35
Figure 19: Credit Limit Distribution ... 36
Figure 20: Payments Distribution ... 36
Figure 21: Minimum Payments Distribution ... 37
Figure 22: Prc Full Payment Distribution ... 37
Figure 23: Tenure Distribution ... 38
Figure 24: Missing Data visualization – Seaborn heatmap ... 40
Figure 25: Missing value count ... 40
Figure 26: Outlier Detection ... 41
Figure 27: Normalization and Standardization ... 42
Figure 28: Classic K-means algorithm ... 45
Figure 29: Spectral Clustering vs K-Means Clustering ... 47
Figure 30: Gaussian Distributions ... 49
Figure 31: DBSCAN Clustering ... 50
Figure 32: BIRCH Overview ... 51
Figure 33: Approach for building a data-driven prediction model ... 52
Figure 34: Decision Tree, learning sine curve ... 54
Figure 35: General design architectural Diagram of the proposed system ... 55
Figure 36: Correlation Matrix with Heat map ... 57
Figure 37: Correlation matrix plot for component loadings ... 58
Figure 38: Eigenvalues; Variance explained by each PC ... 59
Figure 39: Cumulative proportion of variance (from PC1 to PC17) ... 59
Figure 40: Scree Plot ... 59
Figure 41: PCA Variance graph ... 60
Figure 42: Elbow Plot ... 61
Figure 43: Silhouette score Elbow plot ... 62
Figure 44: Cluster result obtained through K-Means ... 62

Figure 45: Customer data distribution among clusters - K-Means clustering 63
Figure 46: 3D plot of cluster visualization(KMeans) 63
Figure 47: Dedrogram 64
Figure 48: Silhouette score of Agglomerative clustering 65
Figure 49: Cluster result obtained through Agglomerative clustering 66
Figure 50: Customer data distribution among clusters - Agglomerative clustering 66
Figure 51: Silhouette score of Spectral clustering 67
Figure 52: Cluster result obtained through Spectral clustering(rbf) 68
Figure 53: Customer data distribution among clusters - Spectral clustering (rbf) 68
Figure 54: Cluster result obtained through Spectral clustering(nearest_neighbors) 69
Figure 55: Customer data distribution among clusters - Spectral clustering (nearest_neighbors) 69
Figure 56:Silhouette score of Spectral clustering 70
Figure 57: Cluster result obtained through GMM based clustering 70
Figure 58: Customer data distribution among clusters - GMM based clustering 71
Figure 59: 3D plot of cluster visualization(GMM) 71
Figure 60: K-Distance Graph 72
Figure 61: Cluster result obtained through DBSCAN clustering 73
Figure 62: Customer data distribution among clusters - DBSCAN clustering 73
Figure 63: Silhouette score of BIRCH clustering 74
Figure 64: Cluster result obtained through BIRCH clustering 74
Figure 65: Customer data distribution among clusters – BIRCH clustering 75
Figure 66: 3D plot of cluster visualization(BIRCH) 75
Figure 67: Plot comparison of Balance of each cluster 77
Figure 68: Plot comparison of Balance_Frequency of each cluster 77
Figure 69: Plot comparison of Purchases of each cluster 78
Figure 70: Plot comparison of One-off Purchases of each cluster 78
Figure 71: Plot comparison of Installments_Purchases of each cluster 78
Figure 72: Plot comparison of Cash Advance of each cluster 78
Figure 73: Plot comparison of Purchases_Frequency of each cluster 79
Figure 74: Plot comparison of Purchase_Installments_Frequency of each cluster 79
Figure 75: Plot comparison of One-off_Purchases_Frequency of each cluster 79
Figure 76: Plot comparison of Cash_Advance_Frequency of each cluster 79
Figure 77: Plot comparison of Cash_Advance_Trx of each cluster 80
Figure 78: Plot comparison of Purchases_Trx of each cluster 80
Figure 79: Plot comparison of Purchases_Tr 80
Figure 80: Plot comparison of Credit_Limit of each cluster 80
Figure 81: Plot comparison of Minimum_Payments of each cluster 81
Figure 82: Plot comparison of Pcc_Full_Payment of each cluster 81
Figure 83: Plot comparison of Tenure of each cluster 81
Figure 84: Feature Importance of Customer Data attributes 83
Figure 85: KPI Features 83
Figure 86: Snake plot of KPI Feature distribution among clusters 84
Figure 87: Bar graph of cluster interpretation 84
Figure 88: pair plot of KPI features of the customer data among four clusters 86
Figure 89: SMOTE on clustered dataset 88
Figure 90: VIF of features 89
Figure 91: Classification report - Multinomial Logistic Regression 89

Figure 92: Classification report - Decision Tree Classifier 90
Figure 93: Classification Report - Random Forest Classifier 90
Figure 94: RF maximum depth and accuracy 92
Figure 95: RF no. of estimators and accuracy 92
Figure 96: Accuracy score after RF model optimization 93
Figure 97: Feature Importance of RF Model 93
Figure 98: Customer Segmentation Prediction Web Application 94
Figure 99: User input parameters 95
Figure 100: Customer segment labels 96
Figure 101: User input customer record 96
Figure 102: Customer segment prediction 97
Figure 103: Prediction probability 97

LIST OF TABLES

Table 1: Information table for customer value matrix Source: (Marcus, C., 1998) 12
Table 2: Customer Segmentation techniques in previous research studies for Customer Segmentation 15
Table 3: Distance Matrices 20
Table 4: Descriptive Analysis of the data columns of customer dataset 29
Table 5: Reduced PC factors 60
Table 6: Clustering Model Evaluation Results 76
Table 7: Mean of each feature for each cluster 82
Table 8: Descriptive analysis of Cluster 0 85
Table 9: Descriptive analysis of Cluster 1 85
Table 10: Descriptive analysis of Cluster 2 85
Table 11: Descriptive analysis of Cluster 3 85
Table 12: Classification Report Evaluation Summary 91

ABBREVATIONS

BIRCH – Balanced Iterative Reducing and Clustering using Hierarchies

CLV – Customer Lifetime Value

CRISP-DM - Cross Industry Standard Process for Data

CRM – Customer Relationship Management

DBSCAN – Density-Based Spatial Clustering of Applications with Noise

GMM – Gaussian Mixture Models

LR – Logistic Regression

ML - Machine Learning

PCA – Principal Component Analysis

RF – Random Forest

RFM - Recency, Frequency, Monetary Value

SMOTE - Synthetic Minority Oversampling

WCSS – Within Cluster Sum of Square

BCSS – Between Cluster Sum of Square

CF Tree – Clustering Feature Tree

VIF – Variance Inflation Factor

CHAPTER 1: INTRODUCTION

Over the years, the commercial world has moved into a rapidly competitive era. All organizations have to do a lot of work as they have to fulfill the needs and expected services of their customers, gather new customers, and also develop their businesses. Organizations must have an interest to invest in the development of customer acquisition, maintenance, and development of strategies. Business intelligence has the main role to act in allowing companies to use technical knowledge to get better knowledge about the customer and programs for outreach. The stable value of a customer to a company plays a core ingredient in decision-making.

In the past few years, every retail industry pays their attention to Customer Relationship Management (CRM) to give better services to their customers when compared with their competitors. Building up a strong relationship with customers helps the enterprises in maximizing profit and customer retention and satisfaction. It is needful for large organizations to identify the potential customers in the huge market by mining the customer data in order to gain the profitable insights.

In the Banking sector, customers are diverse, and they require personalized services from banks where all banks change from the traditional system and implement new changes needed as per the customer preferences. Although the customers are varying from each other only a few attributes about one customer can match with another customer which helps the banks better serve the segment of customers by predicting their wants and needs well in advance. Banks need to focus on the potential of understanding customer data by segmentation using artificial intelligence and machine learning techniques. The segmentation of customer data benefits the banks with the personalization of customer experiences while enhancing and defining the products to make them quickly adapt to their customer needs, habits, and interests.

1.1 Statement of the Problem

In today's world due to the high competition, a loss of the main customer might have a significant effect on the cash inflows and investment of the organization especially in the banking sector which can lead to short-term cash flow problems, or it can even lead the organization into high debt or bankruptcy. Fighting to get individual customers, either to attract new customers or to maintain the current customer audience satisfaction, has become a survival game for each and every industry.

The process of identifying and meeting the needs and requirements of each and every customer in the business is very hard. One of the biggest challenges faced by customer-based organizations is customer cognition, identifying the difference between them, and rating or scoring them. The main reason for this is customers can vary according to their needs, wants, size, demographics, taste, and features, etc. Because of that, it is not a better practice to treat all the customers equally in the business market. Currently, most banks do customer segmentation based on demographics such as age, gender, education, location, etc. A basic segmentation will only allow for fewer predictions and will rely on basic assumptions. But this method needs to be stepped forward to gain an insight into the customer's profile on a more granular level.

It is required for any retail industry or for the banking sector to understand their customer market size and also it is important to have knowledge about customer behaviors for the marketing managers to re-evaluate their strategies with the customers. Furthermore, it is crucial for industries to evaluate the factors which affect the behavior of customers for them to operate successfully. The knowledge of the factors both identical and non-identical factors of separate customer segments help companies to correctly identify the niche characteristics of the customers and also to choose proper marketing tools. When organization employs the customer segmentation criteria, the most effective intelligence can be invented with the use of analytical methodology. Perfectly and accurately done customer segmentation helps to empower any retail businesses, banking sectors or any other businesses to interact with each and every customer in the best efficient approach.

1.2 Motivation

It is very important for industries, especially in the banking sector to focus on keeping good customer relations and also enhancing customer retention over the lifetime of the customer in order to generate higher profits and growth. In order to achieve this target, every organization should have a strong and stable tool to analyze their customers and provide everyone the care they need to keep them alive on their customer's active list or the current list without sending them to the past customers' list. Then the organizations can get a chance to obtain the maximum of their operations budgets by targeting the exact appropriate audiences.

The most successful organizations today are the ones that know their customers well that they can anticipate their needs. This opens up many opportunities and challenges for data science researchers to identify these in-depth insights and group or segment the customers to better serve them. This challenge has been motivated for the adoption of the scheme of customer segmentation or market segmentation, where the customer base is partitioned into smaller groups called segments as in members of individual segment exhibit similar market behaviors and characteristics.

1.3 Research Aims and Objectives

1.3.1 Aims

This research project aims to explore the approaches of using customer segmentation, as a business intelligence tool for the organizations and aims to explore the clustering techniques for the segmentation to help organizations to redeem an intelligible picture of the valuable customer base. This customer segmentation can be applied to any marketing department in banking sector or any retail industry to segment customers into clusters. Customer segmentation will be done by analyzing the customer data to identify the customer groups in order to develop customized relationships, maximize customer benefits, and for the target marketing strategy.

The main target of this project is identifying different customer types and segmenting the customer base into clusters of similar profiles so that the process of target marketing can be executed in an efficient manner. In the banking sector, a strong understanding of customer segmentation will be beneficial to achieve the following goals:

- Lower acquisition costs: here banks can implement more personalized services which can increase the probability of converting more prospects in to customers. And banks can target on specialized groups which yield on high profit margins.
- Increased sales: banks can offer the exact desirable services by customers when knowing their interests, habits and desires at the proper time.
- Decrease churn: when customer satisfaction increases with specialized and exclusive services the loyalty and brand retention will decrease the churn rate.
- Improved marketing campaigns: by using customer segmentation, the banks can decide the most accurate ways to attract new customers by promoting specific products to the ones who are most needful. Which leads to a better understanding on targets will increase sales.

1.3.2 Objectives

This will be achieved by attaining the following objectives:

- Identify factors that contribute most to understand customer's purchasing patterns and to gain an insight of the customer's profile.
- Using these factors, perform unsupervised machine learning algorithms (clustering algorithms) to obtain the segmentation.
- Identify the optimal number of customer groups/clusters through tuning.
- Identify the different customer groups that reflect similarities among customers in each group according to the spending patterns and purchase behaviors.
- According to all different customer groups, identify the most profitable customers.
- Build a model to predict the customer group for a selected customer.

1.4 Scope

Customer segmentation helps businesses to recognize and expose different customer segments that think differently and follow different purchasing strategies. Customer segmentation provides an efficient way of figuring out the customers who vary in terms of preferences, expectations, desires, and attributes. The main purpose of performing customer segmentation is to group people, who have similar interests so that it will help the business to converge in an effective marketing plan. The research will be carried out on the selected customer dataset from the banking sector. The selected customer dataset is from a bank in New York city. The dataset represents the active credit card holders' data of the bank which has been collected during a time period of six months. The dataset consists of around 9000 records of customer level data mentioning the Customer ID, Balance, Balance Frequency, Purchases, One off purchase, Installment purchases, Cash Advance, Purchase frequency, Purchase installment frequency, Cash advance frequency etc. Credit card details and purchasing behaviors will be focused on the research. All together around 18 features will be considered as mentioned above.

1.5 Structure of the Thesis

The thesis documents the research work with five main chapters. The second chapter is the literature survey which includes a background study of the research project referring to the previously published research materials, papers, online articles, books, magazines, etc. An in-depth literature review was performed to identify the previous methodologies followed in similar research studies. The third chapter documents the research methodology which explains the dataset, features, design, and modeling with the technical aspects. The fourth presents the evaluation of the results obtained through performing the research methodology. The final chapter concludes the research project by summarizing the research work along with future works and the limitations of the research project.

CHAPTER 2: LITERATURE REVIEW

The chapter 1 presented the overview of the research project comprising with the problem statement, motivation, research aims and objectives, scope, and the structure of the thesis. This chapter documents an in-depth background study on the research including a critical review of similar research which published previously. There are studies related with the customer segmentation as it is challenging to find perfect segmentation groups and to deduce relationships between customers because the customers status, needs and wants values of the customer will be changing from time to time. For better approach of the research, it is a must, to do a literature survey about recent research which have been carried out under these correlated fields of study which would be a great advantage when it comes to the implementation phase.

2.1 A Literature Review

Customer Relationship Management(CRM)

CRM is a major business approach to develop and secure steady, long-term customer associations. The modern marketing approach strengthens the utilization of CRM as part of the organization's business strategy for building up customer service satisfaction. CRM invariably plays a significant role as a market strategy by providing the organizations with ideal business intelligence for establishing, managing, and developing valuable long-term customer relationships. Several organizations and business institutions have conceived that the importance of CRM and the application of intelligence marketing strategies to achieve competitive advantage among other institutions (Rygielski et al, 2002).

CRM facilitates business enterprises with customer value analysis and also it facilitates target marketing strategies for valuable customers. It also supports business organizations to develop high-quality and long-term customer-company relationships which improve loyalty and also the profits. A more accurate evaluation of customer profitability and the targeting of high-value customers are prime factors that bestow the success of CRM (J. Lee et al, 2005).

CRM portrays a significant role in targeting the customer base. The essential segmentation strategies can be used after the targeted customer base is identified. The CRM strategy is a closed circular ruptures with four dimensions namely, customer identification, customer attraction, customer retention, and customer development. The customer identification is the main part of this structure which clearly supports to the act of grouping or segmenting customers according to their behavior and characteristics, thus the customer segmentation, emerges as a primary function of CRM (Swift, 2000).

Customer Segmentation

Through the years, the commercial world is becoming very competitive. It is very important for organizations to satisfy their customers' needs and wants to enhance their business. However, the task of identifying and satisfying needs and wants of every single customer is a very difficult task. Because of the customers may be varies in their own needs, wants, demography, geography, tastes, preferences, behaviors and so on. Therefore, it is not a good practice to serve all the customers equally in business (Puwanenthiren, 2012).

Customer segmentation means grouping the customers according to various characteristics and behaviors. This is a way for businesses to get a better understanding of their customers. When knowing the differences between the customer groups, it will be more efficient to make strategic decisions on product growth and marketing. The opportunities to segment depend on the customer dataset. There are different methodologies for customer segmentation, and they depend on four types of parameters: Geographic, Demographic, Behavioral, and Psychological (Bhade et al, 2018).

"The purpose of segmentation is the concentration of marketing energy and force on subdivision (or market segment) to gain a competitive advantage within the segment. It's analogous to the military principle of concentration of force to overwhelm energy." Customer segmentation includes geographic segmentation, demographic segmentation, media segmentation, price segmentation, psychographic or lifestyle segmentation, distribution segmentation and time segmentation (Thomas, 2015).

The process of customer segmentation aids in conducting analysis on the needs and wants and also market behavior of customers. And also, this process helps in effective decision making based upon the changing market conditions and the competitors (Bilgic et al, 2015).

In today's economy, to become a successful financial institution it is a must to prioritize their clients. To succeed in this goal the institution must devise a marketing plan with thrive to enhance the client values by targeting the lucrative customer relationships. The Figure 1 depicts the process of marketing strategy, which provides the path to get an insight into the consumers to serve them better.

The first phase of the process is market segmentation, where it divides bank’s market into groups of customers which most probably react similarly to a particular marketing campaign. This division employs banks to identify which group of the customer is more appealing which leads them to focus their efforts on this group (Kotler and Armstrong, 2010). Then the next phases are market differentiation and positioning, in which the decision making happens such as, deciding the most appropriate way to deliver the product to the market where the product will be highlighted and will be promoted the competitive advantage of the bank (Hiziroglu, 2013).

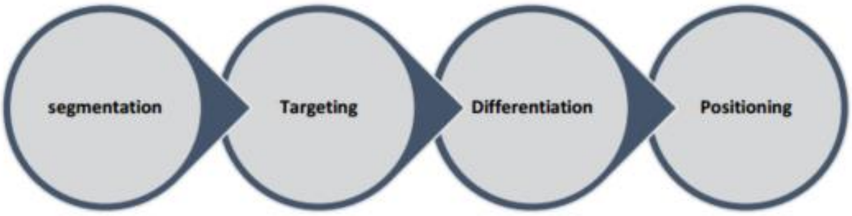

Figure 1: The process of marketing strategy for banks and other financial institutions. Source: (Kotler and Armstrong, 2010)

The Figure 2 depicts another process of bank customer segmentation with further detail. The process is divided into four steps namely, segmentation analysis, then the segmentation assessment, furthermore the segmentation implementation, and finally the segmentation control (Goller, Hogg and Kalafatis, 2002).

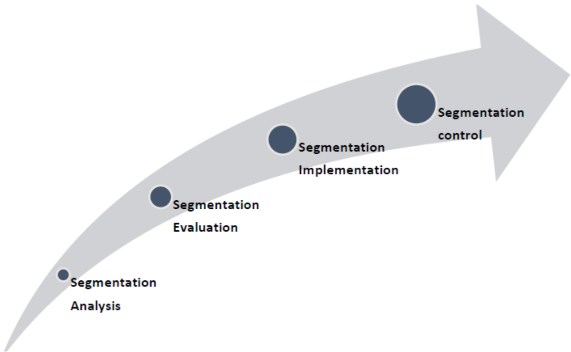

Figure 2: The process of segmentation for financial

Customer Segmentation Based on RFM method

The RFM (Recency, Frequency, Monetary) technique is one of the most commonly used method for customer segmentation in market analysis which can be utilized to identify the customers' behavior by evaluating three dimensions namely, the recency value, the frequency value, and the monetary value. This conventional method is commonly used to identify the behavior of their customers by analyzing the present customer behavior characteristics (Madani, S., 2009).

This method has been utilized in direct markets for about more than 30 years to identify the customer behaviors. Moreover, the RFM model was emphasized to distinguish valuable customers by measuring these three values. These three values are defined in the literature as:

- Recency (R): the recent purchase time of a customer.
- Frequency (F): the total number of purchases customer made during a specific time period.
- Monetary (M): the monetary value customer spent during a specific time period.

There are mass range of studies that have considered RFM method. These previous studies in this research area have highlighted the importance of RFM variables. In the RFM method, it analyzes and ranks a customer numerically for each of the mentioned three categories, ordinarily on a scale of 1 to 5 (where the higher number indicates a better result). According to the analysis, the recognized best customers receive the highest score in each category.

Customer Segmentation Based on CLV method

The CLV (Customer Lifetime Value) method has been utilized for decades in numerous marketing-based companies. The definition of Lifetime Value i.e., LTV is defined as the total number of revenues gained from the customers of a particular company over the company lifetime of transactions following the deduction of the total attraction cost, total selling cost, and total cost of servicing customers. Thus, the result of the calculation represents the time value of money (Hwang et al, 2004).

The CLV analysis method can be decomposed into three main components, which are the current value, the potential value, and customer loyalty. Previous studies that employ the CLV method for customer segmentation has been followed one of these three approaches. The previous studies either segment customers using purely the CLV values, or by using the mentioned components of the CLV method, or have been followed by considering both the CLV and other information which are socio-demographic information and transaction history information, etc. Generally, most of the banking domain-related studies, the last-mentioned approach has been widely followed (Kim, et al., 2006).

In (Sohrabi and Khanlari, 2007), researchers have estimated the customer lifetime value by measuring the RFM variables and further they have clustered an Iranian private bank customers and have proposed a customer retention strategy for treating customers.

Methodology of the Customer Value Matrix

The Customer Value Matrix method has been initiated from the RFM analysis method targeting the small-business retail environments. This method was introduced by Charles Edmundson. It has been noticed that although the RFM employs a straightforward conceptual framework RFM is a complex method and overdue for small retailer businesses. The reason for this has been identified as the results of segmentation based on RFM relent many segments and which has been caused difficulties for marketers to realize which groups can be more suitable for implement a particular strategy.

The initial step of customer value matrix method is to collect the necessary data for the creation of Customer Value Matrix. A customer identification number (Customer ID), the purchase date and the sum of the purchase amount are the data that needs to be extracted from business's database (Marcus, C., 1998). The following step is the customer segmentation process. In the former phase, the average measures for the purchases and the average value of Spending amount needs to be calculated. Finally, each customer in the business is allocated with one of the four results as shown in figure 3. Table 1 depicts the measures that needed to be calculated for the customer segmentation process.

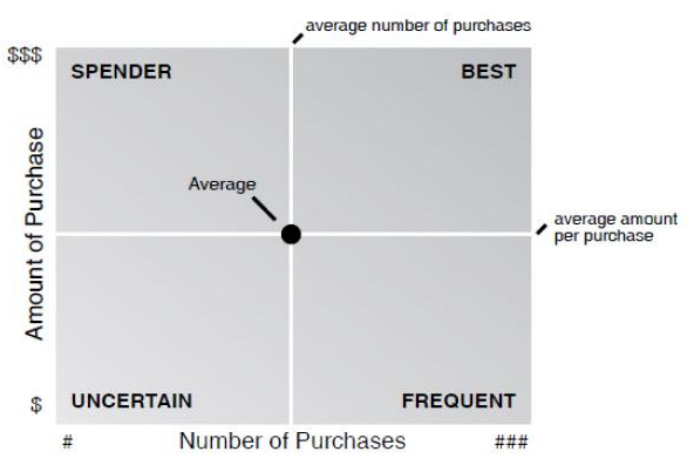

Figure 3: Customer Value Matrix, Source: (Marcus, C., 1998)

Table 1: Information table for customer value matrix Source: (Marcus, C., 1998)

Average number of purchase = Total Number of purchases/ Total number of customers
Total Number of purchases
Total number of customers
Average purchase amount = Total sales/ Total number of customers
Total sales
Total number of customers

Data Science algorithms as a Customer Segmentation Method

The contemporary methods of customer segmentation are underlined with the data science algorithms. Data mining is the finest solution for extricating the meaningful data and information from the raw data in databases typically which are available in numerous and diverse amount of data. It is hard to recognize expressive conclusions through the raw data marketing. The data mining methods and the output results of the process are being used to increase revenue and further to improve the communication including the CRM between organizations and their customer base.

The one effective modern method to perform customer segmentation is by utilizing the data science 'clustering' algorithms. The clustering method is one of the unsupervised learning methods in data science. This method can illuminate the customer segmentation problem by detecting and identifying unexpected or unknown features in the data. This method is capable for apply to a large data source and the performance execution will be fast. Although, there is a drawback of the clustering algorithm where the groups formed from the algorithm might be complex to interpret, and also not be clear about the way to implement the clustering algorithm, as for with which criteria (Blanchard et al. 2019).

When conducting a data-driven market segmentation, usually it is assumed that the market segments exist in the data itself and then it is revealed and described by segmentation (Dolnicar et al. 2018). In this research, the cluster analysis using the case data is also discovered the already existing, but hidden segments.

Clustering Method

Clustering Method is a technique beneficial for exploring the data. It is significantly beneficial where there is more data and there is no obvious natural groupings. In this case, clustering data mining algorithms could be utilized to find the natural groupings that may exist in the data. The cluster analysis identifies the clusters which embedded in the data. A cluster can be described as a collection of data objects that represents a similarity in a certain way to one another. A good clustering method will produce finest clusters ensuring the fact where the clusters employ a lower inter-cluster similarity and higher the intra-cluster similarity (Berkhin, 2012).

The clustering technique is one of the data mining techniques used for a variety of numerous applications, concerning the areas of machine learning, classification, and pattern recognition. There are various clustering algorithms, and those algorithms are varied from one another in accordance with the approach which they accompanied in order to group the objects with relevance to their characteristics (Inaba et al, 1994).

Although there are plenty of algorithms available in clustering technique, still this is a challenging task in data mining (Chang and Bai, 2010). As per the previous research, most of the clustering methods employs with the following general features: (Hammouda, 2001)

- Most of the clustering algorithms were driven by a particular problem domain.
- It is not including any explicit supervision effect and patterns that are organized with respect to a particular optimization criterion.
- All of these methods are adapted to the notion of similarity or distance.

In general sense, as shown in the figure 4 the clustering process includes four fundamental processes which are the Feature Selection, the Clustering Algorithm Design, Clustering Validation and finally the Interpretation of results. These four procedures in clustering algorithm are aligned with a pathway of feedback. Hence the clustering is not a one-way procedure rather this is a repetitive task (Xu et al, 2005).

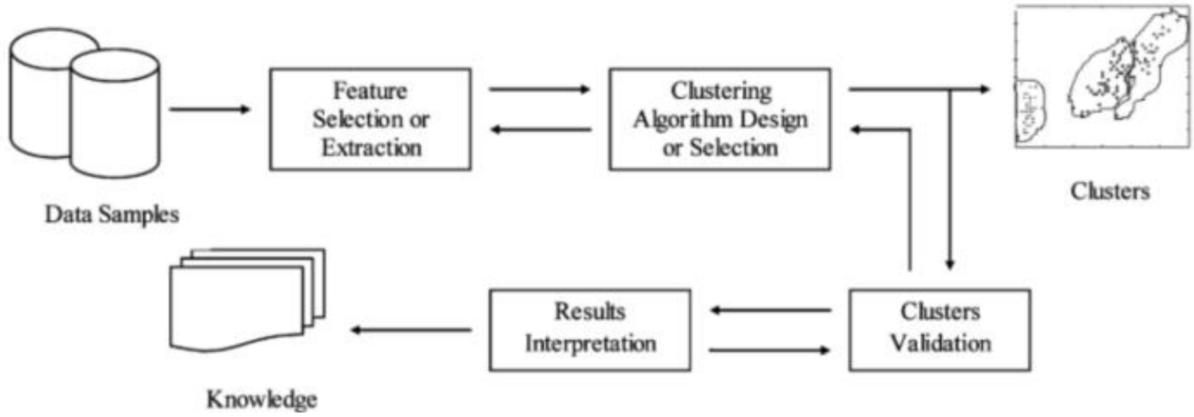

Figure 4: The clustering process

Clustering Algorithms for Customer Segmentation

Clustering techniques helps to disclose internally homogeneous and externally heterogeneous groups or clusters in the dataset. Customers may differ with regard to needs, wants, behavior and characteristics. The main target of clustering techniques is to recognize different types of customer types and to segment the customer base into number of clusters of similar profiles. Thus, the process of target marketing could be executed more effective and efficient manner. Both hierarchical clustering and non-hierarchical clustering algorithms are most commonly used in the process of customer segmentation (Chen et al, 2012).

Generally, the clustering methods falls into two different categories. Mainly, two types of algorithms namely, Hierarchical algorithms and Non-Hierarchical/Partitional algorithms (Yuanli T. and Liangshan S., 2010). Most of the research work have used the clustering techniques for the customer segmentation process. K-means clustering, and Hierarchical Clustering algorithms are widely used for clustering the dataset and for obtaining the extensive usage in customer segmentation.

Table 2: Customer Segmentation techniques in previous research studies for Customer Segmentation

Researcher	Year	Research About(Summary)	Factors
Kalyani Bhade, Vedanti Gulalkari, Nidhi Harwani, Sudhir N. Dhage	2018	A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization customer segmentation, it requires certain parameters that should be considered when segmenting the customers. In a broad sense, the clustering parameters can be classified as geographic, demographic, psychographic and behavioral.	Customer segmentation has been applied using k means, hierarchical, and density-based algorithms. And it is concluded that the K means clustering will provide the highest accuracy in order to segment the customers.
Tushar Kansal, Suraj Bahuguna , Vishal Singh, Tanupriya Choudhury,	2018	Customer Segmentation using K-means Clustering the customer segmentation based on features with datasets that contain 200 records with 2 features is proposed in in this research.	Kmeans clustering, agglomerative clustering, and mean-shift algorithm has been occupied to segment the customers. Also, it has considered two internal clustering measures namely, silhouette score and Calinski-Harabasz index.
Sabbir Hossain Shihab, Shyla Afroge, Sadia Zaman Mishu	2019	RFM Based Market Segmentation Approach Using Advanced K-means and Agglomerative Clustering:A Comparative Study implementation of three clustering algorithms namely k-means, advanced k-means, and agglomerative clustering have been proposed in this research	From the experimental results, it has been observed that agglomerative clustering is not a feasible solution because of its long execution time. And in this research, it has been concluded that for RFM based customer segmentation, advanced k-means clustering is more efficient and feasible.
Tripathi, S., A. Bhardwaj, and E. Poovammal	2018	Approaches to clustering in customer segmentation. Customers with similar means and behavior are grouped together into homogeneous clusters. K-means clustering, Hierarchical clustering: Agglomerative and	It has been concluded that the both K-Means and Hierarchical clustering techniques have some drawbacks. Hierarchical clustering is more suitable for business use, data

		divisive has been used to form the clusters representing the segmentation	visualization forms a major part of efficient data analysis. And K-Means tends to deliver better results in the aspect of performance.
Jan Panuš, Hana Jonášová, Kateřina Kantorová, Martina Doležalová, Kateřina Horáčková,	2016	Customer segmentation utilization for differentiated approach. The research represented another approach of combining the RFM model and ABC analysis, and data mining techniques for clustering the customers for segmentation	It has been concluded that the combination using of data mining techniques for synthesis of data gained from ABC analysis and RFM analysis is suitable for the utilization within CRM approach to customers. For the future work, it has proposed that decision trees or association rules will be considered.
Prabha Dhandayudam, Dr. Ilango Krishnamurthi	2012	An Improved Clustering Algorithm for Customer Segmentation various clustering algorithms results from varying cluster outputs and thus it has been compared the performance of those.	For a better clustering algorithm, within the cluster, customers should behave in a similar manner when compared to the customers in other clusters
Sunitha Cheriyan	2019	Intelligent Sales Prediction Using Machine Learning Techniques. A sales forecasting has been conceptually performed in this research paper by occupying intelligent machine learning models such as, Gradient Boosted Trees, Decision Trees and Generalized Linear Model.	It has been shown that the GBT showed most accuracy than other two techniques. It has been concluded that the business decisions are based on speed and accuracy of data processing techniques. Machine learning approaches highlighted in this research paper can be utilized for an effective mechanism in data tuning and decision making.

Ina Maryani, Dwiza Riana, Rachmawati Darma Astuti	2018	Customer Segmentation based on RFM model and Clustering Techniques with K-Means Algorithm An RFM model has been built based on the customer records, which resulted in 102 customers and the clusters were further clustered into 2 clusters.	Each cluster can be used to improve the market strategy by understanding the customers' behavior in each cluster
Cheng Li	2008	Research on Segmentation implementation process of air cargo Customer based on Data Mining. This research work summarizes the implementation of customer segmentation for the domain aviation cargo based on data mining techniques. Along with describing the hierarchical design strategy and methods of different levels, which improvise a reference value for the airlines to start CRM	This work concludes the segmentation in freight customers and connection with data mining theory can help air cargo business to determine the customers with a real value and analyze their features to maintain CRM them.
Vasilis Aggelis	2005	Customer Clustering using RFM analysis, This research analyzed that a calculation of RFM scoring for the active e-banking customers used for evaluation of the customer's behavior namely, strategic Decision making, future revenue forecasting and conservation of the most important customers.	This research work concludes that the knowledge of RFM scoring of active e-banking users can rank them according to the pyramid model. This result was highlighted using two clustering methods. And identified that the e-banking unit of a bank may efficiently identify the most important customers.

The above table 2 shows the brief descriptions of the literature of customer segmentation techniques used in previous research that was studied in this research study.

Cluster Result Validity

Cluster validity is a comprehensive subject area with limitless arguments where the conception of 'good' clustering is purely related to the domain of applications and to certain requirements (Halkidi and Vazirgiannis, 2001). Distinct clusters will be obtained with the usage of various parameter values from a clustering algorithm with the given. Thus, it is essential to decide the clustering that fits perfectly with the dataset and the business case.

Cluster validity techniques that are applied decide numerous aspects of cluster validity and are generally classified into three types. External Index; this is to measure the extent that which cluster labels are matched with the externally supplied class labels. For example, entropy is one of the external index measures. Internal Index; this is to measure the integrity of a clustering technique without pre-specified external labels or benchmarks. The Sum of Squared Error (SSE) is one of the internal validity measurements. Relative Index; this is to compare two different clustering approaches or two different clusters. Generally, for cluster validation, an external or internal evaluation is utilized in most scenarios (Kumar, 2005).

2.2 Research Gap

According to the above conducted literature survey, it is evident that the customer segmentation has been considered as a crucial concept in the financial and other organizations, which guarantees significant benefits that would eventually result in higher market revenues. Hence, it is not astonished the fact that organizations will be driven to utilize the strategy within their day-to-day operations, since it facilitates them to gain a better insight of their customers and to focus on the prime group that bring them more profit. The conventional approaches of customer segmentation are mainly utilized by through categorizing techniques based on the experiences, analysis of statistics or elementary partitioning (Xin-a Lai, 2009). These approaches cannot cater the requirements of farther complex analysis which businesses are facing in the present days.

The conventional models like RFM and CLV can only be capable of catering a limited number of selection parameters. As mentioned in above literature RFM model only based on the calculation of the recency, frequency, and monetary values. However, calculations based on these limited number of metrics may not always guarantee better results. Rather there must be numerous metrices to recognize the behavior of clients. Furthermore, organizations are rapidly changing and developing and with corporate to evolving of businesses the market campaigns and strategies also has to be evolved to cater the necessary requirements. With the numerously growth of customers' data and with the heavy usage of management information systems, the traditional or conventional customer segmentation approaches cannot cater to analyze large amount of customer data. It is almost a peculiar task to identify most significant information in the process of decision-making. Therefore, machine learning models can be utilized for the process of customer segmentation more effectively and efficiently to cater the requirements.

As reviewed in above section most of the previous research have considered conventional models like RFM and CLV for the customer segmentation. In addition to that there are studies that has been used clustering methods to segment the customers. There is no single way for organizations to segment their customer base. All these clustering techniques has various advantages and disadvantages relevant to the exact implementation and the input dataset. Therefore, customer segmentation is not a facile task to determine the most suitable algorithms and techniques to employ a given specific problem domain.
Most of the previous research on customer segmentation were focused mainly on K-Means and Hierarchical clustering approaches and there was no adequate research conducted as a comparative study of several clustering algorithms focusing on customer segmentation on the Banking Sector. All clustering techniques can be used in multiple ways, and most of the studies were concluded the research with the presentation of customer segmentation results and has been mentioned the prediction as a future work or limitation of the studies. This study focuses on filling these gaps by employing a hybrid approach with comparing several unsupervised machine learning models to segment the customers and focuses on come up with a predictive model by utilizing supervised machine models.

2.3 Presentation of Scientific Material

Computation of proximity and the Distance matrix

The selection of a suitable metric will affect the shape of the clusters, the data elements might be close to each other according to one distance and farther away when employs to another distance.

The table 3 depicts the matrices which are most commonly used for clustering algorithms.

Table 3: Distance Matrices

Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
Maximum distance	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1} (a-b)}$ where S is the Covariance matrix

Evaluation of clustering results will be a complex task as clustering itself, Cluster evaluation approaches are based internal evaluation criteria and external evaluation criteria. Internal evaluation approach is the clustering results are evaluated based on the data that used for clustering itself. And external evaluation is where the clustering results are evaluated based on the data that was not used for the clustering purpose. Those are referred to class labels and external benchmarks.

In general, most of the indices that used for internal clustering validation are based on compactness or cohesion and separation.

WSS/WCSS

Compactness or Cluster Cohesion is a measurement of how closely related are the objects in the cluster. A good compactness indicator of a cluster is a lower within-cluster variation.

Cluster Cohesion can be measured by using the within cluster sum of squares (SSE):

$$\boldsymbol{WSS} = \sum_{\boldsymbol{i}} \sum_{\boldsymbol{x} \in \boldsymbol{C}_i} (\boldsymbol{x} - \boldsymbol{m}_i)^2$$

BSS/BCSS

Cluster Separation is a measurement of how distinct or well separated a cluster is from the other clusters.

Separation can be measured by the between cluster sum of squares.

$$BSS = \sum |C_i|(m - m_i)^2$$

Where m_i is the mean of points in C_i and $|Ci|$ is the size of cluster i.

Silhouette Coefficient

Silhouette coefficient is used to determine the degree of separation between the clusters. A model with a proper defined clusters can be determined by a higher Silhouette Coefficient score. The Silhouette Coefficient is composed of two scores: mean intra-cluster distance and the mean nearest-cluster distance. The function computes the mean Silhouette Coefficient using these two scores. The mean intra cluster distance represents the distance between the sample and all other data points in the same cluster. The mean nearest-cluster distance represents the distance between the sample and all other data points in the next nearest cluster.

$$S = (b - a) \,/ max\,(a, b)$$

Where, a is the mean intra-cluster distance and b is the mean nearest-cluster distance.

The coefficient may take values within the interval [-1, 1]. Value closest to 0 represents the sample is too close to the neighboring cluster and a value closest to 1 represents the sample is much far from the neighboring cluster. And if the coefficient is -1, that means the sample is assigned to the wrong clusters.

Davies-Bouldin index

The Davies-Bouldin index calculated the measure of average similarity of each cluster relative to its most similar cluster. This similarity measure id is defined as a ratio of within-cluster distances to the between-cluster distances. Which means that the score is better when clusters are farther apart and less dispersed. The minimum score is defined as zero, when the score is lower it indicates of a better clustering.

Calinski and Harabasz score

The Calinski and Harabasz is also known as the Variance Ratio Criterion of the clusters. This score is defined as the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all the clusters among the cluster pool, when the score is higher it indicates of a better clustering performance (Yanchi Liu et al, 2010).

Dunn Index

The Dunn index is used to identify the dense and proper separated cluster. This can be computed as the ratio between the minimal inter-cluster distances to maximal intra-cluster distance. For each segment of cluster, the Dunn index can be computed by:

$$D = \frac{\min_{1 \leq i < j \leq n} d(i,j)}{\max_{1 \leq k \leq n} d'(k)},$$

Where, d(i,j) is the distance between cluster i and cluster j, and d '(k) is the intra-cluster distance of cluster k.

In the External Evaluation, the approach utilizes pre-known class labels and external benchmarks. They are consist of a set of pre-classified items and these are typically created by the experts.

RAND Index

The Rand index is used to calculate the similarity of the clusters obtained by the clustering algorithm to the predefined benchmark classifications. It can be calculated by:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

Where True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

F-Measure

The F-measure is used to balance the contribution of the false negatives by weighting the recall through a parameter that is $\beta > 0$. Precision and recall is defined by:

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$

F-measure can be computed by:

$$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

Where $\beta = 0$ and F0 = P.

V-Measure

The V-measure a type of numerical mean between the homogeneity and the completeness. A proper homogeneous clustering can be referred to a one where each cluster contains data-points belonging to the same class label. The homogeneity factor depicts the closeness of this type of a perfect clustering algorithm.

On the other hand, completeness factor depicts the closeness of the clustering algorithm where every data-point belongs to the same class are clustered into the same cluster. For N data samples, C class labels, K clusters and a_{ck} number of data-points belongs to class c and cluster k. The homogeneity h is given by:

$$h = 1 - \frac{H(C,K)}{H(C)}$$

Where.

$$H(C,K) = -\sum_{k=1}^{K}\sum_{c=1}^{C} \frac{a_{ck}}{N} log(\frac{a_{ck}}{\sum_{c=1}^{C} a_{ck}})$$

and

$$H(C) = -\sum_{c=1}^{C} \frac{\sum_{k=1}^{K} a_{ck}}{C} log(\frac{\sum_{k=1}^{K} a_{ck}}{C})$$

The completeness C is given by:

$$c = 1 - \frac{H(K,C)}{H(K)}$$

where,

$$H(K,C) = -\sum_{c=1}^{C}\sum_{k=1}^{K} \frac{a_{ck}}{N} log(\frac{a_{ck}}{\sum_{k=1}^{K} a_{ck}})$$

and

$$H(K) = -\sum_{k=1}^{K} \frac{\sum_{c=1}^{C} a_{ck}}{C} log(\frac{\sum_{c=1}^{C} a_{ck}}{C})$$

Therefore the weighted V-Measure is given by:

$$V_\beta = \frac{(1+\beta)hc}{\beta h + c}$$

The factor β can be adjusted according to the homogeneity or the completeness of the clustering algorithm.

In supervised learning, several evaluation metrices are used to evaluate the performance of the predictive model. The evaluation process of a classification model is primarily based on the calculations of number of correct predictions and number of incorrect predictions against the number of samples in the test records (Tan et al., 2014).

The primary evaluation metrices for classification model can be calculated as below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = true\ positive\ rate = \frac{TP}{TP + FN}$$

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall}$$

The notation TP indicates the number of true positives (TP) which means the number of instances that the model is correctly predicted the positive class. Likewise, the notation TN indicates true negatives (TN) i.e., the number of occurrences of the model where the model correctly predicted the negative class. Further, the notation FP indicates False positive (FP) which is the total number of instances that the model prediction is false, and it predicts the positive class. Moreover, the notation FN indicates the false negatives (FN) that the model prediction is false, and it predicts the negative class.

Accuracy can be described as the number of observations where the model correctly predicts the total number of observations in other words the proportion of the number of correct predictions and the total number of model predictions. Precision can be described as the total number of correct positive predictions predicted by the model proposed to the total number of positive predictions. Recall also known as the true positive rate can be described as the correct proportion of actual positive observations predicted by the model and the F1 score describes an aggregated measure of precision and recall calculated employing the harmonic mean (Lindholm et al., 2021).

CHAPTER 3: METHODOLOGY

This chapter explains all aspects of the concepts along with the research design and research methodology furthermore the cognization process of the whole research study. The research methodology comprises with the solutions to the developed research aims and objectives mentioned in the introduction chapter.

The Cross-industry process for data mining (CRISP-DM) process is depicted in figure 5, adopted as the main framework for the research methodology. The research aims to follow as closely as possible the below-depicted methodology. This procedure includes the applicable steps for the research work where the process is a voyage through a series of research phases.

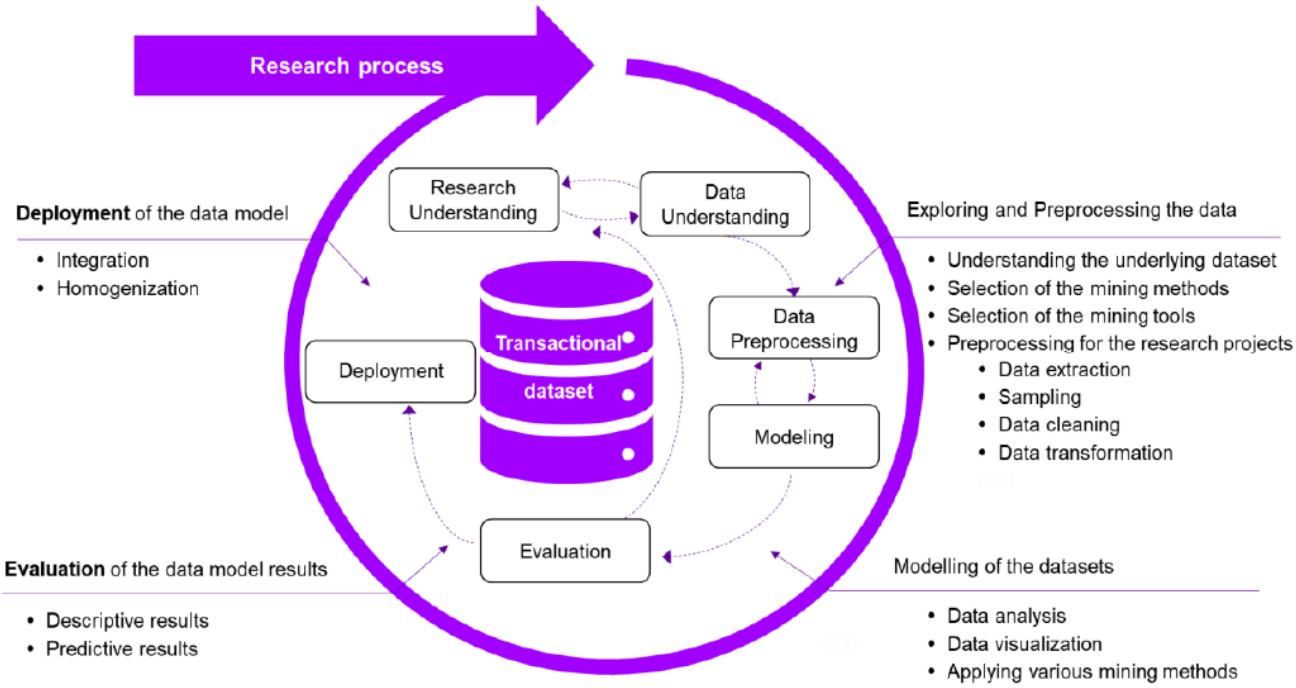

Figure 5: The CRISP-DM Process, Source: (Chapman et al. ,2000)

The main phases of research described in this dissertation can be concluded under six main phases:

1. Research understanding
2. Data understanding
3. Data preprocessing
4. Modeling
5. Evaluation
6. Deployment

3.1 Research understanding phase

The detail of this phase is described in-depth in the chapter 1, the customer segmentation aims to segment the customer base into different groups, this can be contemplated as a prominent asset for organizations most importantly in financial sector as it can be applied as an intelligent business strategy to extend the customer profitability among the whole pool of customers. It is vital for the organization to formulate an effective strategy for business expansion. This will provide the organization far more clear concepts about which clients have the highest retention rate. Especially, in the banking sector it is vital to gain more insight about the customers' behavior and to know the most preferred or loyal customers to the bank. This helps the bank to improve the customer retention rate to focus marketing strategies on a particular customer segment

3.2 Data understanding phase

According to the methodology, the targeted dataset is selected to prepare the data for modelling. The dataset summarizes the usage behavior of about 9000 active credit cardholders of a New York City bank during the time period of 6 months. The data represents on the customer level with 18 behavioral variables. This dataset is used to extract segments of customers depending on their behavior patterns provided in the dataset, to focus marketing strategy of the bank on a particular segment.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8950 entries, 0 to 8949
Data columns (total 18 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   CUST_ID                           8950 non-null   object 
 1   BALANCE                           8950 non-null   float64
 2   BALANCE_FREQUENCY                 8950 non-null   float64
 3   PURCHASES                         8950 non-null   float64
 4   ONEOFF_PURCHASES                  8950 non-null   float64
 5   INSTALLMENTS_PURCHASES            8950 non-null   float64
 6   CASH_ADVANCE                      8950 non-null   float64
 7   PURCHASES_FREQUENCY               8950 non-null   float64
 8   ONEOFF_PURCHASES_FREQUENCY        8950 non-null   float64
 9   PURCHASES_INSTALLMENTS_FREQUENCY  8950 non-null   float64
 10  CASH_ADVANCE_FREQUENCY            8950 non-null   float64
 11  CASH_ADVANCE_TRX                  8950 non-null   int64  
 12  PURCHASES_TRX                     8950 non-null   int64  
 13  CREDIT_LIMIT                      8949 non-null   float64
 14  PAYMENTS                          8950 non-null   float64
 15  MINIMUM_PAYMENTS                  8637 non-null   float64
 16  PRC_FULL_PAYMENT                  8950 non-null   float64
 17  TENURE                            8950 non-null   int64  
dtypes: float64(14), int64(3), object(1)
memory usage: 1.2+ MB
```

Figure 6: Information of the data columns of customer dataset

In this research project, several kinds of machine learning models were applied in the targeted customer dataset. Statistical analysis and data preprocessing methods were employed in this study. Mainly, Google Colaboratory and Jupyter Notebook are the web environments used for data processing and analysis.

The Python Pandas Library is utilized for loading the data. Afterwards, the info function used to depict the number of records and the data types as shown in figure 6. Then the Numpy library of Python is used for basic quantitative analysis of the data. Central tendency, range, standard deviation, mean, max and min values are calculated using descriptive statistics and matplotlib, plotly and Seaborn were utilized through the study. Further, Python machine learning libraries were employed in the study for cluster analysis, model training, and model evaluation and comparison.

The table 4 depicts the initial descriptive analysis of the raw data.

Table 4: Descriptive Analysis of the data columns of customer dataset

	count	mean	std	min	25%	50%	75%	max
BALANCE	8950.0	1564.474828	2081.531879	0.000000	128.281915	873.385231	2054.140036	19043.13856
BALANCE_FREQUENCY	8950.0	0.877271	0.236904	0.000000	0.888889	1.000000	1.000000	1.00000
PURCHASES	8950.0	1003.204834	2136.634782	0.000000	39.635000	361.280000	1110.130000	49039.57000
ONEOFF_PURCHASES	8950.0	592.437371	1659.887917	0.000000	0.000000	38.000000	577.405000	40761.25000
INSTALLMENTS_PURCHASES	8950.0	411.067645	904.338115	0.000000	0.000000	89.000000	468.637500	22500.00000
CASH_ADVANCE	8950.0	978.871112	2097.163877	0.000000	0.000000	0.000000	1113.821139	47137.21176
PURCHASES_FREQUENCY	8950.0	0.490351	0.401371	0.000000	0.083333	0.500000	0.916667	1.00000
ONEOFF_PURCHASES_FREQUENCY	8950.0	0.202458	0.298336	0.000000	0.000000	0.083333	0.300000	1.00000
PURCHASES_INSTALLMENTS_FREQUENCY	8950.0	0.364437	0.397448	0.000000	0.000000	0.166667	0.750000	1.00000
CASH_ADVANCE_FREQUENCY	8950.0	0.135144	0.200121	0.000000	0.000000	0.000000	0.222222	1.50000
CASH_ADVANCE_TRX	8950.0	3.248827	6.824647	0.000000	0.000000	0.000000	4.000000	123.00000
PURCHASES_TRX	8950.0	14.709832	24.857649	0.000000	1.000000	7.000000	17.000000	358.00000
CREDIT_LIMIT	8949.0	4494.449450	3638.815725	50.000000	1600.000000	3000.000000	6500.000000	30000.00000
PAYMENTS	8950.0	1733.143852	2895.063757	0.000000	383.276166	856.901546	1901.134317	50721.48336
MINIMUM_PAYMENTS	8637.0	864.206542	2372.446607	0.019163	169.123707	312.343947	825.485459	76406.20752
PRC_FULL_PAYMENT	8950.0	0.153715	0.292499	0.000000	0.000000	0.000000	0.142857	1.00000
TENURE	8950.0	11.517318	1.338331	6.000000	12.000000	12.000000	12.000000	12.00000

Then the Exploratory Data Analysis (EDA) is performed on the loaded dataset to gain a better understanding of the dataset. The python libraries namely, Matplotlib and Seaborn are used for data visualization. A histogram is commonly used to visualize the distribution of numerical data. When exploring the dataset, it is vital to get understanding of the distribution of certain numerical variables of the data. Box plots are used to detect the outliers of the dataset. Figure 7 to figure 23 depicts the distribution of features of the dataset and the histograms and boxplots utilized to visualize the distribution.

1. Cust_Id

This attribute indicates the customer id which is the identification of each credit card holder. This is a categorical variable, and this is unique for the customer.

2. Balance

This attribute indicates the amount of Balance left in customer's account to make the purchases.

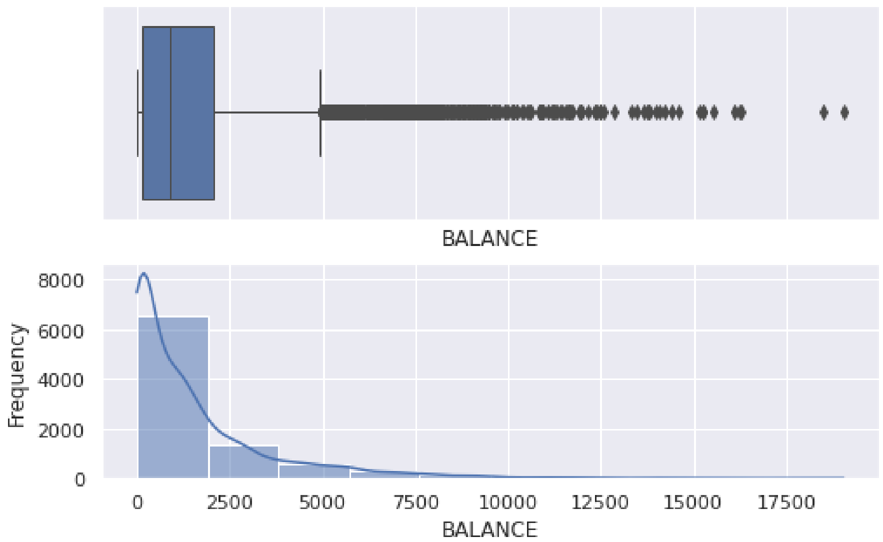

Figure 7: Balance Distribution

3. Balance_Frequency

This indicates how frequently the Balance is updated by the customer, records between 0 and 1 (1 = balance frequently updating, 0 = not frequently updating the balance).

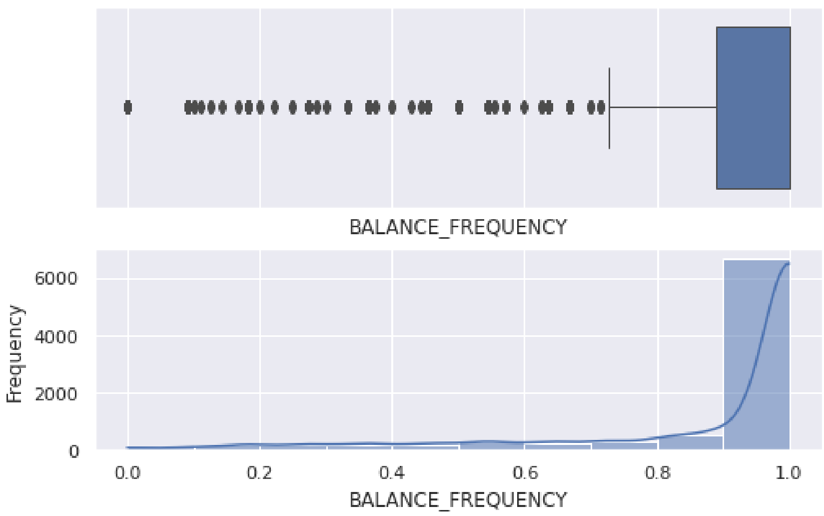

Figure 8: Balance Frequency Distribution

4. Purchases

This attribute represents the amount of purchases made from the customer's account.

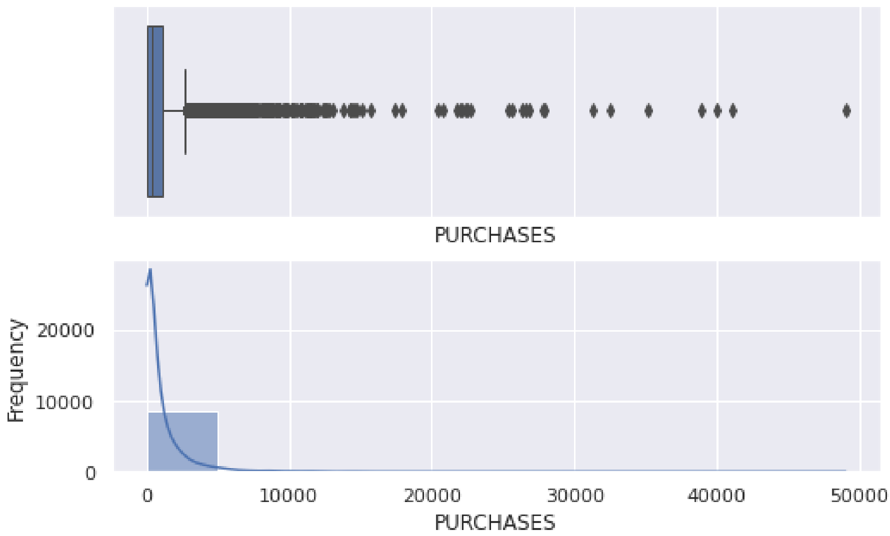

Figure 9: Purchases Distribution

5. One-off_Purchases

This attribute indicates the maximum purchase amount done by the customer in one-go.

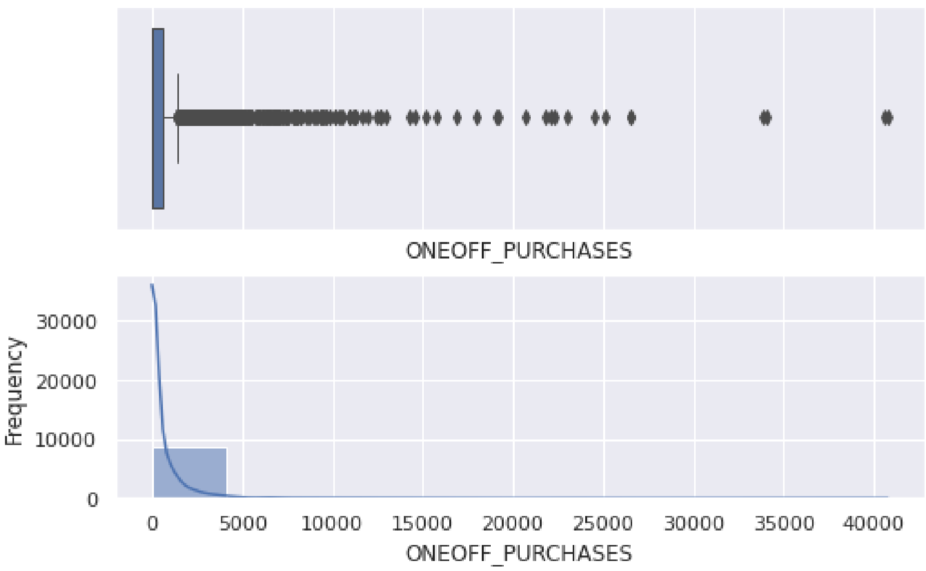

Figure 10: One-off Purchases Distribution

6. Installments_Purchases

This indicates the amount of purchase done by the customer in installments.

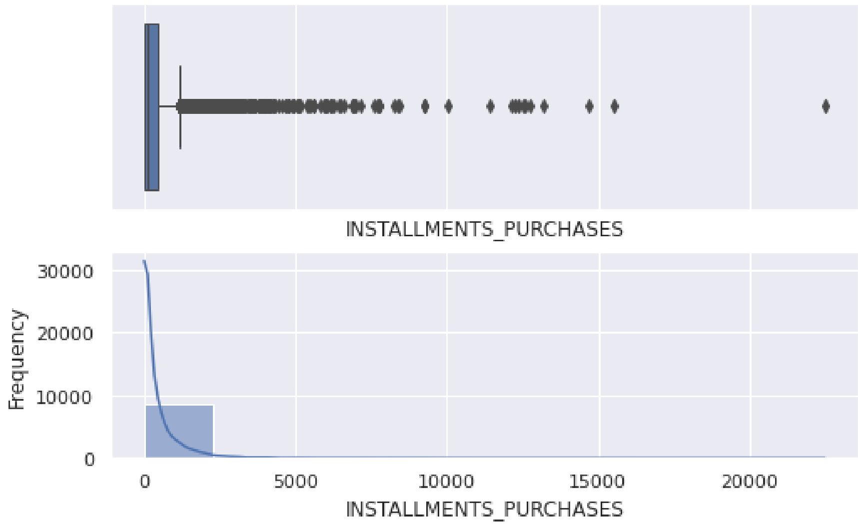

Figure 11: Installments Purchases Distribution

7. Cash_Advance

This indicates the cash in advance given by the customer.

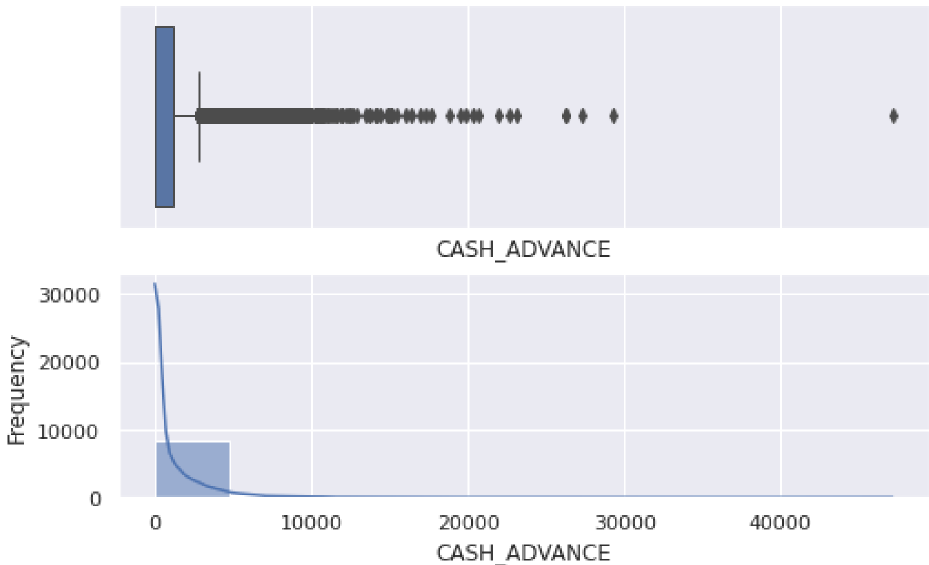

Figure 12: Cash Advance Distribution

8. Purchases_Frequency

This indicates how frequently the Purchases are being made by the customer, the recorded score is between 0 and 1 (1 = frequently purchasing, 0 = not frequently purchasing).

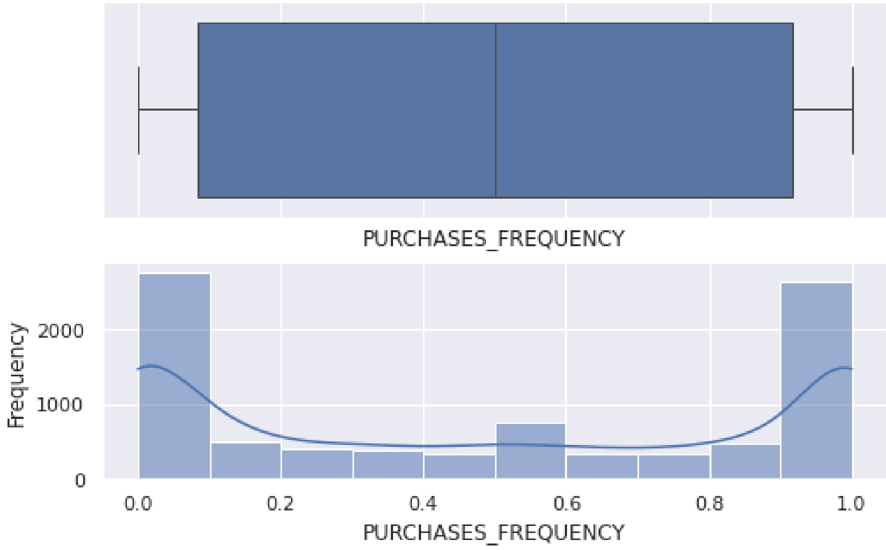

Figure 13: Purchases Frequency Distribution

9. Oneoff_Purchases_Frequency

This indicates how frequently the Purchases are being made in one-go (1 = frequently purchasing in one-go, 0 = not frequently purchasing in one-go).

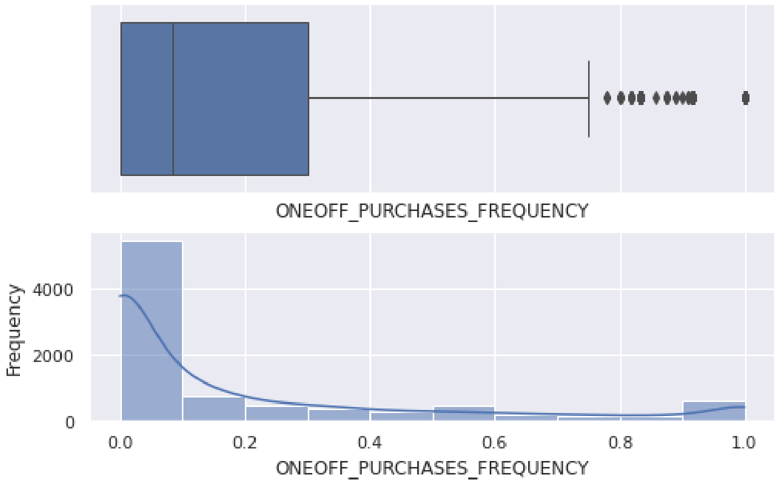

Figure 14: One-off Purchases Distribution

10. Purchases_Installments_Frequency

This indicates how frequently the purchases in installments are being done by the customer (1 = frequently purchased in installments, 0 = not frequently purchased in installments).

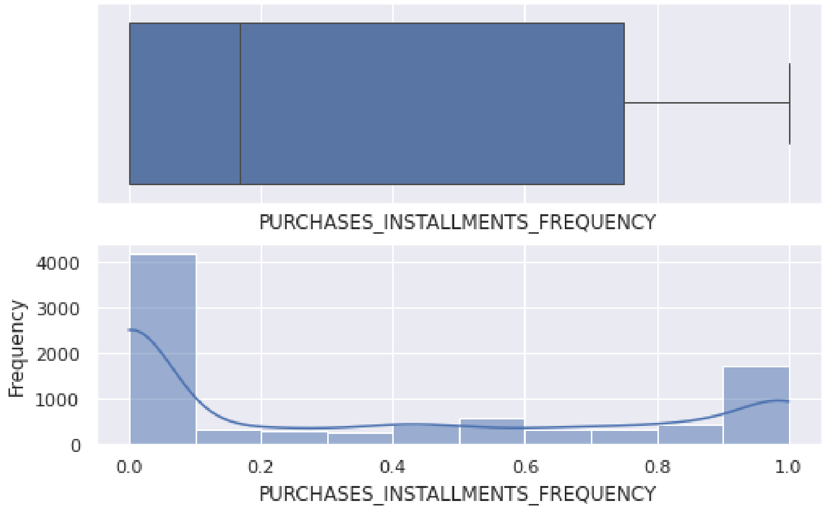

Figure 15: Purchases Installments Distribution

11. Cash_Advance_Frequency

This indicates how frequently the cash in advance being paid by the customer.

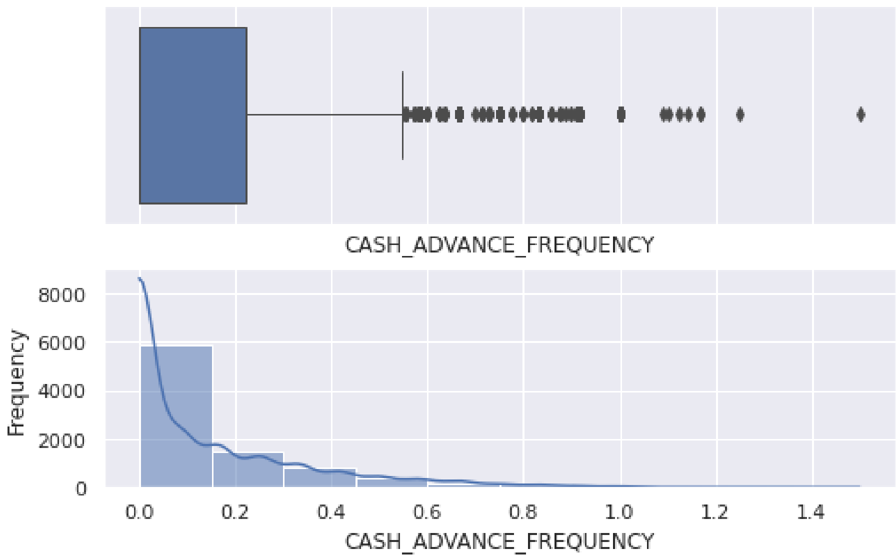

Figure 16: Cash Advance Frequency Distribution

12. Cash_Advance_Trx

This indicates the number of Transactions made with "Cash in Advanced" by the customer.

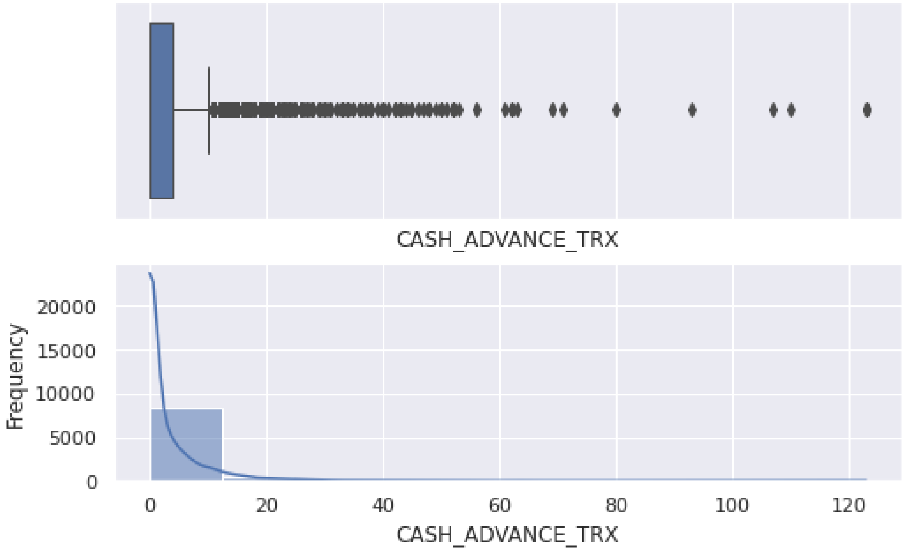

Figure 17: Cash Advance Trx Distribution

13. Purchases_Trx

This indicates the number of purchase transactions made with the credit card by the customer.

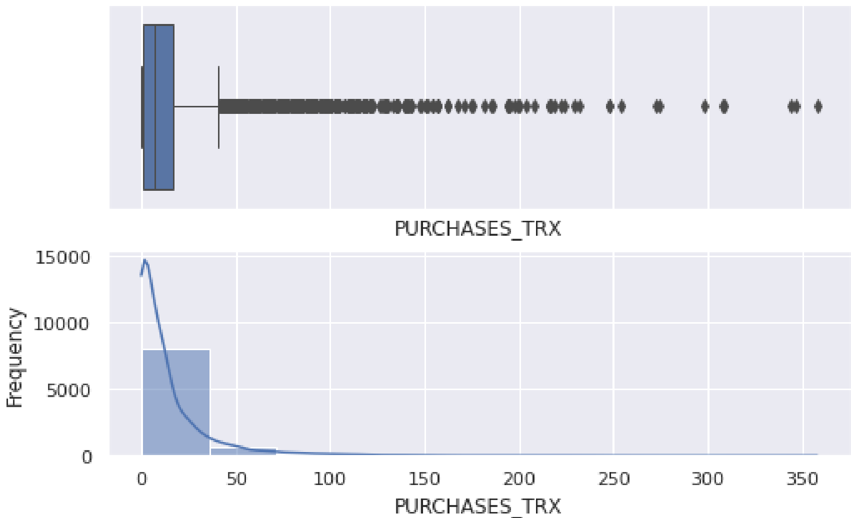

Figure 18: Purchases Trx Distribution

14. Credi_Limit

This indicates the limit of the Credit Card for the customer.

Figure 19: Credit Limit Distribution

15. Payments

This indicates the amount of Payment done by the customer.

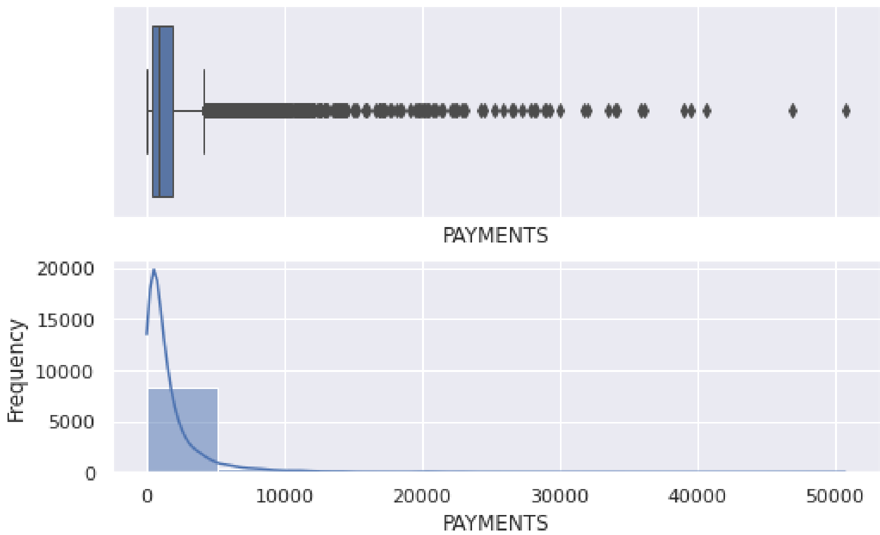

Figure 20: Payments Distribution

16. Minimum_Payments

This indicates the minimum amount of payments made by the customer.

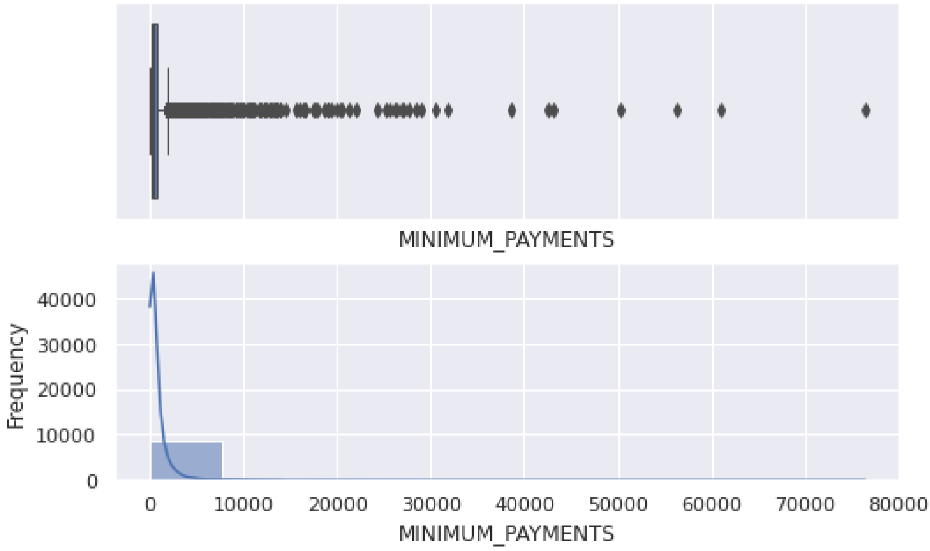

Figure 21: Minimum Payments Distribution

17. Prc_Ful_Payment

This attribute indicates the percent of full payment paid by the customer.

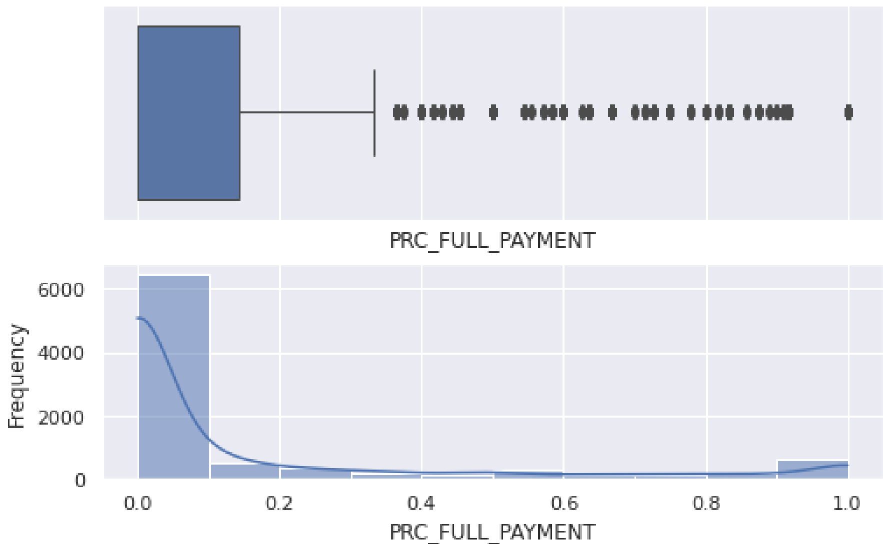

Figure 22: Prc Full Payment Distribution

18. Tenure

This indicates the tenure of the credit card service for the customer which is the pre-agreed time period for the customer to repay the principal and interest in full to the bank.

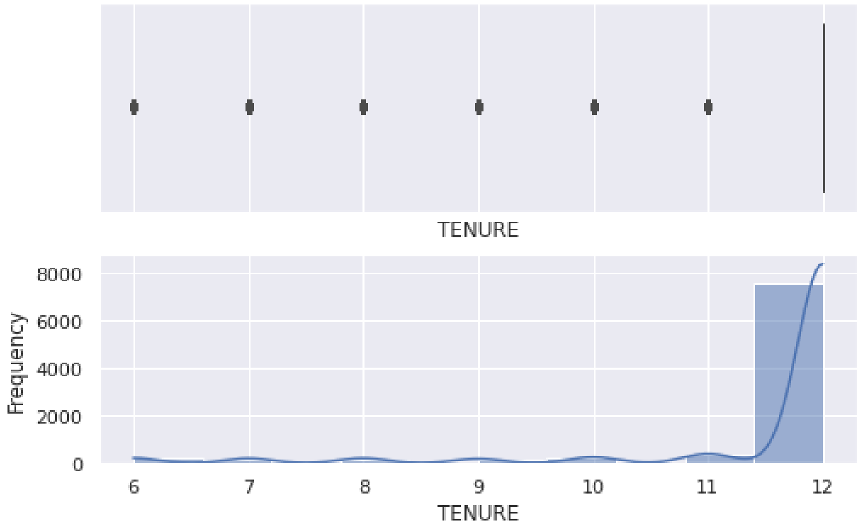

Figure 23: Tenure Distribution

3.3 Data Preprocessing

Before employing the targeted data to train the models, the data pre-processing is a required step since the real-world data is not always clean and well formatted. Generally, the raw datasets include null data points and also data may be collected from various measurements. These issues can be accommodated by applying data pre-processing techniques. This is a very crucial phase as it comprises with the numerous activities of converting the raw data into the final processed dataset. The resultant dataset from the preprocessing phase then can be fed into the machine learning models to obtain the separate clusters of the customers. In the Data preprocessing phase, Data Reduction, Data cleaning, imputing missing values, removing outliers, Creation of new variables, Data normalization and the data transformation are the activities which will be performed.

3.3.1 Handling Missing Values

When performing the exploratory data analysis using the dataset, one of the earliest things to detect is the existence of the missing values in the dataset. These missing values will be caused by reducing the quality of the dataset, furthermore, reducing the accuracy of the models that are trained on that data. Hence, the missing values handling is a very crucial part of data preprocessing. Missing data is an ordinary issue in datasets and yields to affect in a negative manner on the conclusions drawn from the data. However, it is not a good practice to remove the data records that contain missing values directly since the size of the dataset will then be lesser which means there will be less data for the model. Seaborn graphs were plotted for the customer dataset to identify the missing values and corresponding features (Kelleher et al., 2015).

There are several data imputation methods to overcome the missing value problem.

- Forward Fill (ffill)

  In this strategy, the value preceding the occurrence of the missing value is selected to fill the missing value. This technique is commonly used with time series data.

- Back Fill (bfill)

  Intuitively, this strategy chooses the valid value of the succeeding data record to fill up the missing value.

- Filling with the Mean

  Generally, this is the most commonly used method and here, the mean value of the attribute will be used to fill up the missing value. This can be used with the numeric columns.

According to figure 24, there are multiple missing values in Minimum_Payments attribute on the dataset and according to figure 25 there are several missing values in Credit_Limit and Minimum_Payments attributes.

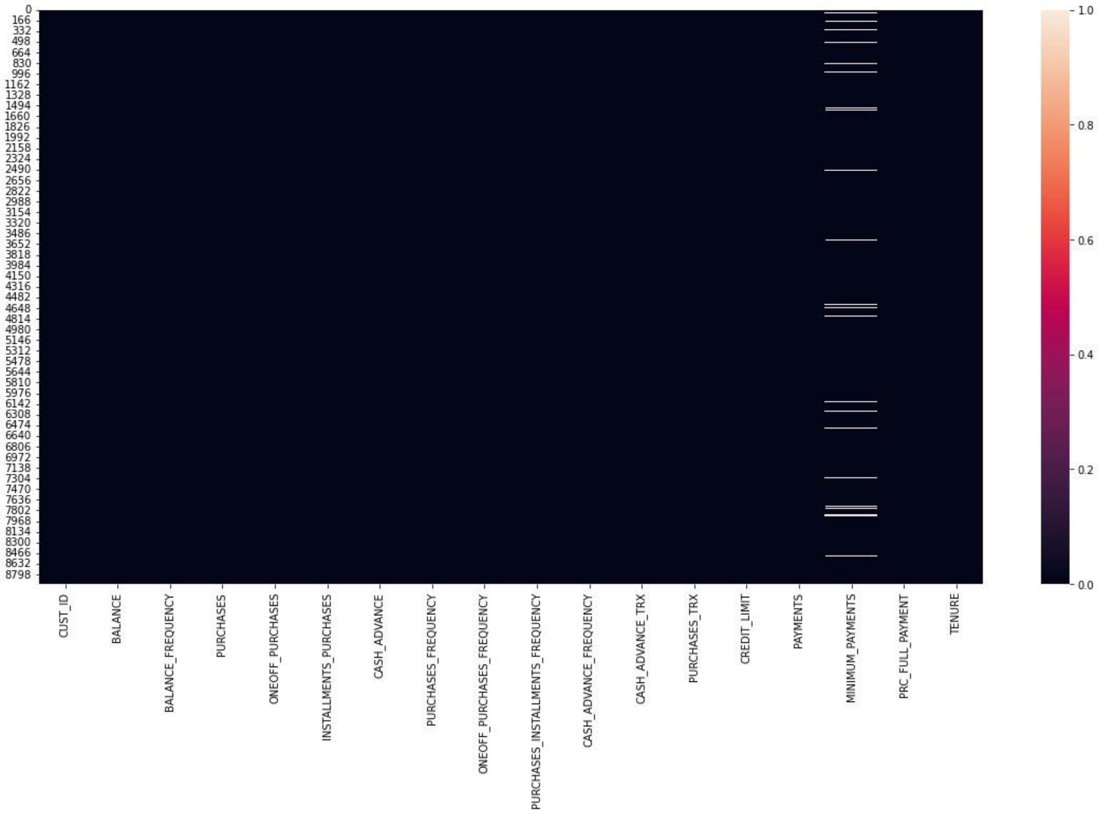

Figure 24: Missing Data visualization – Seaborn heatmap

.

```
CUST_ID                               0
BALANCE                               0
BALANCE_FREQUENCY                     0
PURCHASES                             0
ONEOFF_PURCHASES                      0
INSTALLMENTS_PURCHASES                0
CASH_ADVANCE                          0
PURCHASES_FREQUENCY                   0
ONEOFF_PURCHASES_FREQUENCY            0
PURCHASES_INSTALLMENTS_FREQUENCY      0
CASH_ADVANCE_FREQUENCY                0
CASH_ADVANCE_TRX                      0
PURCHASES_TRX                         0
CREDIT_LIMIT                          1
PAYMENTS                              0
MINIMUM_PAYMENTS                    313
PRC_FULL_PAYMENT                      0
TENURE                                0
dtype: int64
```

Figure 25: Missing value count

The missing values in Credit_Limit and Minimum_Payments attributes in the customer dataset are imputed using filling with the mean approach since both the columns are numeric columns and this approach is the most commonly used approach to handling the missing values.

3.3.2 Outlier Detection

The dataset contains outliers which are significantly differ and far from the other observations. Which are the data points that falls outside of the overall distribution of the data. There are multiple ways to detect the outliers in the data. By using mathematical formulas, statistical approaches or visualization tools can detect the outliers in the data. Scatter plots, Box plots and Histograms are some popular outlier detection methods, and these can be used to identify the outliers of the data. However, outlier treatment will depend on the algorithms using for building the models.

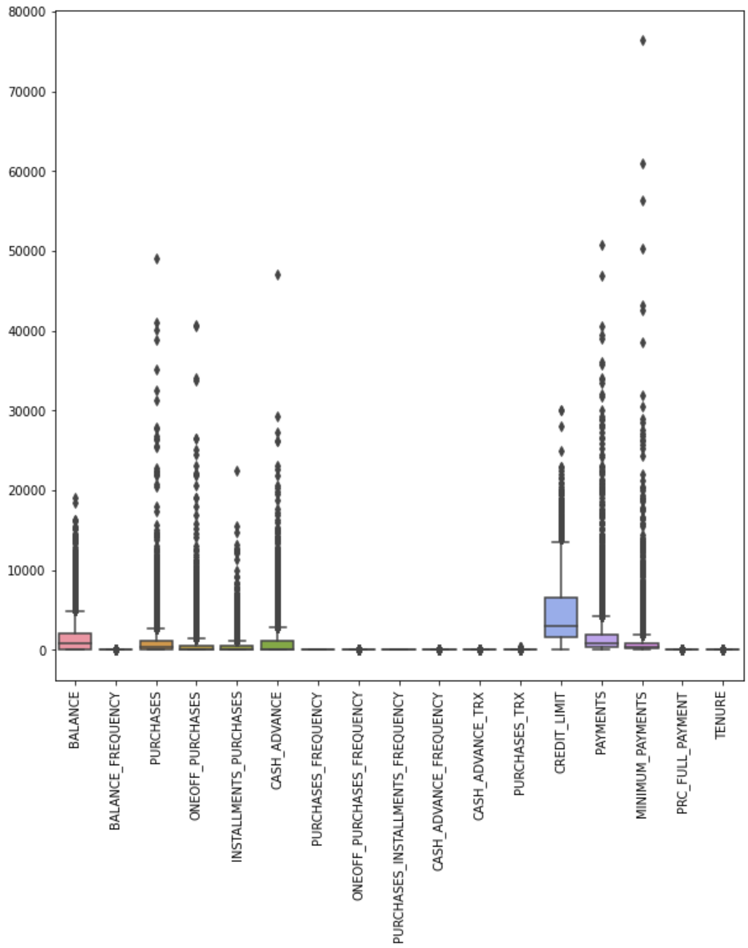

Figure 26: Outlier Detection

As shown in figure 26, the customer dataset is detected with the outliers and the outliers handled through the log transformation without deleting the records with outliers. Further standardization and normalization applied for further processing.

3.3.3 Normalization and Standardization

Scaling is one of the most crucial steps in data preprocessing phase in machine learning. Most commonly used feature scaling techniques are normalization and standardization. Normalization is used to transform the dataset features that are on different ranges of values to a common scale. Normally the scale will be ranged from 0 to 1 or on some cases -1 to 1. The scaling is very important as if the ranges are relatively far apart it will badly affect to the learning process. There are various normalization methods available namely, the standard scaler, the min-max scaler etc. In the standard scaler scaling occurs independently on every feature by calculating the relevant statistics on the dataset. In min-max scaling approach, the estimator scales each feature individually using the minimum and the maximum value in the dataset. However, this will depend on the algorithms using for training the models. The normalization and standardization applied to customer data and figure 27 depicts the box plots of features in the customer data after this process.

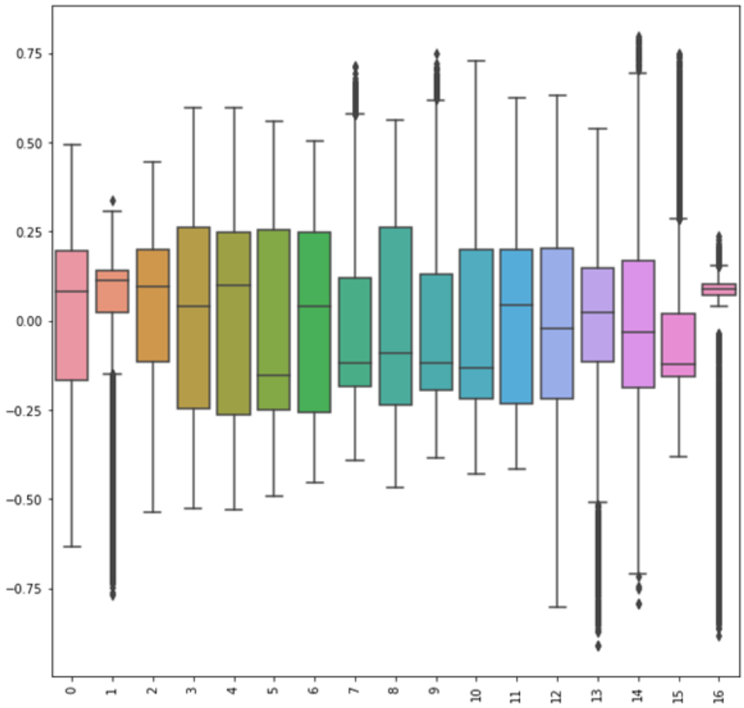

Figure 27: Normalization and Standardization

3.3.4 Feature Selection

Feature selection is another significant factor to be considered in the data pre-processing phase which significantly counterfeits the performance of the machine learning model. There are widely used feature selection techniques in machine learning which are Uni-variate Selection, Bivariate Analysis, Furthermore the Feature Importance Analysis and Correlation Matrix with Heat map (Shaikh, 2018).

3.3.5 Dimensionality Reduction

Generally, most of the clustering algorithms are not competent for handling high-dimensional data and these algorithms are more efficient and accurate when the number of features is relatively small which means approximately below 10 number of attributes (Han, et al., 2011).

One of the widely used approaches for dimensionality reduction along with clustering is to use Principal Component Analysis (PCA) to extricate the important components from the original dataset, which are then used to perform the clustering. The PCA is an unsupervised learning technique that deems for a significant ratio of the variation in the dataset along with projecting the data into a low-dimensional feature space (James, et al., 2013). Thus, achieving the principal components of the data which are a series of projections of the data that are mutually uncorrelated from each other and ordered in variance. The PCs of a dataset in $\mathbb{R}p$ provide a sequence of best linear approximations to that particular dataset, of all ranks $q \leq p$ (Hastie, et al., 2009).

3.3.5 Encoding

The datasets consist of different data types such as numerical, categorical, etc. Nevertheless, when the dataset is used in machine learning techniques or deep learning techniques, the categorical data has to be encoded to an applicable numeric form before the customer data are utilized for modeling.

3.4 Modeling

In this research six different unsupervised machine learning models will be trained on the same customer dataset. Six different clustering algorithms will be utilized in this study to perform the customer segmentation for the targeted customer dataset. For each model, the cluster analysis, distributions of the features among the clusters will be discussed and visualized.

The customer segmentation results of each model will be visualized. Further, Evaluation of results and comparison between the performances of each machine learning model will be presented. Finally, after evaluation of the cluster results obtained through training the clustering models the most accurate model will be selected to build the prototype of the cluster prediction system. This will be achieved using the supervised machine learning algorithms. Therefore, this research methodology employs a hybrid approach.

3.4.1 Customer Segmentation Model Building

The modeling of the customer segmentation utilized unsupervised machine learning algorithms and here six clustering algorithms were selected for the model training to obtain the customer groups or the clusters. For this k-means clustering, agglomerative clustering, spectral clustering, gaussian mixture model-based clustering, DBSCAN clustering, and BIRCH clustering were utilized for the cluster analysis and comparison. The definitions of selected clustering algorithms and their factors are briefly described below.

K-Means Clustering

K-Means clustering algorithm is one of the most popular unsupervised machine learning algorithms. Conventionally, unsupervised algorithms are used with unlabeled datasets to obtain the inferences from the data. K-means algorithm is a Partitional Clustering approach which divides the data objects into non overlapping groups.

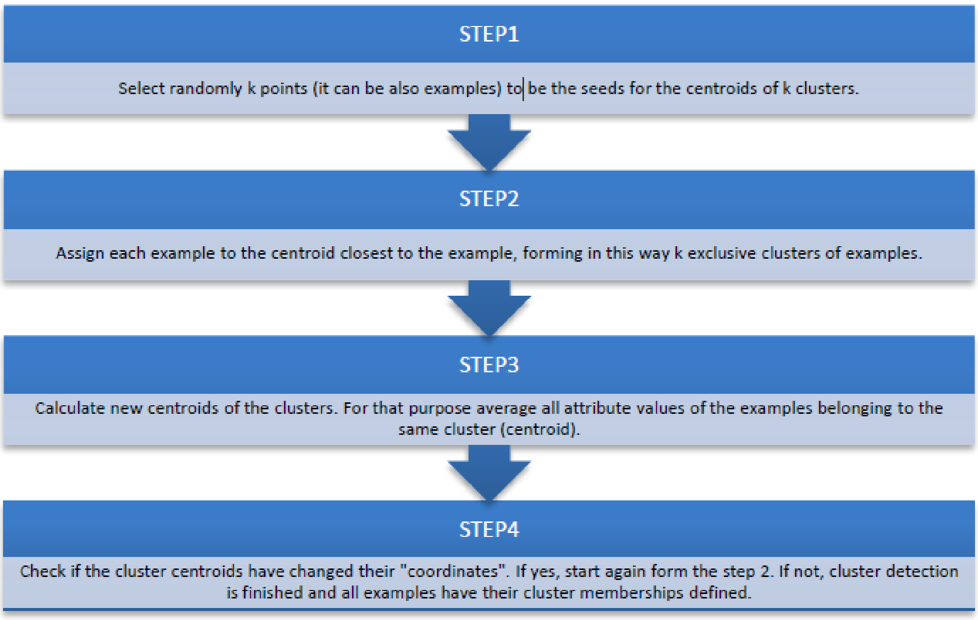

Figure 28: Classic K-means algorithm

The main objective of this algorithm is to obtain K groups of similar data together and to discover the patterns of data. The figure 28 depicts the main flow of the kmeans clustering. The K-means algorithm works iteratively to assign each and every data point in the dataset to one of K groups based on the provided features. Data points are clustered based on similarity of features. The inputs for the K-means algorithm are the number of clusters K and the provided data set. The output of the algorithm is a group of k number of clusters.

In the first iteration, the initial cluster centers are selected arbitrarily from the dataset, and the algorithm then iterates between two steps until there are no changes in the cluster centers. This step proceeds iteratively through reassigning each object to the cluster in accordance with the similarity of the object to the cluster based on the mean value in the objects in that cluster. And then the cluster means will be recalculated accordingly for the objects in each cluster (Arora et al, 2016).

Hierarchical Clustering

Hierarchical clustering determines the cluster assignments by building a hierarchy of clusters. These algorithms produce a tree-like diagram including hierarchy of points called a dendrogram. By cutting the dendrogram at a specified depth, clusters can be formed. This will result in k number of groups of smaller dendrograms. There are two strategies for the algorithm; this can be implemented by either using a bottom-up or using a top-down approach:

- Agglomerative clustering represents the bottom-up approach. In this algorithm, it is not required to pre-specify the number of clusters. This will treat each data point as a singleton cluster and then it successively merges or agglomerates the two points that are the most similar until all cluster points have been merged into a single cluster that includes all the data.
- Divisive clustering represents the top-down approach. I is not needed to specify the number of clusters in this algorithm as well. This approach requires a procedure for splitting the cluster which contains all the data. It starts with splitting the least similar clusters recursively until single data points have been split into a singleton cluster (Salvador et al, 2004).

Spectral Clustering

Spectral clustering is a clustering technique that aids with roots in graph theory, where this approach is used to obtain communities of nodes in a graph. This identification is based on the edges that connect the nodes. This is a flexible method that provides a method to cluster the non-graph data as well.

In this clustering technique, the data points are treated as nodes of a graph. Hence, spectral clustering is a graph partitioning problem. Then the nodes are mapped to a low-dimensional space that can be segregated easily to form the clusters. The main objective of spectral clustering is to cluster the data that is connected but not necessarily compact or clustered within convex boundaries.

Spectral clustering employs information from the eigenvalues (spectrum) of special matrices built from the graph or the data set. Figure 29 depicts the difference of Spectral Clustering and K-means Clustering.

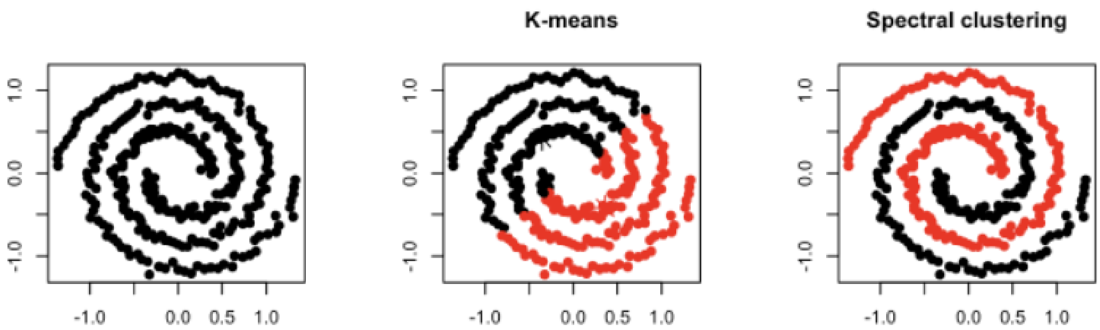

Figure 29: Spectral Clustering vs K-Means Clustering

When comparing the kmeans and spectral clustering method, kmeans is based on the Compactness clustering approach where the data points that lie close to each other will be group in the same cluster and these points are compact around the cluster center.
Spectral Clustering is based on Connectivity clustering approach where the data points that are connected or immediately next to each other will be fall in the same cluster. Whether the distance between two points is smaller, the two points will not be in the same cluster if they are not connected to each other. The main three steps involved in spectral clustering are respectively, constructing a similarity graph, projecting the data into a lower-dimensional space, and clustering the data.

First the algorithm forms a distance matrix from the given data points. Then the distance matrix will be transformed into an affinity matrix - A. Affinity metric determines the closeness of two points. Then the Degree matrix - D and the Laplacian matrix - L will be computed ($L = D - A$). Next step of this algorithm is to find the eigenvalues an eigenvectors of L. Then a matrix will be formed with the eigenvectors of k largest eigenvalues which computed from the previous step. Then this algorithm normalizes the vector to obtain the clusters of the data points in k-dimensional space.

Gaussian Mixture Models

The Gaussian Mixture Models (GMMs) are probabilistic models that suppose in the dataset there will be a certain number of Gaussian distributions and assume that the clusters can be represented by these distributions. Therefore, a Gaussian Mixture Model aims to group the data points that belong to a single distribution together. In other words, the GMMs tend to model the dataset as a mixture of several Gaussian Distributions. This is the core idea of the Gaussian Mixture Model. This model utilizes the soft clustering technique for assignment of data points to the Gaussian distributions.

In one dimensional space, the probability density function of a Gaussian distribution is given by:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where, μ describes the mean and σ2 describes the variance.

For two-dimensional space Gaussian distribution, the probability density function is given by:

$$f(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left[-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right]$$

Where, x, μ and Σ are respectively the input vector, 2D mean vector and 2×2 covariance matrix.

Above function can be generalized for d-dimensions where, the multivariate Gaussian model could have x and μ as vectors of length d, and Σ could be the *d x d* covariance matrix.

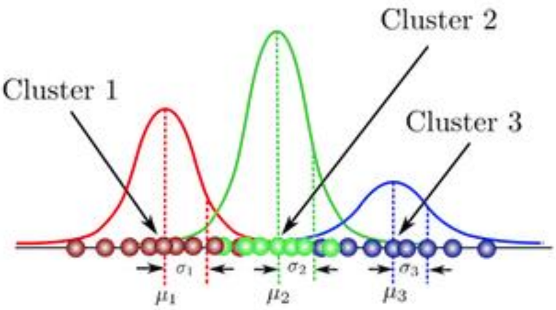

Figure 30: Gaussian Distributions

Therefore, for a dataset with d features, there could be a mixture of k Gaussian distributions where, k represents the number of clusters and each of these clusters have a certain mean vector and variance matrix.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

This is a base algorithm of the Density-Based Clustering approach. As the name implies DBSCAN algorithm can discover distinctive groups or clusters of various shapes and sizes from a large amount of data containing noise and outliers. This algorithm is constructed on the assumption that a cluster in the data space is a contiguous region of a high point density where it is separated from other clusters by the contiguous regions of low point density. DBSCAN groups the data points which are 'densely grouped' into a single cluster. This algorithm can identify clusters in large spatial datasets by observing the local density of these data points (Xu et al, 2008).

The most important feature of DBSCAN clustering is that it is robust to outliers. And it is not required to pre-specify the number of clusters for the algorithm. The DBSCAN requires only two parameters which are the epsilon and the minPoints. The parameter epsilon is used to represent the radius of the circle which will create around each data point to check the density. The parameter MinPoints is used to represent the minimum number of data points required inside that circle for that data point to be classified as a Core point. In the higher dimensions, the circle would be a hypersphere and the epsilon would be the radius of that hypersphere, and minPoints would be the minimum number of data points required inside that hypersphere.

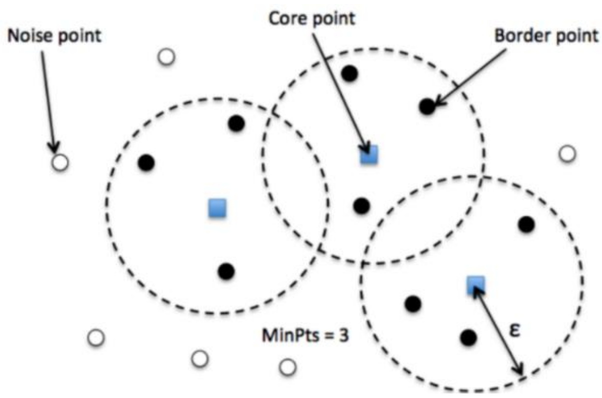

Figure 31: DBSCAN Clustering

As depicted in figure 31, DBSCAN provides three types of points when the clustering is complete. This algorithm creates a circle of epsilon radius around each and every data point and then it classifies them into Core point, Border point, and Noise. A data point will become a Core point if the circle around it contains at least 'minPoints' number of data points. If the number of data points is less than the minPoints, then it will be classified as a Border Point, and if there are no other data points around any data point within an epsilon radius, then it will be treated as a Noise point.

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

BIRCH Algorithm is a scalable clustering technique which is based on hierarchy clustering and this algorithm only requires scanning the database for one time. Thus, this method is a fast technique when working with large datasets. This clustering method has four main phases namely, scanning the data into memory, condensing or resizing the data, further the global clustering, and refining clusters. This is mainly based on the clustering feature trees(CF trees). Furthermore, this algorithm employs a tree structures summary to perform the clustering of the data. This tree structure produced through the BIRCH algorithm is known as the CF tree.

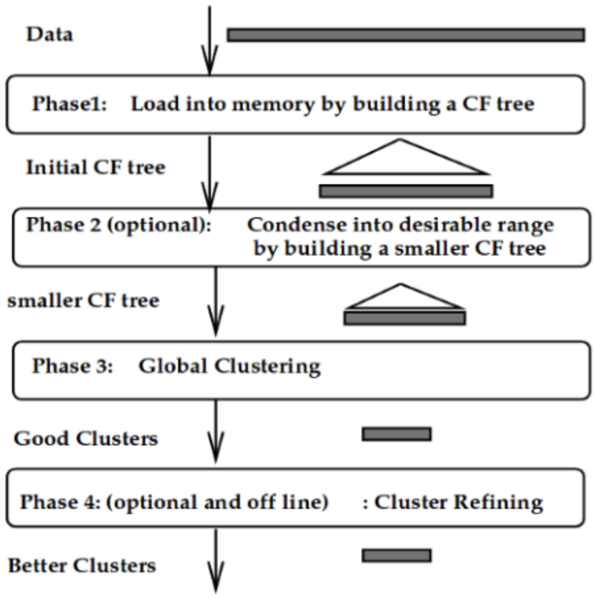

Figure 32: BIRCH Overview

As shown in the figure 32, this algorithm compresses the input data into a set of CF tree nodes. Further, those nodes will have many sub-clusters, and these are known as the CF subclusters. Then in the next phase these subclusters will be grouped into larger clusters and as a result this will produce an overall smaller CF-tree. In the global clustering phase almost any of the clustering algorithm can be applied to cluster the features rather than the data points. The final phase includes the steps of correcting the inaccuracies that been caused by applying the clustering algorithm to the summary of coarse data. Also this phase includes detecting and removing the outliers as well.(Anthony D. and Joana A., 2018)

As described in this section these six selected unsupervised machine learning algorithms in this study the clustering algorithms will be applied and trained on the targeted customer dataset for cluster analysis and cluster the unlabeled dataset. Afterwards, the results of cluster analysis and performance will be evaluated of these six different models. The internal cluster evaluation techniques will be used to evaluate the cluster validity results of the clustered obtained through training these models.

The most performed and most accurate clustering algorithm will be selected to interpretation of the cluster results and to gain the insight of the customer profiles according to the customer segmentation results.

Finally, the resultant clustered dataset obtained through the above phase will be utilized to build the prototype of predicting the customer segmentation of the given data input according to the obtained cluster results. For the predicting system supervised machine learning algorithms will be used to building the predictive model.

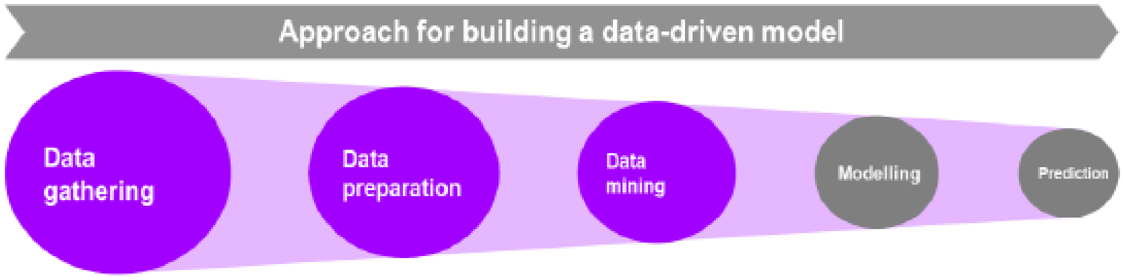

Figure 33: Approach for building a data-driven prediction model

Then the most accurate model will be used to build the prototype of the cluster predicting system. The following supervised machine learning models will be trained on the segmented customer dataset. The labels or the targets will be the cluster results obtained through training the clustering algorithms. The Figure 33 depicts the approach using for the prediction model.

3.4.2 Customer Segment Prediction Model Building

In the following section, the definitions of selected multiclass classification algorithms and their factors are briefly described. These algorithms will be used to solve the classification problem of labelling the future customers into one of the identified customer segments.

Multinomial Logistic regression

Logistic regression is one of the popular classification algorithms used for numerous domains such as traditional statistics, social science and medicine fields. This algorithm is used for binary classification problems in which there will be two classes, nevertheless this algorithm can be extended to solve multiclass classification problems. Multinomial logistic regression is the extended version of the logistic regression algorithm. This multiclass classification algorithm can be utilized to classify more than two categories of the dependent or targeted variable. Same as the binary logistic regression, multinomial logistic regression also uses the maximum likelihood estimation to evaluate the probability of categorical outcome. All the probabilities are non-negative values, and the sum is equal to one. (Osborne, 2012)

Decision Tree

Decision tree is a supervised learning technique which can be used for classification and regression. This can be utilized for predictions for a target variable employing a tree shaped learning process, which are known as the decision rules. This algorithm infers the target class labels along with a set of sequential if-else statements that are repeated with the features. This is a powerful learning technique for the reason that this algorithm can be applied for any non-linear relation. As shown in figure 34, the decision tree algorithm is learning from the data which approximate a sine curve with the decision rules that is a set of if-then-else statements. When the tree is deeper, the decision rules will be complex thus means a fitter model (Sawicz et al., 2014)

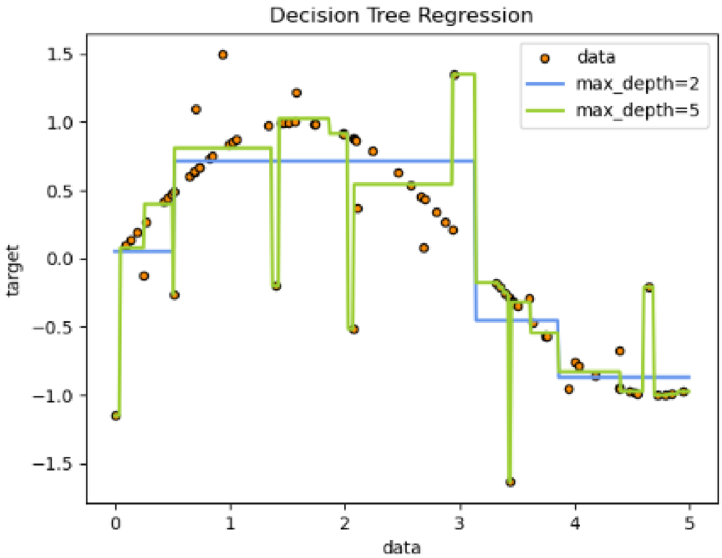

Figure 34: Decision Tree, learning sine curve

Random Forest

Random Forest is also known as random decision forests is a powerful machine learning method used in classification and in regression. This algorithm employs an ensemble learning method where many classifiers will be combined to provide the solution for complex problems. In classification this ensemble algorithm produces the output which is the class chosen by most of the trees in the forest.

The forest is generated from the random forest algorithm is trained using bagging or rather bootstrap aggregating. Bagging is an ensemble meta-algorithm in machine learning, and this is used to improve the accuracy of the algorithms. Further this is used in random forest to improve the accuracy of the algorithm. Thus, this algorithm holds unexcelled in accuracy among other machine learning algorithms, further this algorithm is capable of handling a broad amount of data effectively and efficiently (James, et al., 2013).

The following high-level architectural diagram shown in figure 35 illustrates the hybrid approach model which utilizes both the unsupervised learning model(ULM) and supervised machine learning model (SLM) of the proposed system.

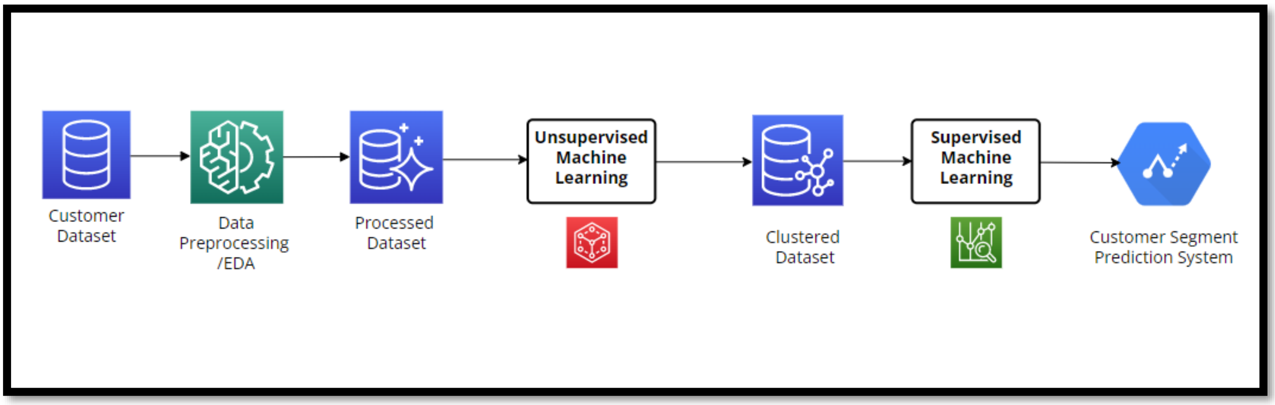

Figure 35: General design architectural Diagram of the proposed system

The prototype of the customer segmentation prediction application is principally based on the cluster analysis of the customer dataset. The raw customer dataset will be preprocessed before feeding into the cluster analysis. Afterwards, the resultant dataset of the cluster analysis provides the capability for applying the supervised learning models where the obtained clusters will be the target label of the dataset. The class imbalance handling, and necessary splitting dataset are performed on the clustered dataset before training the prediction model.

This chapter comprised with the methodology of the research project and the upcoming chapter will discuss the evaluation and results of the research study.

CHAPTER 4: EVALUATION AND RESULTS

In this chapter, the evaluation of the analysis with the results obtained from the research project, the evaluation metrics used to assess the implemented system, cluster analysis, and insight into the customer segments are explained. Afterward, a comparison of the machine learning models is presented and discussed. Furthermore, the prediction modeling using supervised learning models is discussed and evaluated.

4.1 Evaluation of the Results

4.1.1 Correlation Matrix with Heat Map

The term Correlation is a statistical measure that measures the strength of a relationship among two variables. When the correlation is positive means that the two variables are moving in the same direction which means that when one variable increases, the other variable is also increasing relatively. On the other hand, when a correlation is negative means that the two variables are moving in the opposing directions i.e., when one variable increases, the other variable will decrease. A correlation matrix is analyzed in the study to recognize the relations with the attributes of the customer dataset.

The figure 36 presented the correlation between independent and dependent variables. As shown in the figure, the value in each square depicts the correlation among variables in the dataset and as depicted above the values are ranging from -1 to +1 where negative values represent a negative correlation between the attributes and the positive values represent a positive correlation between the attributes. From the above correlation matrix below relationships can be identified.

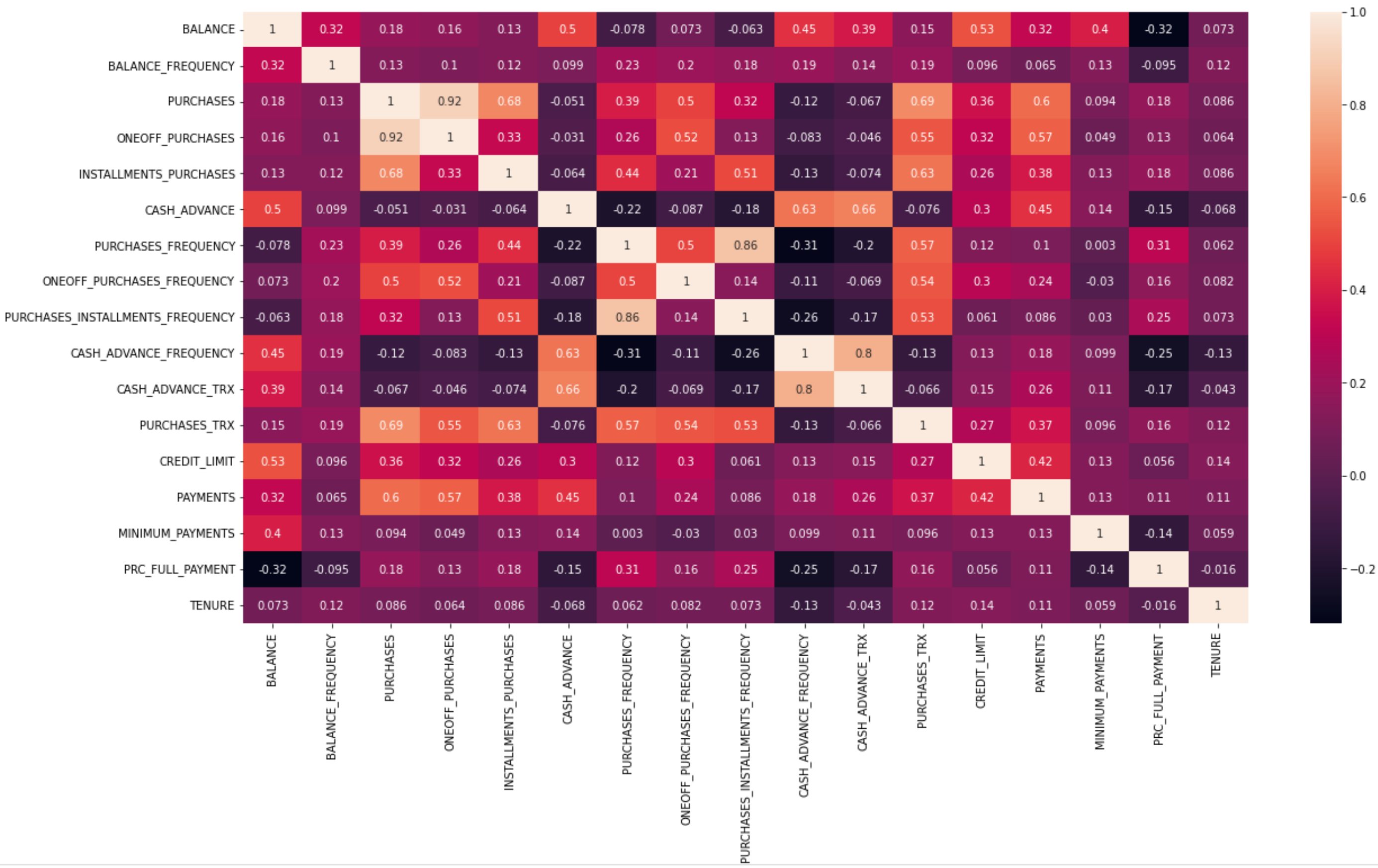

Figure 36: Correlation Matrix with Heat map

- Balance has a higher level of correlation with Cash Advance, Credit Limit and Cash Advance Frequency
- Payments variable has a high correlation with Purchases and one-off Purchases
- Tenure has a negative correlation with Cash Advance and Cash Advance Frequency variables
- Purchases, one-off purchases, and installment purchases are highly correlated.
- Customers don't make full payments when the balance is high
- Purchase frequency feature and cash advance frequency feature are negatively correlated. That is as the purchase frequency is high, the number of times cash is paid in advance is less and vice-versa

4.1.2 PCA Analysis

PCA is a dimensionality reduction method, and this can be used to reduce the dimensionality of highly interrelated variable by transforming the variables in the dataset into a new set of variables, these newly constructed variables are known as the principal components. Since the customer dataset is highly correlated and has many features PCA analysis is performed as a dimensionality reduction method to recognize the hidden patterns of the dataset by alleviating the variances.

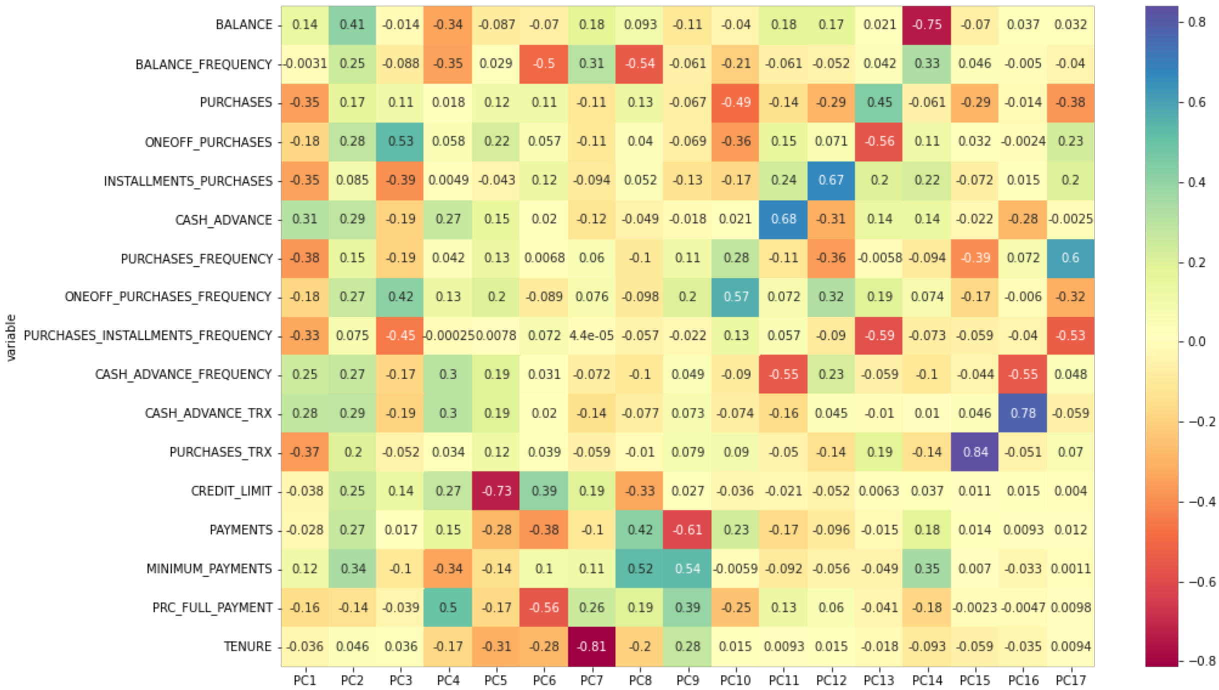

Figure 37: Correlation matrix plot for component loadings

Generally, this method follows the technique of feature extraction. Figure 37 depicts the correlation matrix plot aids for component loadings. Here, both the positive and negative measures in the component loadings reflect the positive and negative correlation of the variables with the PCs, and the loadings represent the covariances/correlations among the original variables and the unit-scaled components.

The PCA method process a computation of the eigenvalue decomposition using an estimation value of the covariance matrix of the given data set and then this method uses the most essential eigenvectors for projecting the feature space into a lower dimension space.

```
array([0.34607901, 0.19545989, 0.1167744 , 0.08215483, 0.06243788,
       0.04240659, 0.0362757 , 0.02892916, 0.02611924, 0.01845825,
       0.0116047 , 0.00970836, 0.00807757, 0.00608579, 0.00306995,
       0.00244759, 0.00176523])
```

Figure 38: Eigenvalues; Variance explained by each PC

Figure 38 depicts the calculated eigenvalue for each PC explaining the variance ration of PC1 to PC17. Figure 39 depicts a cumulative value of the eigenvalues of each PCs.

```
array([0.34682325, 0.54270347, 0.65972899, 0.74206049, 0.80463264,
       0.84713042, 0.88348414, 0.91247551, 0.93865092, 0.95714887,
       0.96877852, 0.97850776, 0.9866027 , 0.99270158, 0.99577813,
       0.99823098, 1.        ])
```

Figure 39: Cumulative proportion of variance (from PC1 to PC17)

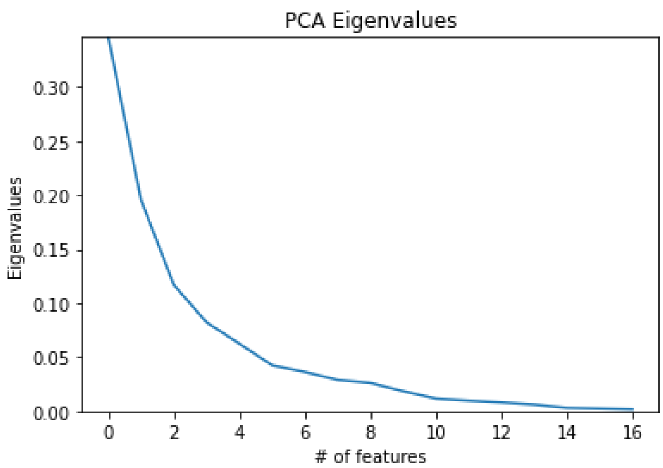

Figure 40: Scree Plot

The scree plot depicted in figure 40 helps to identify the number of PC components of the customer data set that explain most of the variation in the data. In PCA analysis, a scree plot is used to depict a line plot of the eigenvalues of the principal components.

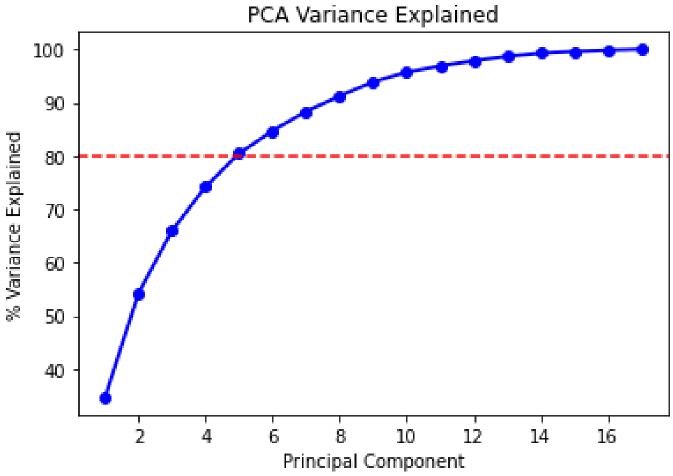

Figure 41: PCA Variance graph

After the analysis of the PCA results, the acceptable variance level is decided for the customer dataset. As shown in figure 41, in this study an 80% of variance is considered and the number of PC components extend up to PC5.

The table 5 describes the reduced five principal component factors along with the features of the customer dataset.

Table 5: Reduced PC factors

index	PC1	PC2	PC3	PC4	PC5
BALANCE	0.14164904396443556	0.4132382025118417	-0.01390964082332927	-0.33983071200944975	-0.08667535586118791
BALANCE_FREQUENCY	-0.003117057555804617	0.2505926485810462	-0.08753631763962938	-0.3480956583611293	0.028581951862876365
PURCHASES	-0.3503498292171746	0.16628614203740488	0.10807595472460137	0.017864896261481716	0.11662857468739657
ONEOFF_PURCHASES	-0.17849668845740535	0.2847565084518162	0.5254178955630745	0.057905456284979735	0.22147127296024952
INSTALLMENTS_PURCHASES	-0.34699867944000845	0.08500934047715519	-0.3921643313445479	0.004925923502626286 5	-0.04295424795079349
CASH_ADVANCE	0.31433993469042903	0.2890499156868187	-0.18692859903237996	0.2692267353620819	0.1532552696220159
PURCHASES_FREQUENCY	-0.3786266697482561	0.14643981287452965	-0.18875761073608543	0.042149105885884317	0.12941460911342567
ONEOFF_PURCHASES_FREQUENCY	-0.17973434593559642	0.2651732174621992	0.4183262684088333	0.1285090932054404	0.19973892934832213
PURCHASES_INSTALLMENTS_FREQUENCY	-0.3348168453283325	0.07540064535076474	-0.45062012778544885	-0.00024787916700212687	0.007795605212204118
CASH_ADVANCE_FREQUENCY	0.25169678608202944	0.26745610641788187	-0.17415646931113163	0.30072286231969514	0.192899098813578
CASH_ADVANCE_TRX	0.27984999525803433	0.2908541798489771	-0.19094402709552194	0.297778457475886 7	0.19402182112249655
PURCHASES_TRX	-0.36667081558407033	0.20051943375409517	-0.05166657908454568	0.033628182760228456	0.11743457692431206
CREDIT_LIMIT	-0.03818264279498699	0.25144047996530583	0.13724193363421786	0.2739268739526156	-0.7293545504326445
PAYMENTS	-0.027956696487885166	0.2653307363961612	0.01744961103348211	0.14993463040959507	-0.28317490214540336
MINIMUM_PAYMENTS	0.11589567072557326	0.34410574225919005	-0.10306241835154252	-0.33781733382303447	-0.14305832951624203
PRC_FULL_PAYMENT	-0.15890095170954258	-0.1379378684904534	-0.03869631779264317	0.4987600397909028	-0.1698261887391823
TENURE	-0.03607747943584342	0.04562577179956811	0.03550964722174031	-0.16804240455429478	-0.3112616155589269

4.1.3 Clustering Model Analysis

In this section, the cluster analysis evaluation is presented with the distribution and obtained cluster results through training six different clustering algorithms.

4.1.3.1 K-Means Clustering

The hyperparameter tuning is performed on the final processed dataset to obtain the optimal number of clusters for the K-Means clustering algorithms. And for other clustering algorithms similar evaluation performed in order to identify the optimal number of clusters.

As shown in figure 42, an elbow plot graphs the WCSS value against the no of clusters. When analyzed the graph we can determine that the graph is rapidly changing at the point of 4 and this point is treated as the elbow point of the graph. After this point the WCSS value starts to decrease.

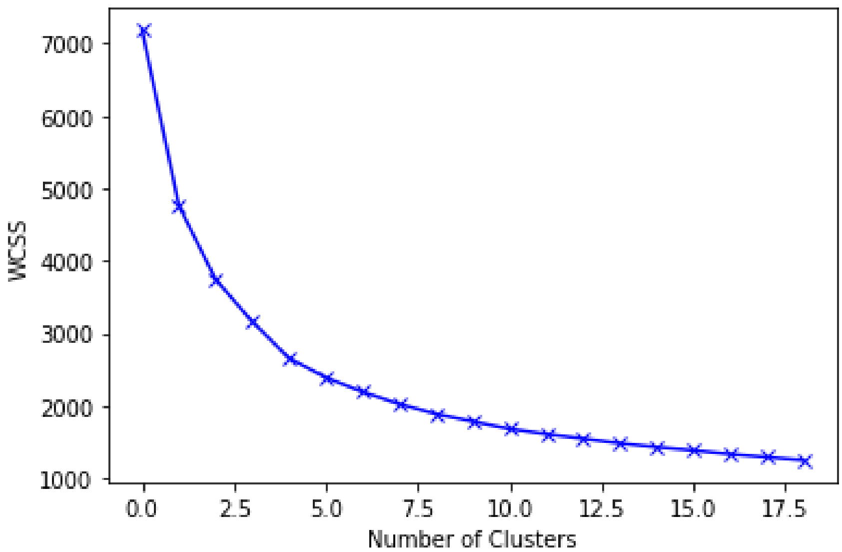

Figure 42: Elbow Plot

For further analysis, as shown in figure 43 an Elbow plot with the Silhouette score is depicted for the K-Means clustering algorithm and from this graph we can identify that the optimal value for k is 4 for the clustering the customer dataset.

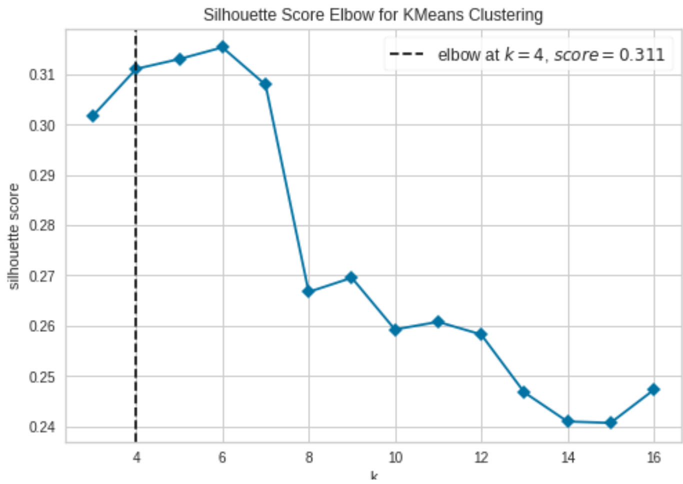

Figure 43: Silhouette score Elbow plot

After deciding the optimal value for the clustering algorithm, the K-Means clustering algorithm is applied to the customer dataset and the cluster result is shown in figure 44. PC1 and PC2 used for visualization of clusters 2D plot.

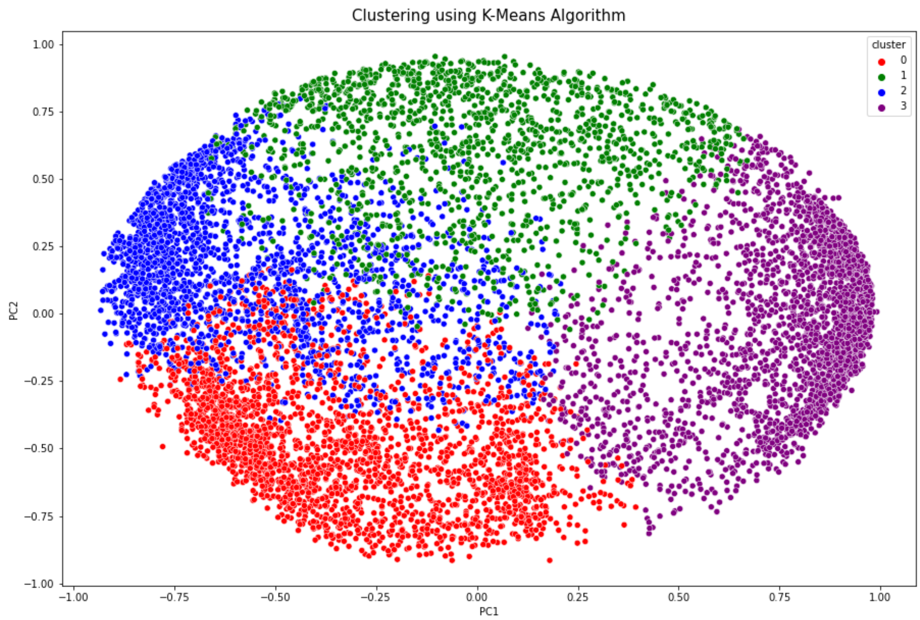

Figure 44: Cluster result obtained through K-Means

According to the obtained result, the data distribution among 4 clusters is shown in figure 42 and for the cluster 0 there are 2532 customers, for the cluster 1 there are 1489 customers, for the cluster 3 there are 2712 customers, and further cluster 3 has a total of 2712 customers.

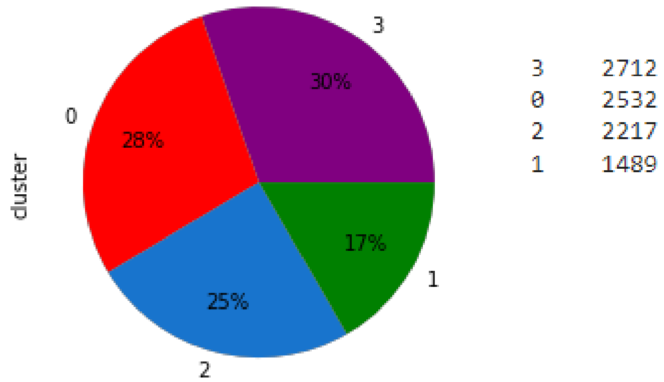

Figure 45: Customer data distribution among clusters - K-Means clustering

According to the obtained result, the data distribution among 4 clusters is shown in figure 45 and for the cluster 0 there are 2532 customers, for the cluster 1 there are 1489 customers, for the cluster 3 there are 2712 customers, and further cluster 3 has a total of 2712 customers. Figure 46 depicts the 3D plot of the clusters obtained using PC1, PC2 and PC3.

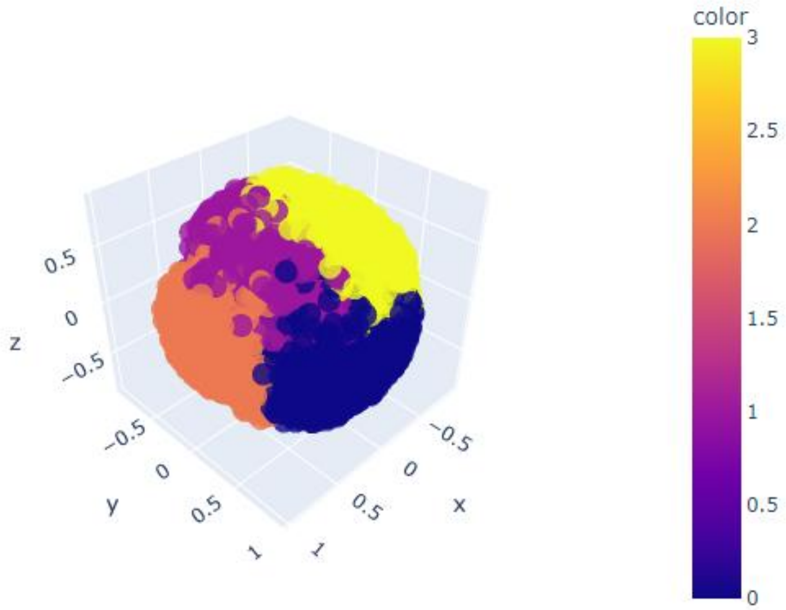

Figure 46: 3D plot of cluster visualization(KMeans)

4.1.3.2 Hierarchical Clustering

Hierarchical clustering produces a hierarchy of clusters when performed on the given dataset. Generally, the merges and the splits of the hierarchy are persevered in a greedy manner. This can be visualized by using a dendrogram which is a tree-like diagram. Dendrogram records the sequence of merges or splits of the clusters. Figure 47 depicts the dendrogram used to divide a cluster of data into many different clusters.

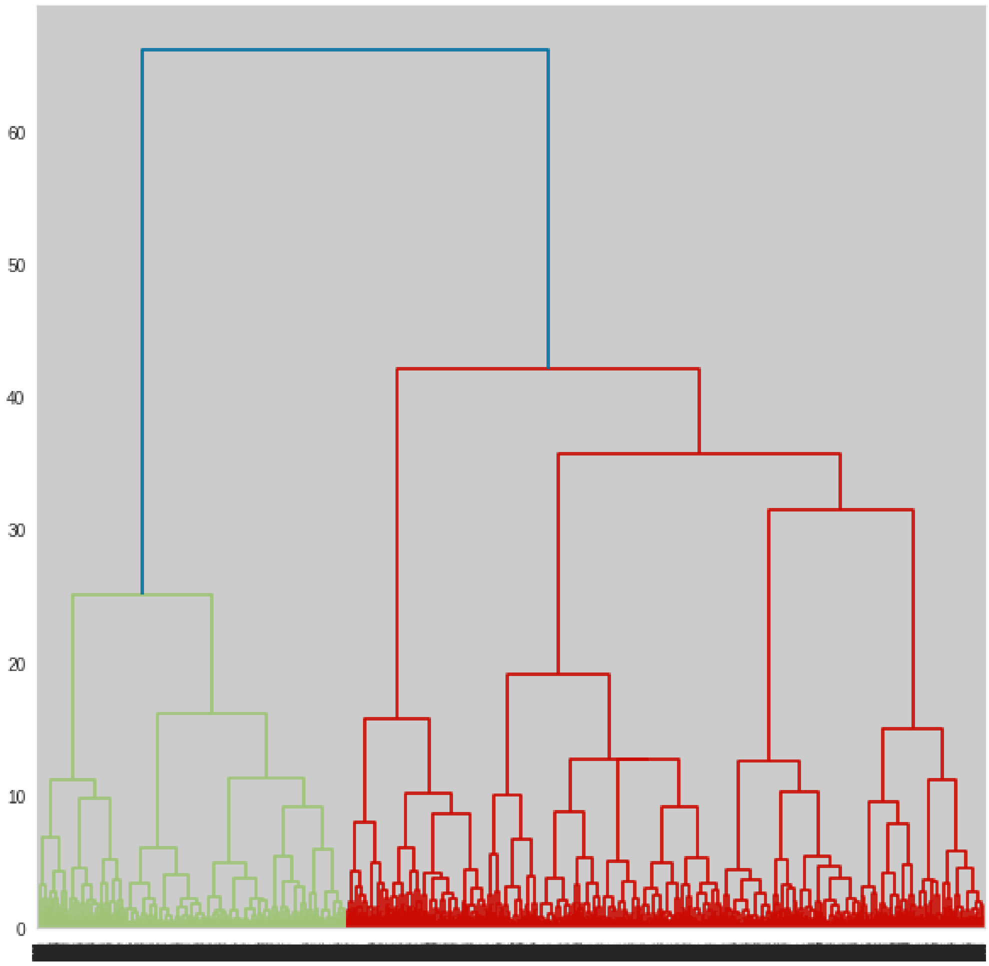

Figure 47: Dedrogram

The bottom-up approach of the hierarchical clustering is called Agglomerative clustering and this clustering algorithm is used for cluster analysis of the customer dataset.

The optimal number of clusters obtained through hyperparameter tuning before performing the agglomerative clustering algorithm on the processed dataset.

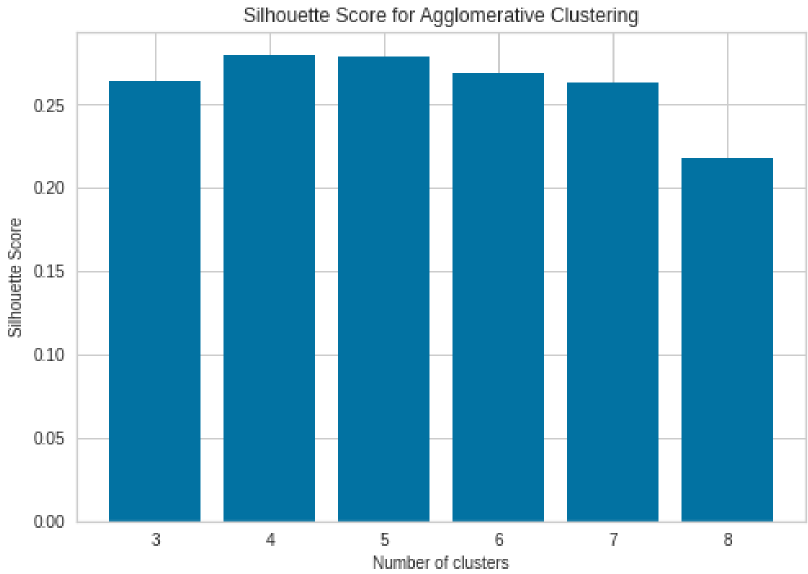

Figure 48: Silhouette score of Agglomerative clustering

The figure 48 depicts the Silhouette score obtained for Agglomerative clustering to determine the optimal number of clusters before performing the clustering and according to the plotted bar graph we selected 4 as optimal number of clusters. As mentioned in the chapter 2 literature survey the higher silhouette score depicts a better clustering.

When the Silhouette Coefficient is high, it indicates that the object is properly matched or more similar to its own cluster on the other hand the object of the cluster is poorly matched or dissimilar to its neighboring clusters. When the number of clusters equal to 4 the Silhouette score is higher according to the above graph. Hence, the no of clusters selected to be 4 for the cluster analysis.

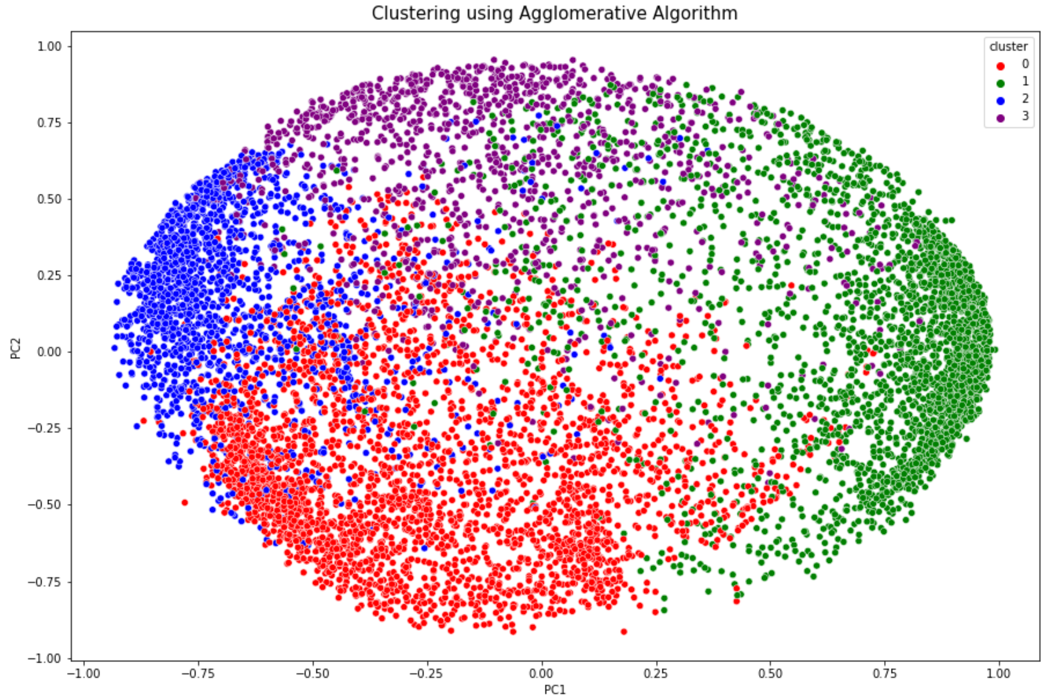

Figure 49: Cluster result obtained through Agglomerative clustering

After deciding the optimal value for the clustering algorithm, the agglomerative clustering algorithm is applied to the customer dataset and the cluster result is shown in the figure 49. PC1 and PC2 used for visualization of clusters 2D plot. Figure 50, pie chart depicts the distribution of data among four different clusters.

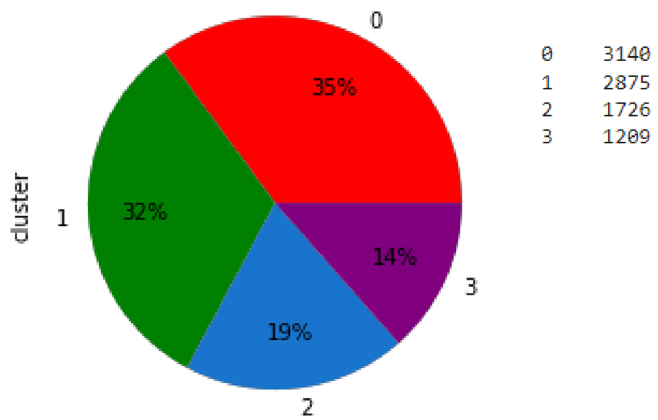

Figure 50: Customer data distribution among clusters - Agglomerative clustering

4.1.3.3 Spectral Clustering

The optimal number of clusters is obtained through hyperparameter tuning before performing the spectral clustering algorithm on the processed customer dataset.

Figure 51: Silhouette score of Spectral clustering

The figure 51 depicts the Silhouette score obtained for Spectral clustering to determine the optimal number of clusters before applying the spectral clustering model to obtain the customer segments. When the number of clusters equal to 4 the Silhouette score is higher according to the above graph. Hence, the no of clusters selected to be 4 for the cluster analysis.

Here, two different spectral clustering models were trained with two different values for the affinity matrix parameter. As mentioned in the former chapter spectral clustering employs a graph called affinity matrix also known as adjacency matrix where the rows and columns represent the nodes of the graph. We have considered a Gaussian kernel and Euclidian distance for parameter affinity.

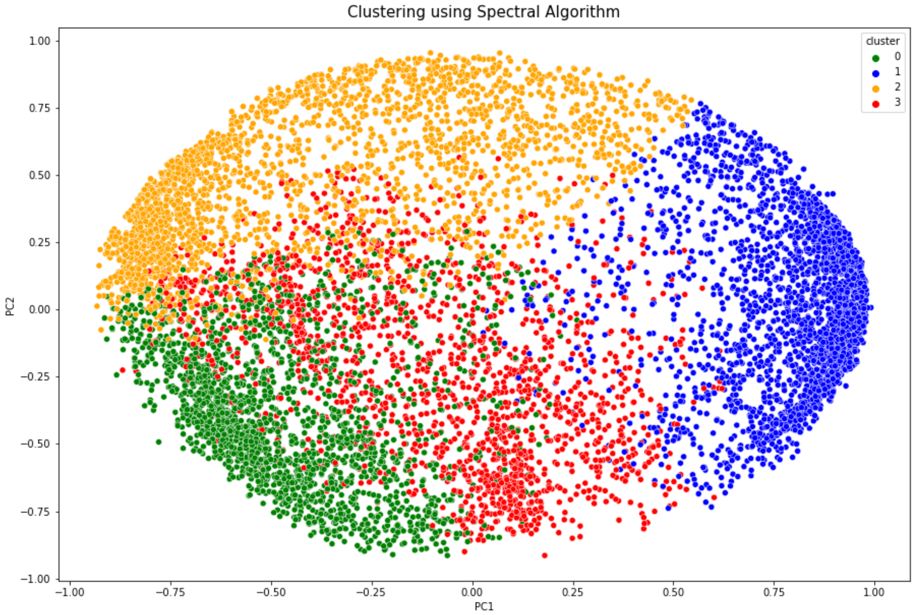

Figure 52: Cluster result obtained through Spectral clustering(rbf)

Figure 52 depicts the cluster results obtained through performing the spectral clustering model with Gaussian Kernel. As shown in figure 53, Cluster 0 has about 20% distribution of the customer data, cluster 1 has 23% distribution, cluster 2 has 29% distribution of data, and cluster 3 has 27% distribution of data according to the obtained cluster results.

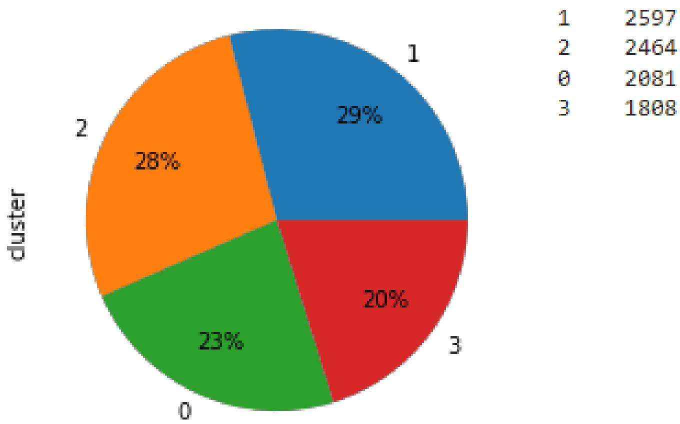

Figure 53: Customer data distribution among clusters - Spectral clustering (rbf)

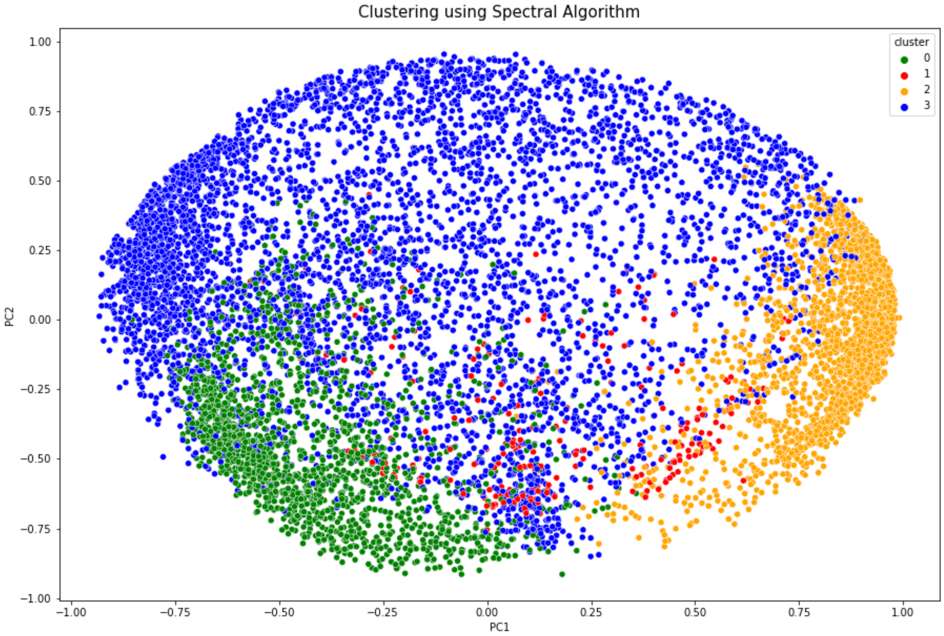

Figure 54: Cluster result obtained through Spectral clustering(nearest_neighbors)

Figure 54 represents the cluster results obtained by training the spectral clustering model with Euclidean Distance. As shown in figure 55, Cluster 0 has about 21% distribution of the customer dataset, cluster 1 has only 3% distribution, cluster 2 has 23% distribution, and cluster 3 has 54% distribution of data according to the obtained cluster results. We can identify that the spectral clustering with Gaussian kernel has a better distribution and is a fitter results by analyzing the obtained cluster results.

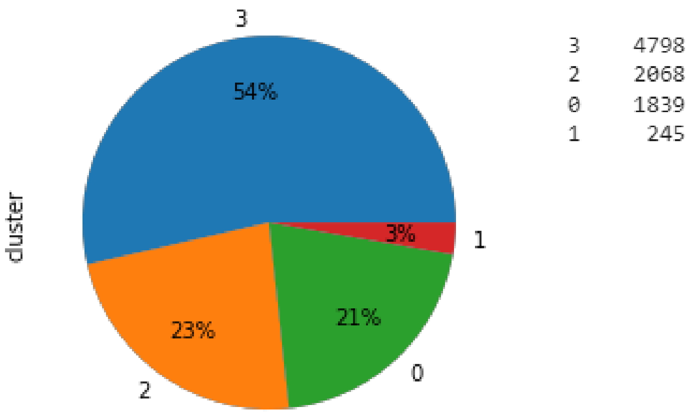

Figure 55: Customer data distribution among clusters - Spectral clustering (nearest_neighbors)

4.1.3.4 Gaussian Mixture Model

Hyperparameter tuning is performed on the processed customer dataset for Gaussian Mixture Based Clustering to identify the optimal number of clusters before training the cluster model.

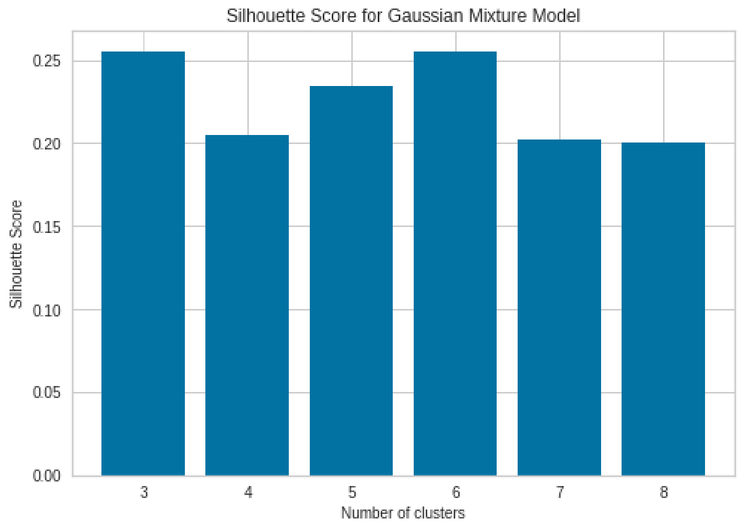

Figure 56:Silhouette score of Spectral clustering

As shown in figure 56, the Silhouette score has a maximum value when the number of clusters is equal to 3. Hence, we selected the optimal number to be 3 for the GMM based clustering.

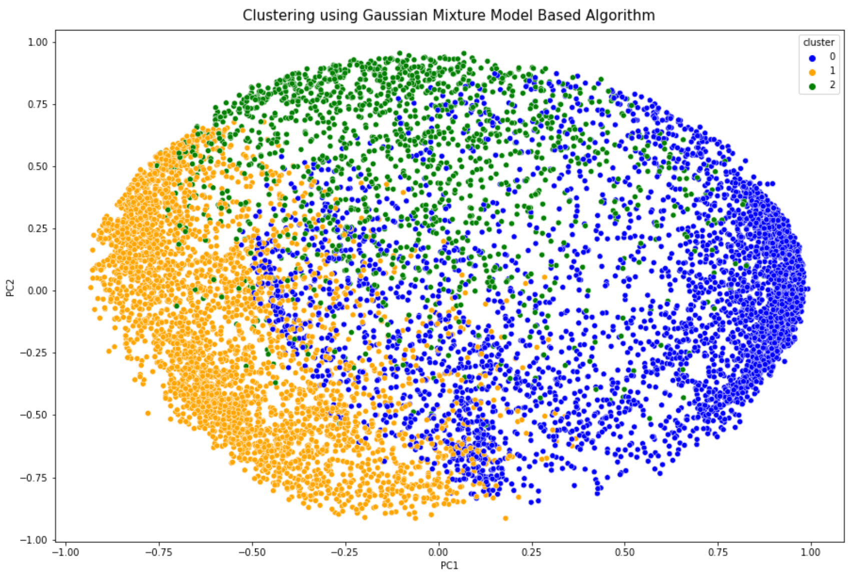

Figure 57: Cluster result obtained through GMM based clustering

Figure 57 depicts the cluster results obtained through GMM based clustering and the cluster visualization is done by the aid of PC1 and PC2 to in the 2D space. As shown in figure 58, Cluster 0 has about 44% distribution of the customer dataset, cluster 1 has ABOUT 40% distribution, and cluster 2 has 17% distribution according to the obtained cluster results.

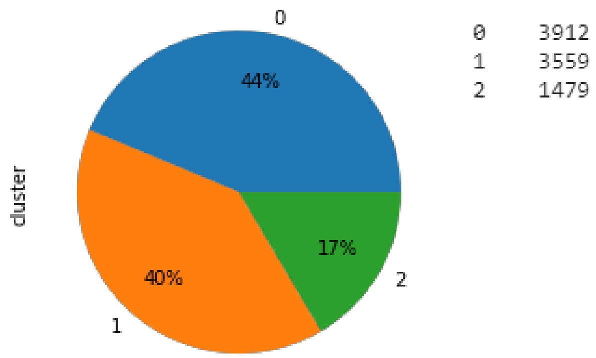

Figure 58: Customer data distribution among clusters - GMM based clustering

Figure 59 depicts the 3D plot of the clusters obtained through performing the GMM based clustering using PC1, PC2 and PC3.

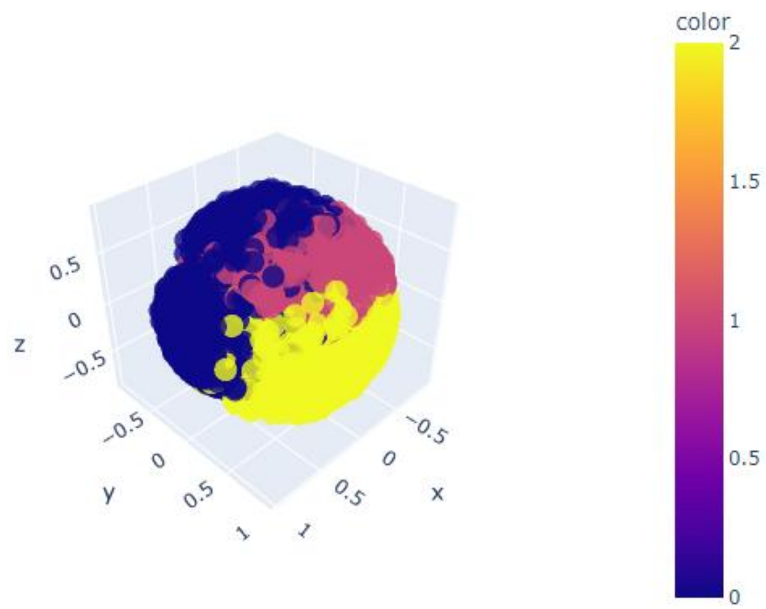

Figure 59: 3D plot of cluster visualization(GMM)

4.1.3.5 DBSCAN Clustering

DBSCAN algorithm does not require to determine the number of clusters for performing the clustering. This algorithm only requires two parameters which are the epsilon and minPoints.

The value for parameter minPoints has to be at least greater than the number of dimensions in the dataset. i.e., minPoints>=Dimensions+1. Hence the value for the 18 is selected as the value for parameter.

The value for parameter epsilon can be determined by plotting the K-Distance graph. The maximum point curve is the graph is selected as the value for the parameter epsilon.

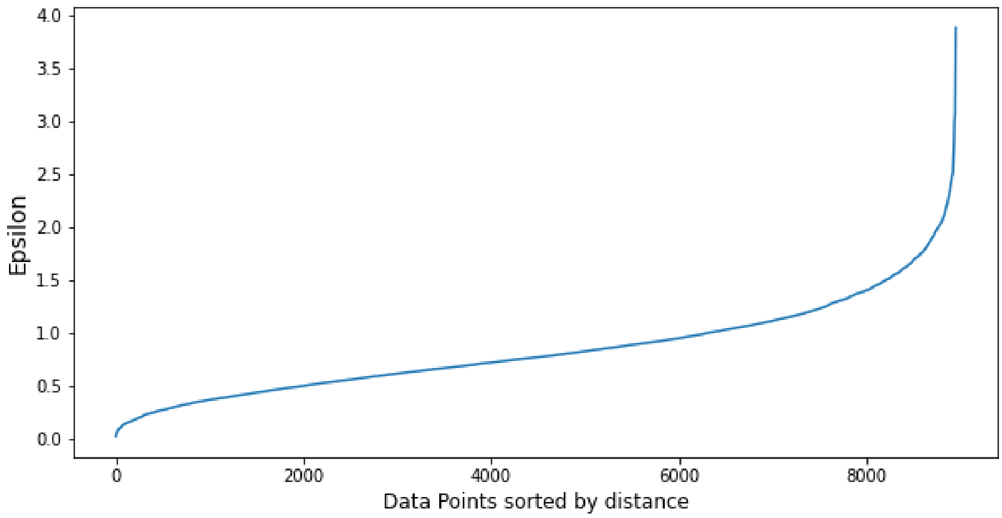

Figure 60: K-Distance Graph

The figure 60 shows the plotted K-Distance graph for the processed dataset. According to this graph the maximum curve point is at 1.8 value hence, this value is selected as the optimum value for parameter epsilon. Then using these two parameters DBSCAN clustering is performed on the customer dataset.

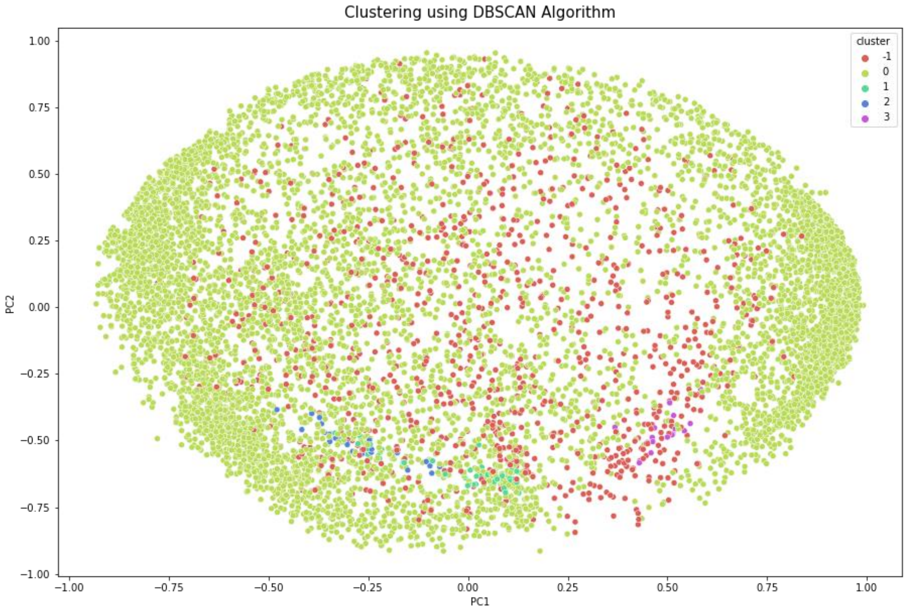

Figure 61: Cluster result obtained through DBSCAN clustering

The results obtained by performing the DBSCAN algorithm is shown in figure 61, according to the above result the customer base is clustered up to four different clusters. The above 2D plot is visualized using PC1 and PC2 components. We can identify that the distribution of clusters is not appealing when analyzing cluster distribution which is depicted in figure 62. 88% of the customer dataset is grouped as cluster 0 and 11% for cluster -1. The other cluster distributions are relatively small when compared to cluster 0.

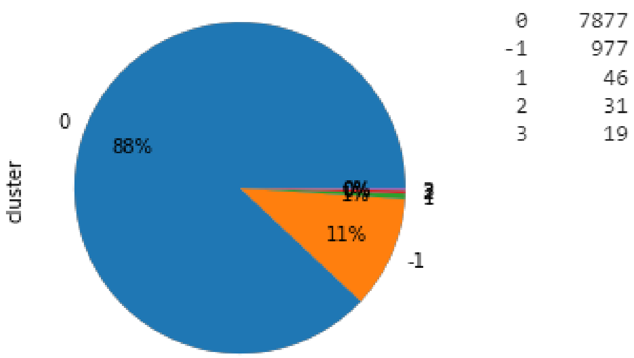

Figure 62: Customer data distribution among clusters - DBSCAN clustering

4.1.3.6 BIRCH Clustering

The process of hyperparameter tuning is performed on the customer dataset before training the BIRCH clustering model to identify the optimal number of clusters.

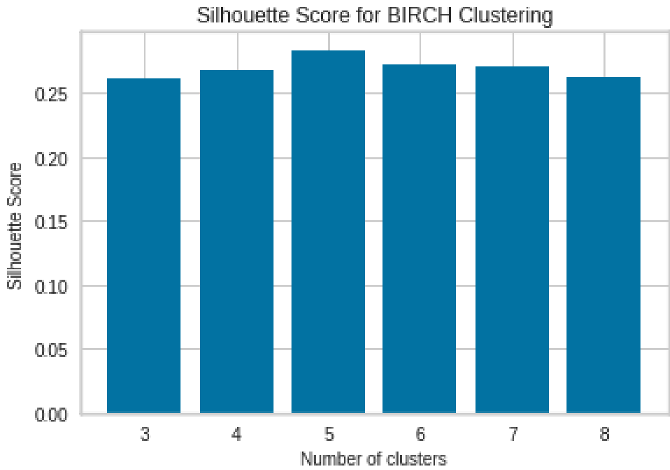

Figure 63: Silhouette score of BIRCH clustering

The result of the Silhouette score for number of clusters is shown in figure 63, the Silhouette score has a maximum value when the number of clusters is equal to 5. Hence, we selected the optimal number to be 5 for training the BIRCH clustering model.

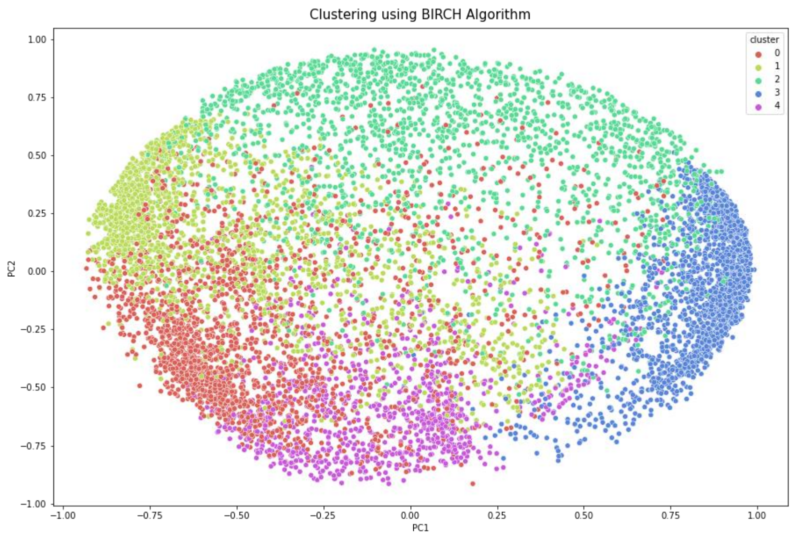

Figure 64: Cluster result obtained through BIRCH clustering

The cluster results obtained by performing the BIRCH algorithm is shown in figure 64. The cluster visualization is depicted using the PC1 and PC2 components for 2D space visualization.

To identify the distribution of the customer base among these five different clusters from the distribution pie chart was plotted as shown in figure 65.

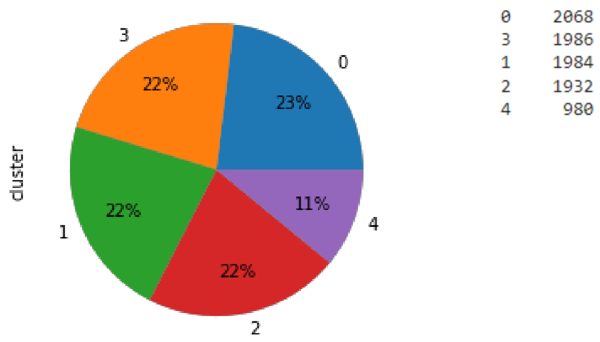

Figure 65: Customer data distribution among clusters – BIRCH clustering

According to the distribution chart the cluster 0 has 23% distribution of the customer data, cluster 1, cluster 2 and cluster 3 has 22% of the distribution, and cluster 4 has 11% of the distribution of the customers. Figure 66 depicts the 3D plot of the obtained cluster results using the PC1, PC2 and PC3.

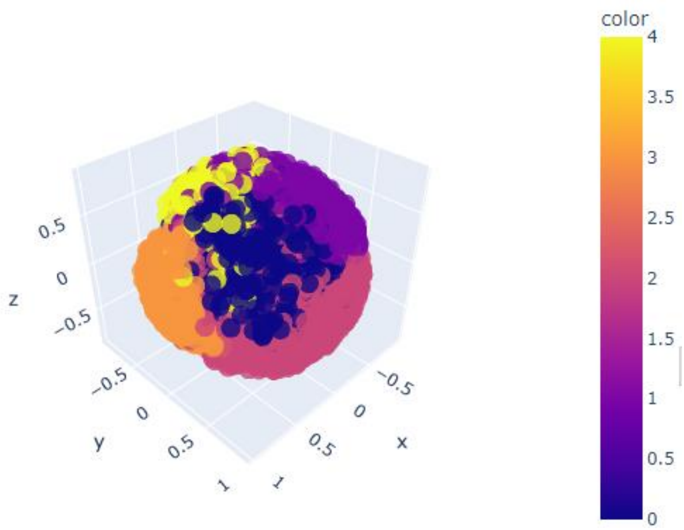

Figure 66: 3D plot of cluster visualization(BIRCH)

4.1.4 Cluster Results Evaluation and Interpretation

The clustering evaluation can be divided into two main types: External Measures and Internal Measures. The external measures are evaluated with the ground truth labels of the dataset and the internal measures are evaluate based on the cohesion and separation of the clusters. In this research project internal cluster analysis method is used to evaluate the cluster results as the customer dataset does not have any ground truth labels and this is purely based on unsupervised learning.

Table 6: Clustering Model Evaluation Results

Clustering Model	**Davies-Bouldin index**	**Silhouette score**	**Calinski & Harabasz score**
KMeans Clustering	1.3369	0.3042	3759.5100
Agglomerative Clustering	1.5085	0.2281	2806.1391
Spectral Clustering(rbf)	1.3062	0.3055	3690.4564
Spectral Clustering (nearest-neighbor)	1.6023	0.1208	1979.3887
Gaussian Mixture Model	1.4002	0.2576	3291.0562
DBSCAN Clustering	2.9476	-0.2817	42.2742
BIRCH Clustering	1.6386	0.1966	2391.1380

The cluster evaluation results are shown in the table 6 and according to the obtained evaluation results, the highest silhouette score value reported in Spectral clustering model with Gaussian kernel. The lowest Davies-Bouldin index is also reported in the Spectral clustering with Gaussian kernel model. Highest Calinski & Harabasz score is reported in K-Means algorithm, the second highest value is reported in Spectral algorithm(rbf).

As discussed in chapter 2 through inter cluster evaluation methods, the better clustering result will be achieved through higher Silhouette score, with a lower Davies-Bouldin index and with a higher Calinski and Harabasz score. Therefore, the best clustering analysis from the six different clustering models that trained on the same processed customer dataset is the Spectral clustering Model with the Gaussian Kernel. The cluster results obtained through spectral clustering is used to interpret the cluster results and to gain an insight about the different customer profiles in the customer dataset.

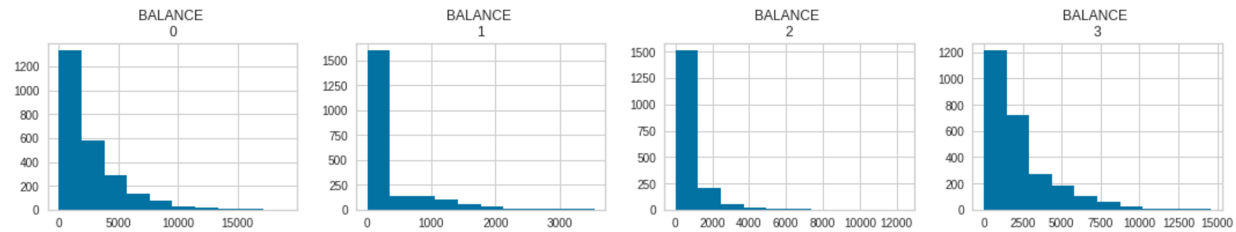

Figure 67: Plot comparison of Balance of each cluster

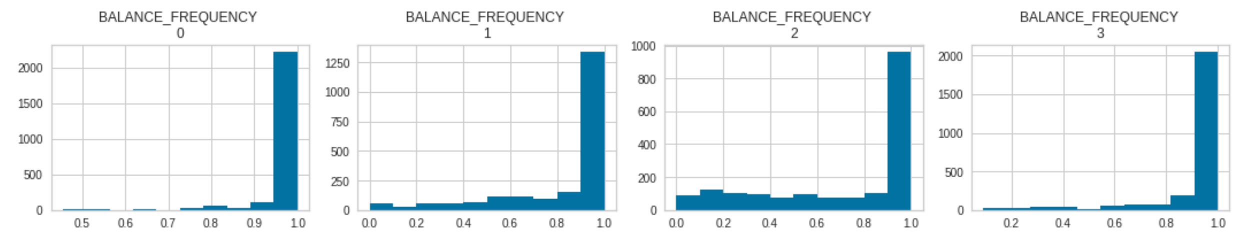

Figure 68: Plot comparison of Balance_Frequency of each cluster

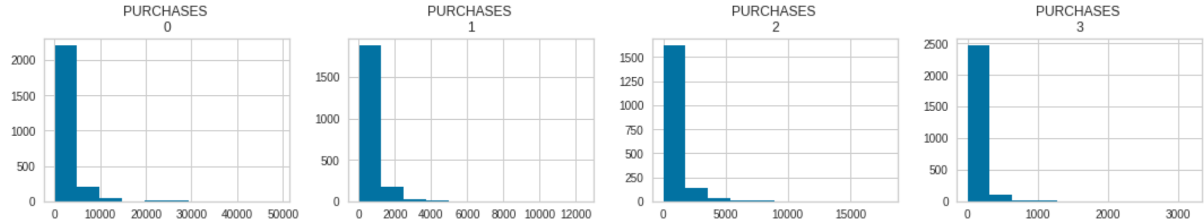

Figure 69: Plot comparison of Purchases of each cluster

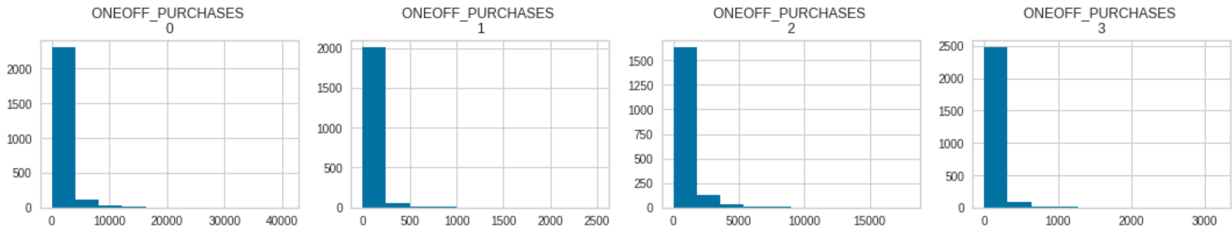

Figure 70: Plot comparison of One-off Purchases of each cluster

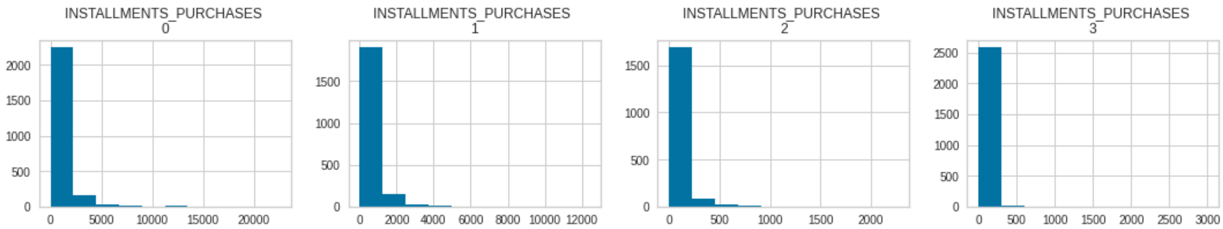

Figure 71: Plot comparison of Installments_Purchases of each cluster

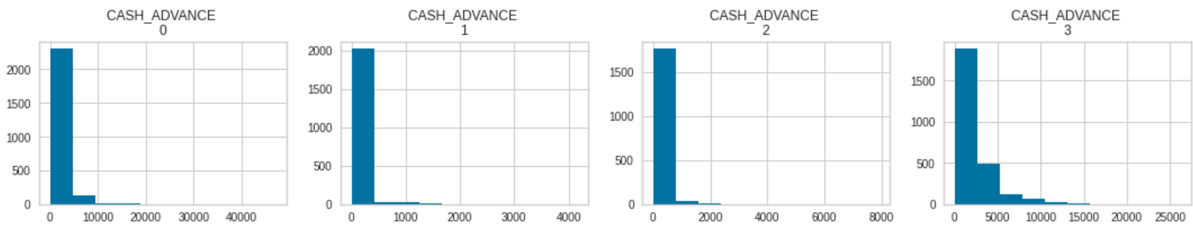

Figure 72: Plot comparison of Cash Advance of each cluster

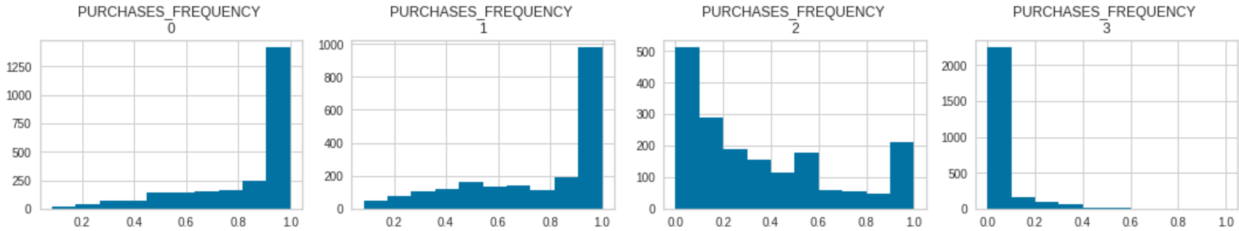

Figure 73: Plot comparison of Purchases_Frequency of each cluster

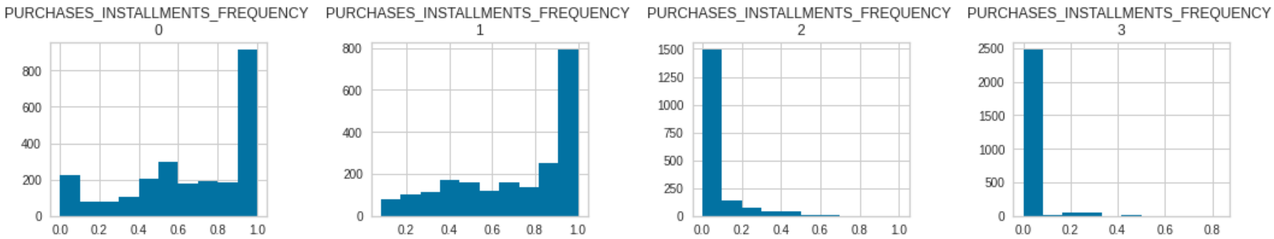

Figure 74: Plot comparison of Purchase_Installments_Frequency of each cluster

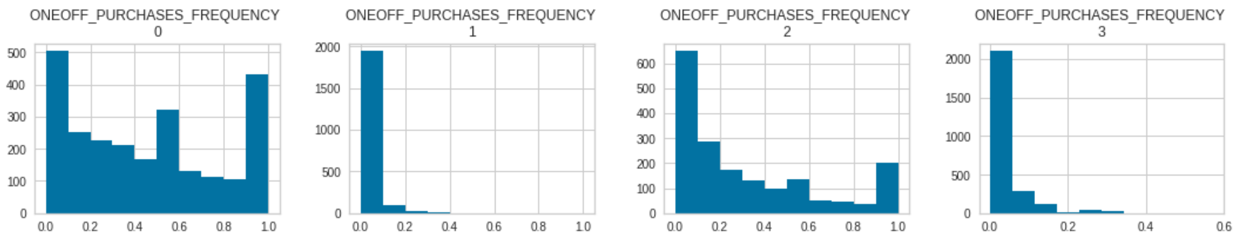

Figure 75: Plot comparison of One-off_Purchases_Frequency of each cluster

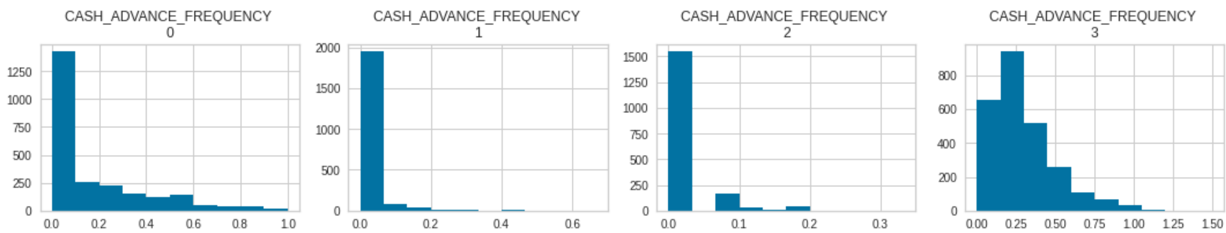

Figure 76: Plot comparison of Cash_Advance_Frequency of each cluster

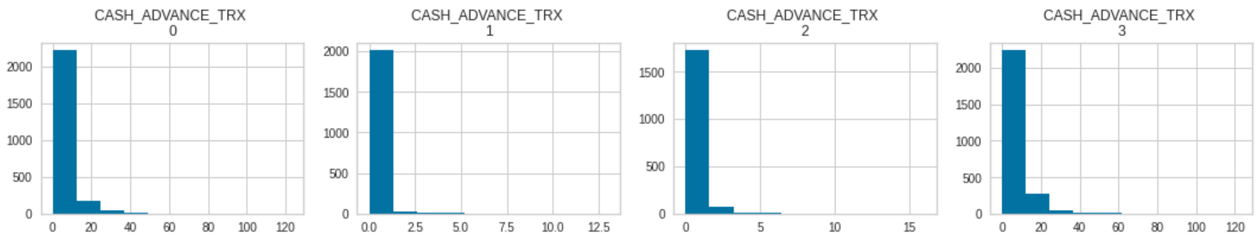

Figure 77: Plot comparison of Cash_Advance_Trx of each cluster

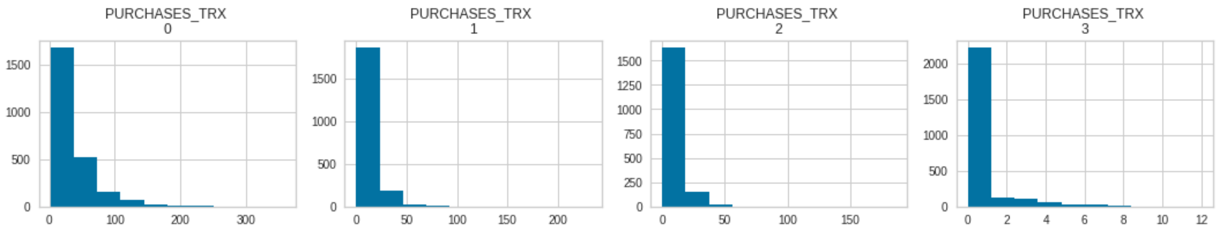

Figure 78: Plot comparison of Purchases_Trx of each cluster

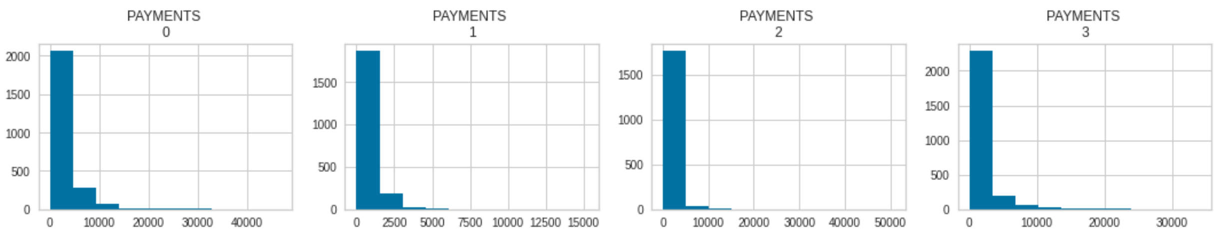

Figure 79: Plot comparison of Purchases_Tr

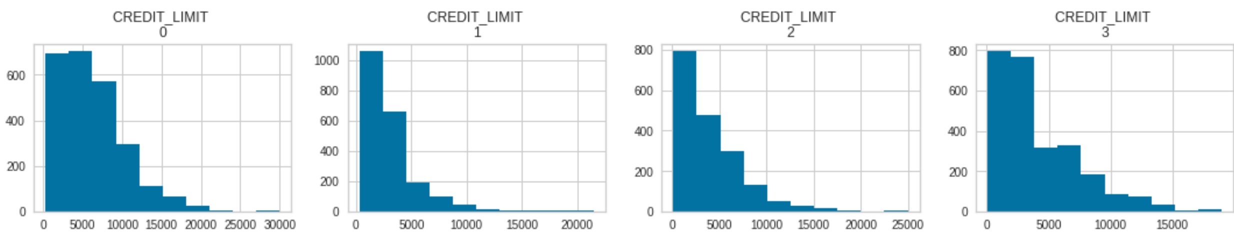

Figure 80: Plot comparison of Credit_Limit of each cluster

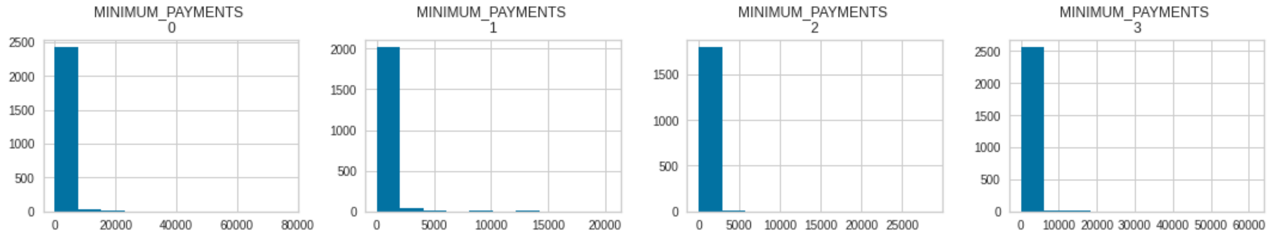

Figure 81: Plot comparison of Minimum_Payments of each cluster

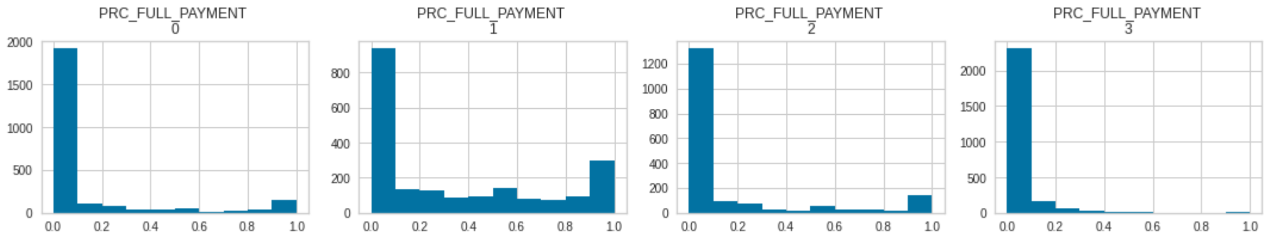

Figure 82: Plot comparison of Pcc_Full_Payment of each cluster

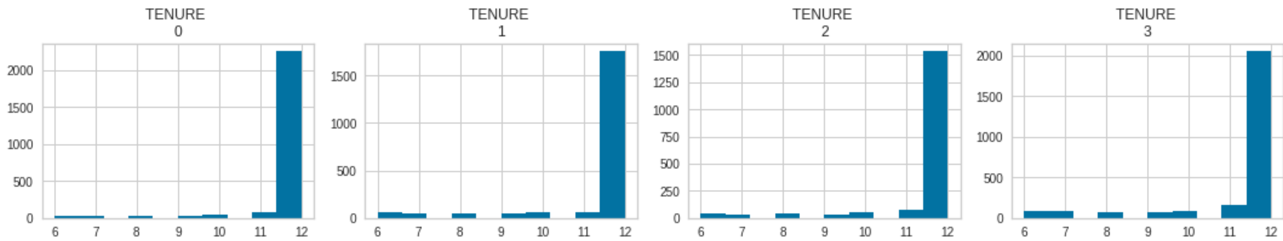

Figure 83: Plot comparison of Tenure of each cluster

The feature distribution of customer data among obtained four clusters are shown in above figure 67 – 83. This cluster distribution results are obtained through performing the spectral clustering to the customer dataset.

Table 7: Mean of each feature for each cluster

cluster	0	1	2	3
BALANCE	314.550920	2286.373265	2555.690207	615.341374
BALANCE_FREQUENCY	0.829610	0.919054	0.982063	0.729298
PURCHASES	577.522806	55.512307	2515.461498	793.471079
ONEOFF_PURCHASES	33.058914	46.788987	1528.161843	744.810337
INSTALLMENTS_PURCHASES	545.326012	8.770169	987.429935	48.909569
CASH_ADVANCE	38.840180	2094.923488	1267.391942	64.546459
PURCHASES_FREQUENCY	0.759972	0.042419	0.830681	0.359611
ONEOFF_PURCHASES_FREQUENCY	0.021750	0.026653	0.453022	0.321499
PURCHASES_INSTALLMENTS_FREQUENCY	0.717049	0.013797	0.663642	0.054475
CASH_ADVANCE_FREQUENCY	0.009538	0.288528	0.166855	0.016180
CASH_ADVANCE_TRX	0.148486	6.807855	4.321834	0.242810
PURCHASES_TRX	12.912542	0.622256	36.200081	7.726217
CREDIT_LIMIT	3004.811813	4171.263463	6291.535280	4224.112309
PAYMENTS	716.375993	1685.447032	3083.214106	1132.031822
MINIMUM_PAYMENTS	438.316632	1064.433736	1339.723237	418.749543
PRC_FULL_PAYMENT	0.333485	0.033179	0.129876	0.152425
TENURE	11.453628	11.338082	11.750000	11.530973

The table 7 describes the mean value for each feature among the four different clusters. The means values were identified since the mean values presents a good indication of the distribution of customer data among four different customer groups. For further analysis the feature importance of customer data attributes is plotted. Figure 84 depicts the feature importance identified through the feature importance analysis.

Feature importance analysis was utilized in the cluster analysis to gain a better insight into customer profiles. The obtained results are used to determine the different customer groups when performing the customer profiling of the different groups. These important features are useful to distinguish the customer segments conveniently.

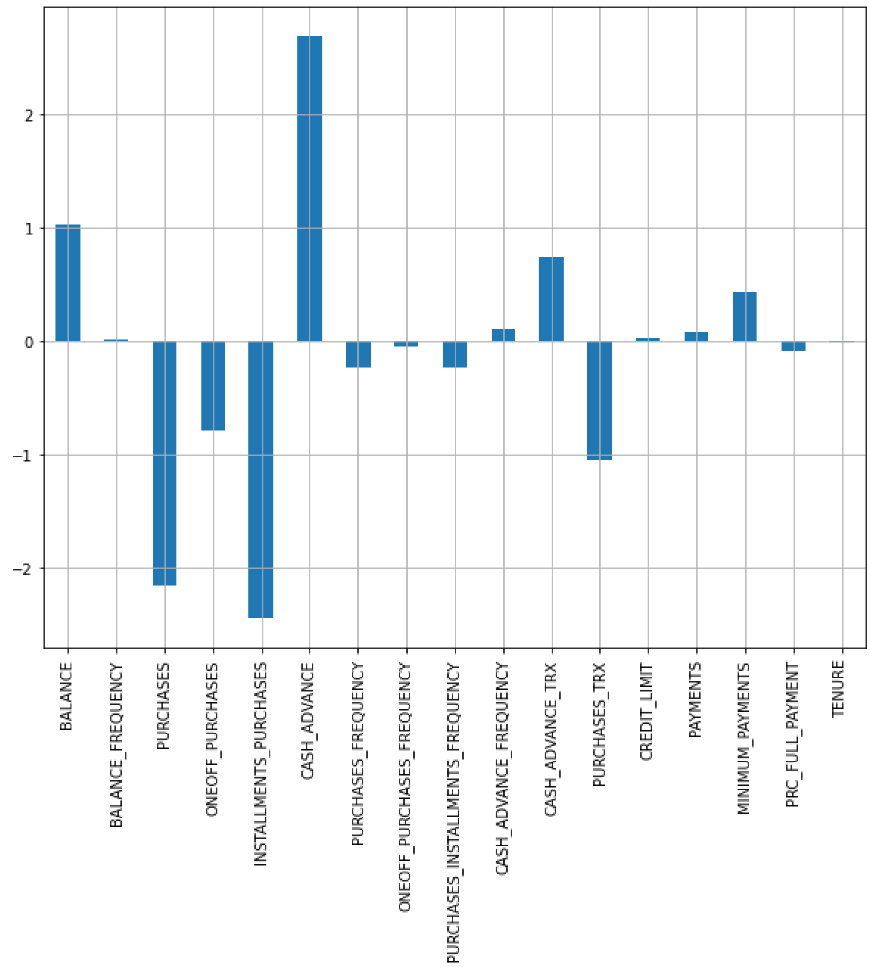

Figure 84: Feature Importance of Customer Data attributes

According to the obtained results of feature importance analysis, the following features were selected as the KPI index of the customer dataset for better analysis of the data distribution.

```
['BALANCE',
 'PURCHASES',
 'ONEOFF_PURCHASES',
 'INSTALLMENTS_PURCHASES',
 'CASH_ADVANCE',
 'CASH_ADVANCE_TRX',
 'PURCHASES_TRX',
 'CREDIT_LIMIT',
 'PAYMENTS',
 'MINIMUM_PAYMENTS']
```

Figure 85: KPI Features

Figure 86 depicts a snake plot of the data of the customer data, the KPI features were plotted for gain an insight of different customer profiles.

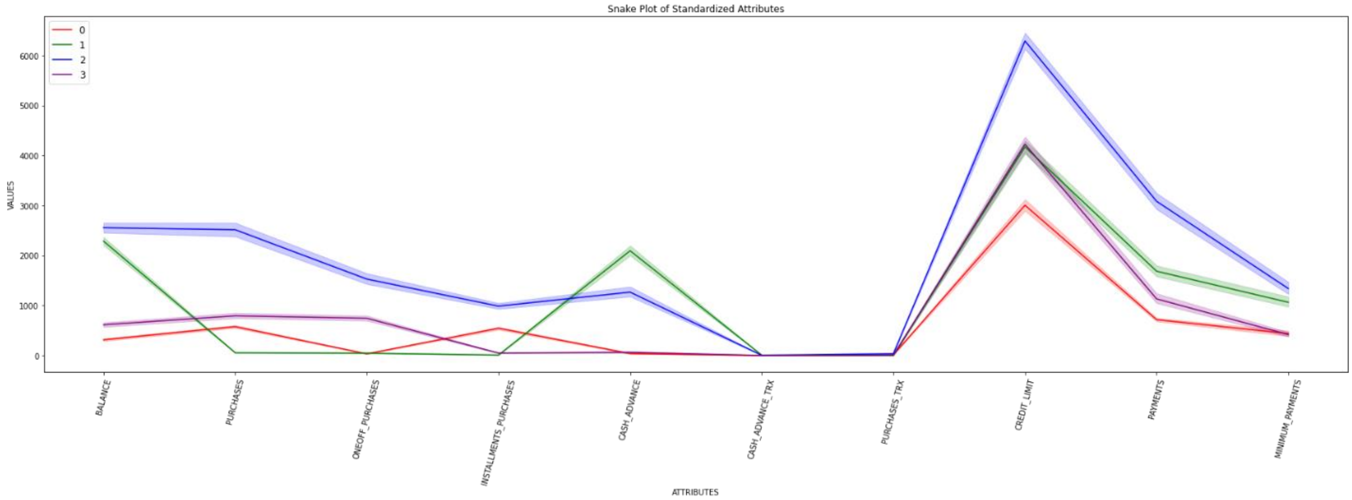

Figure 86: Snake plot of KPI Feature distribution among clusters

As shown in figure 87, a bar graph is plotted to gain a better insight into each customer segment. According to the plotted snake plot cash_advance_trx feature and purchase_trx feature distribution look similar among four clusters hence, these two attributes were disregarded when plotting the bar graph.

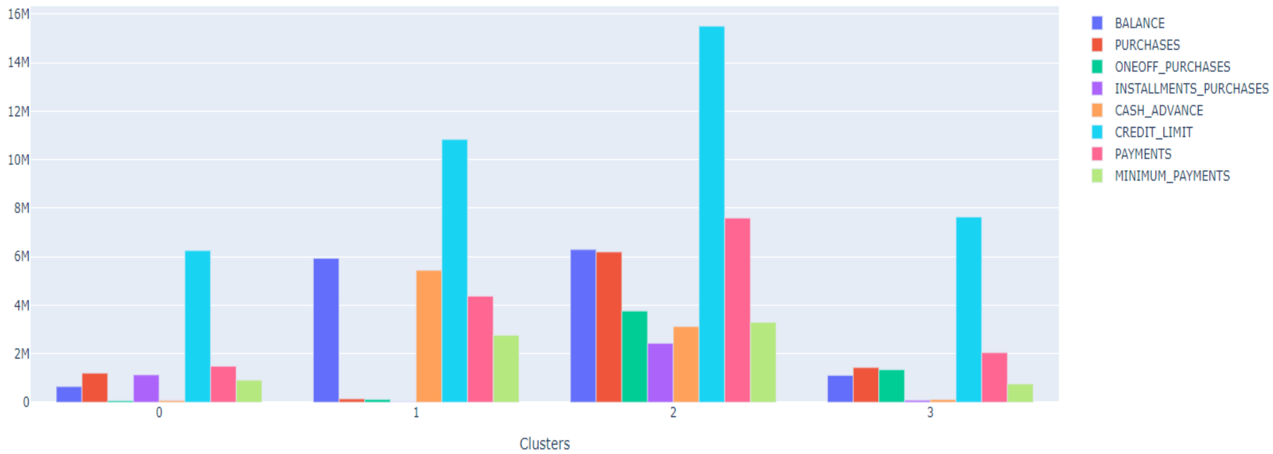

Figure 87: Bar graph of cluster interpretation

Table 8: Descriptive analysis of Cluster 0

index	count	mean	std	min	25%	50%	75%	max
BALANCE	2081.0	314.5509196641038	523.3306991131004	0.0	25.178303	72.518075	300.427627	3543.905366
PURCHASES	2081.0	577.5228063431042	643.5077272871512	12.0	208.6	390.0	694.91	12375.0
ONEOFF_PURCHASES	2081.0	33.0589139836617	127.03213661303917	0.0	0.0	0.0	0.0	2501.0
INSTALLMENTS_PURCHASES	2081.0	545.3260115329169	618.6403832804892	12.0	200.0	366.66	643.41	12375.0
CASH_ADVANCE	2081.0	38.840179851513696	217.54583792518517	0.0	0.0	0.0	0.0	4158.990631
CASH_ADVANCE_TRX	2081.0	0.14848630466122056	0.7354739876236286	0.0	0.0	0.0	0.0	13.0
PURCHASES_TRX	2081.0	12.912542047092744	11.381498199941625	0.0	7.0	11.0	14.0	232.0
CREDIT_LIMIT	2081.0	3004.81181347333	2569.8897718270623	300.0	1200.0	2100.0	4000.0	21500.0
PAYMENTS	2081.0	716.3759932825565	809.4040957904755	0.0	245.689379	475.523262	897.514706	15246.11594
MINIMUM_PAYMENTS	2081.0	438.31663236308316	1228.2295651938364	0.019163	127.210691	166.30613	224.709953	20316.09631

Table 10: Descriptive analysis of Cluster 1

index	count	mean	std	min	25%	50%	75%	max
BALANCE	2597.0	2286.3732648737	2122.3988036703454	1.691842	899.526526	1576.305029	2933.387442	14581.45914
PURCHASES	2597.0	55.51230650750867	173.55635969914798	0.0	0.0	0.0	0.0	3191.0
ONEOFF_PURCHASES	2597.0	46.78898729303042	159.79029954208443	0.0	0.0	0.0	0.0	3191.0
INSTALLMENTS_PURCHASES	2597.0	8.77016942626107	70.75682490595884	0.0	0.0	0.0	0.0	3000.0
CASH_ADVANCE	2597.0	2094.9234881008856	2499.583507619766	0.0	450.089413	1340.127945	2806.959645	26194.04954
CASH_ADVANCE_TRX	2597.0	6.807855217558721	8.479902819673708	0.0	2.0	4.0	9.0	123.0
PURCHASES_TRX	2597.0	0.6222564497497112	1.4286353862059236	0.0	0.0	0.0	0.0	12.0
CREDIT_LIMIT	2597.0	4171.263462503658	3316.506789652552	50.0	1500.0	3000.0	6000.0	19000.0
PAYMENTS	2597.0	1685.44703209742	2673.276252647527	0.0	389.817084	804.570403	1755.16459	34107.07499
MINIMUM_PAYMENTS	2597.0	1064.4337359981268	2570.6099059610556	8.56154	293.203915	542.04143	1053.773309	61031.6186

Table 9: Descriptive analysis of Cluster 2

index	count	mean	std	min	25%	50%	75%	max
BALANCE	2464.0	2555.6902072199678	2556.372061885625	12.423203	746.2621622500001	1693.298307	3531.93873425	19043.13856
PURCHASES	2464.0	2515.461497564935	3436.942494837084	65.82	785.82	1599.5549999999998	2967.615	49039.57
ONEOFF_PURCHASES	2464.0	1528.1618425324677	2746.5932842174834	0.0	255.24	813.01	1748.95	40761.25
INSTALLMENTS_PURCHASES	2464.0	987.429935064935	1419.1387623768105	0.0	234.9075	581.2	1227.6275	22500.0
CASH_ADVANCE	2464.0	1267.3919424431817	2555.296870935949	0.0	0.0	175.1407095	1658.2445605	47137.21176
CASH_ADVANCE_TRX	2464.0	4.321834415584416	7.95163479179824	0.0	0.0	1.0	5.0	123.0
PURCHASES_TRX	2464.0	36.20008116883117	36.63711037003665	2.0	13.0	25.0	45.0	358.0
CREDIT_LIMIT	2464.0	6291.535279745535	4123.278891798015	300.0	3000.0	6000.0	8500.0	30000.0
PAYMENTS	2464.0	3083.2141056554383	4002.968931552677	0.0	1042.4476415	1892.0123800000001	3614.8546895	46930.59824
MINIMUM_PAYMENTS	2464.0	1339.7232365123195	3200.2295997004153	41.854466	251.92691	626.1612565	1336.6660972500001	76406.20752

Table 11: Descriptive analysis of Cluster 3

index	count	mean	std	min	25%	50%	75%	max
BALANCE	1808.0	615.3413741321904	987.0379303203318	0.0	36.5979715	177.13030049999998	890.8079185	12323.84536
PURCHASES	1808.0	793.471078539823	1151.5345583810654	0.0	124.10499999999999	435.94	1012.9475	17945.0
ONEOFF_PURCHASES	1808.0	744.8103373893805	1141.0238608830332	0.0	99.0	392.815	950.0074999999999	17945.0
INSTALLMENTS_PURCHASES	1808.0	48.9095685840708	140.84091202755883	0.0	0.0	0.0	23.0	2272.26
CASH_ADVANCE	1808.0	64.54645880365045	304.9494774833176	0.0	0.0	0.0	0.0	7894.578816
CASH_ADVANCE_TRX	1808.0	0.24280973451327434	0.815730284696234	0.0	0.0	0.0	0.0	16.0
PURCHASES_TRX	1808.0	7.726216814159292	10.536033969204269	0.0	1.0	4.0	11.0	186.0
CREDIT_LIMIT	1808.0	4224.112309463697	3436.1990924135257	150.0	1500.0	3000.0	6000.0	25000.0
PAYMENTS	1808.0	1132.0318224054204	2146.043174080264	0.0	258.67542649999996	590.5946005000001	1293.5550845	50721.48336
MINIMUM_PAYMENTS	1808.0	418.7495434344156	968.0235964164357	0.05588	123.724716	192.0155795	465.89675724999995	28483.25483

Table 8-11 depicted above describes the statistical analysis of the four different clusters. This information were utilized identify the descriptive statistical analysis of the data distribution between the obtained four different clusters. As shown in above tables, the mean value, standard deviation, minimum value, first quartile, median, third quartile, and finally the maximum value of the customer records of the selected features are reported. The pair plot shown in figure 87 depicts the pairwise relationships among the selected KPI features of the customer dataset. The different colors represent the four different clusters of the customer data. The cluster label value was used as the hue parameter to depict the visualization of the distribution of features among the clusters.

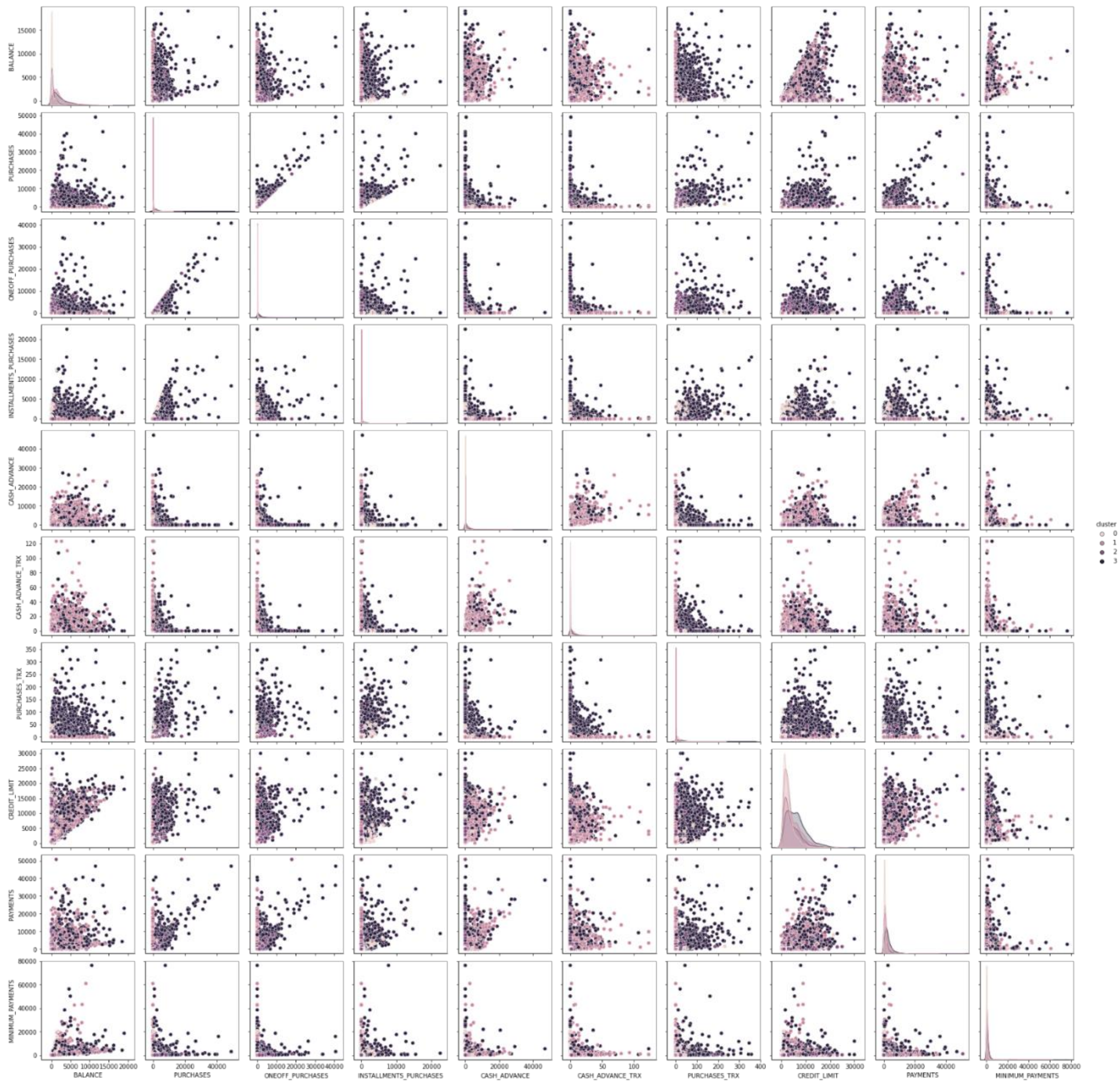

Figure 88: pair plot of KPI features of the customer data among four clusters

According to the results obtained through cluster analysis, cluster evaluation, cluster interpretation and descriptive statistical analysis the four different customer groups are identified from the targeted customer dataset and the unique features of these four customer segments are identified.

Cluster 0:

This customer group reported the lowest balance amount, lowest payments amount, lowest cash advance rate, lowest credit limit and purchases are also relatively low. Hence this customer group is treated as New Customers.

Cluster 1:

This customer group reported lowest purchases where the installment purchases and one-off purchases also reported as lowest along other customer groups. But this customer group has a higher credit-limit and scores the highest number of minimum payments and cash advance amount. This customer group rarely spends their money hence, this customer group is treated as Money Hoarders.

Cluster 2:

This customer group reported the highest credit limit, highest purchases, highest balance, highest payments, and highest one-off purchases. This customer group indicated as valuable customer group with higher amount of money and purchases. Hence these customers identified as the most valuable customer group and treated as Prime Customers.

Cluster 3:

This customer group reported the second lowest balance amount, second lowest payments, second highest purchases, second highest one-off purchases, higher credit limit. This group does not report any highest or lowest feature. Hence, this customer group is treated as Average Customers.

4.1.5 Prediction Model Evaluation

The resultant customer segmentation obtained through performing clustering model were utilized to implement the prediction model to predict the future customer into identified segments. Supervised models were modeled on the resultant clustered customer dataset to evaluate the prediction models.

Before training the classification models SMOTE(Synthetic Minority Oversampling Technique) is applied on the dataset to handle the class imbalance of the four different clusters. This technique is applied since the imbalance classes effects to lessen the accuracy of the classification model. Figure 89 depicts the result of the clustered dataset after applying the SMOTE technique.

```
Before SMOTE:
{0: 2081, 1: 2597, 2: 2464, 3: 1808}
================================================
After SMOTE:
{0: 2597, 1: 2597, 2: 2597, 3: 2597}
```

Figure 89: SMOTE on clustered dataset

The obtained dataset is split into train set and test set before training the classification models. Most of the previous researchers have split the dataset where the 80% of dataset belong to the training dataset and the 20% of the dataset was treated as the test dataset. Hence the commonly used ratio of 80:20 split was considered for the clustered customer dataset before performing the classification modelling.

4.1.4.1 Multinomial Logistic Regression

Before training the logistic regression model , VIF (Variance Inflation Factor) analysis is performed on the SMOTE dataset obtained to identify and remove the high variance features from the dataset. The logistic regression model supposes the gestures do not have strong multicollinearity and are independent variables.

	feature	VIF
2	PURCHASES	63109.104754
3	ONEOFF_PURCHASES	35674.287546
4	INSTALLMENTS_PURCHASES	10986.915648
6	PURCHASES_FREQUENCY	27.610552
8	PURCHASES_INSTALLMENTS_FREQUENCY	17.741031
1	BALANCE_FREQUENCY	17.332062
16	TENURE	15.597801
7	ONEOFF_PURCHASES_FREQUENCY	6.121900
9	CASH_ADVANCE_FREQUENCY	4.812560
12	CREDIT_LIMIT	4.447391
11	PURCHASES_TRX	4.325595
0	BALANCE	3.959362
10	CASH_ADVANCE_TRX	3.882798
13	PAYMENTS	3.632108
5	CASH_ADVANCE	3.425188
15	PRC_FULL_PAYMENT	1.733024
14	MINIMUM_PAYMENTS	1.409559

Figure 90: VIF of features

The figure 90 described the VIF of each feature in the dataset and the features with high VIF value is removed from the dataset before applying the logistic regression model. Class labels are the four clusters obtained through the cluster analysis.

```
Accuracy: 0.8888354186717998
Multinomial Logistic Regression
              precision    recall  f1-score   support

           0       0.88      0.93      0.90       520
           1       0.93      0.96      0.94       520
           2       0.88      0.84      0.86       519
           3       0.87      0.82      0.85       519

    accuracy                           0.89      2078
   macro avg       0.89      0.89      0.89      2078
weighted avg       0.89      0.89      0.89      2078
```

Figure 91: Classification report - Multinomial Logistic Regression

The figure 91 shows the classification report obtained for Multinomial Logistic Regression model. Here an accuracy of 0.89 is reported with other classification metrices.

4.1.4.2 Decision Tree Classifier

The figure 92 shows the classification report obtained for Decision tree classifier for the customer dataset. Class labels represent the four different clusters obtained through clustering analysis. Here an accuracy of 0.92 is reported with precision, recall, f1-score, and support for each class. We can identify that the Decision Tree classifier accuracy is higher than the Logistic Regression algorithm.

```
Accuracy: 0.9278152069297402
Decision Tree Classifier
              precision    recall  f1-score   support

           0       0.92      0.95      0.94       520
           1       0.95      0.95      0.95       520
           2       0.91      0.88      0.90       519
           3       0.92      0.93      0.93       519

    accuracy                           0.93      2078
   macro avg       0.93      0.93      0.93      2078
weighted avg       0.93      0.93      0.93      2078
```

Figure 92: Classification report - Decision Tree Classifier

4.1.4.3 Random Forest Classifier

The Random Forest classifier is applied to the balanced dataset and the figure 93 depicts the classification report obtained for the RF classifier. Here, an accuracy of 0.96 is reported with the precision, recall, f1-score, and support metrices for each class label.

```
Accuracy: 0.9615014436958614
Random Forest Classifier
              precision    recall  f1-score   support

           0       0.95      0.97      0.96       520
           1       0.99      0.97      0.98       520
           2       0.94      0.95      0.95       519
           3       0.96      0.96      0.96       519

    accuracy                           0.96      2078
   macro avg       0.96      0.96      0.96      2078
weighted avg       0.96      0.96      0.96      2078
```

Figure 93: Classification Report - Random Forest Classifier

Table 12: Classification Report Evaluation Summary

		Multinomial Logistic Regression	Decision Tree Classifier	Ensemble Random Forest Model
accuracy		0.8888	0.9278	0.9615
precision	0	0.88	0.92	0.95
	1	0.93	0.95	0.99
	2	0.88	0.91	0.94
	3	0.87	0.92	0.96
recall	0	0.93	0.95	0.97
	1	0.96	0.95	0.97
	2	0.84	0.88	0.95
	3	0.82	0.93	0.96
F1-score	0	0.90	0.94	0.96
	1	0.94	0.95	0.98
	2	0.86	0.90	0.95
	3	0.85	0.93	0.96

Table 12 depicts the summary of the evaluation metrices of the three supervised learning models. According to the results obtained for classification modelling the RFM ensemble learning model reports the highest accuracy score. Hence, the Random Forest Model is selected for customer segmentation prediction modelling to predict the customer segment of the future customers.

The hyperparameter optimization is performed on the algorithm before building the random forest model to obtain a highest accuracy for the prediction model. Through hyperparameter tuning the optimal maximum depth parameter and optimal number of estimators are determined.

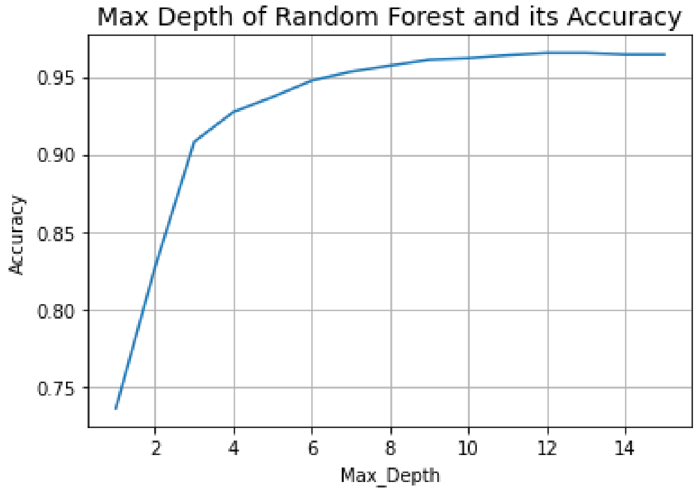

Figure 94: RF maximum depth and accuracy

According to the figure 94, the optimal maximum depth is the point in the graph where the RF model accuracy tends to stop improving. Here, it is identified that the accuracy of the model stops improving at the maximum depth of 12. Hence, the 12 is selected as the optimal depth for the RF model.

The figure 95 depicts the no of estimator against the rf model accuracy in order to identify the optimal number of estimators for the random forest model.

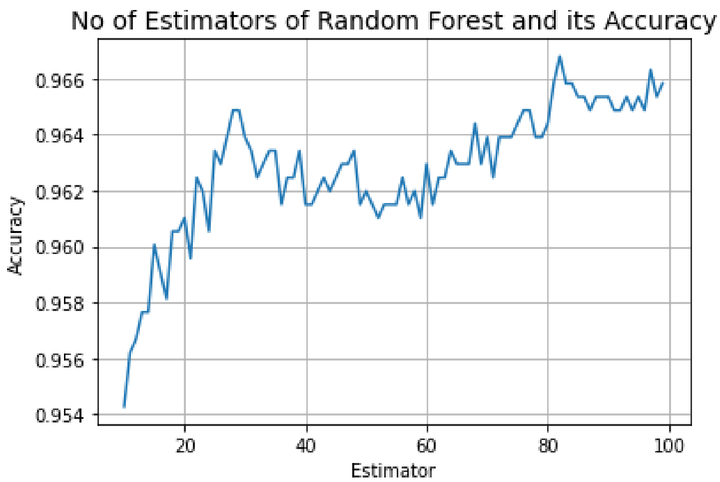

Figure 95: RF no. of estimators and accuracy

According to the above results it is identified that the highest accuracy point is reported at the point where the number of estimators is equal to 82. Hence, the optimal number of estimators is selected as 82 for optimize the prediction model accuracy.

As shown in figure 96, after optimizing the RF model with hyperparameter tuning the model records an accuracy of 0.97.

```
Random Forest Model with the optimal max depth and optimal no of estimators
Accuracy: 0.9667949951876804
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       520
           1       0.99      0.97      0.98       520
           2       0.95      0.97      0.96       519
           3       0.96      0.96      0.96       519

    accuracy                           0.97      2078
   macro avg       0.97      0.97      0.97      2078
weighted avg       0.97      0.97      0.97      2078
```

Figure 96: Accuracy score after RF model optimization

The figure 97 depicts the feature importance of the trained RF model. Feature importance in the random forest classifier is plotted to recognize the most significant features in the prediction model.

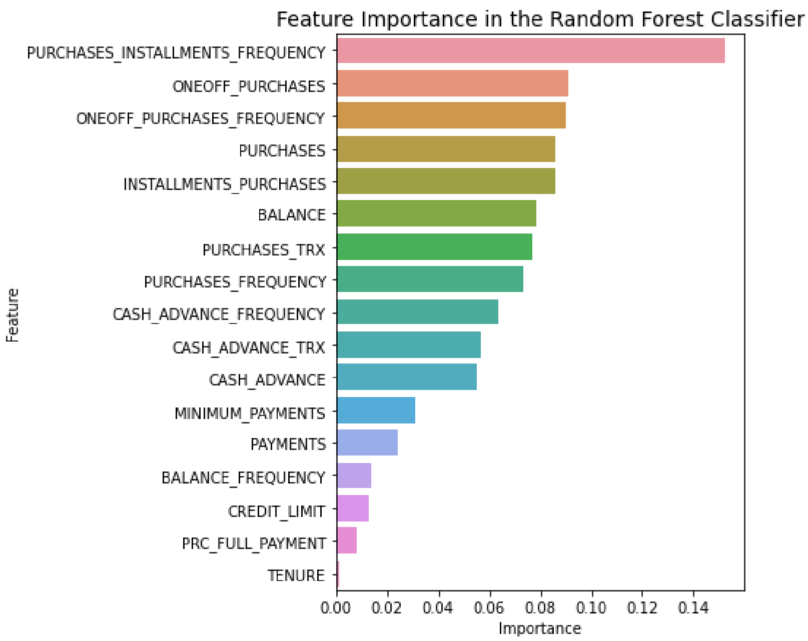

Figure 97: Feature Importance of RF Model

4.2 Customer Segmentation Prediction System

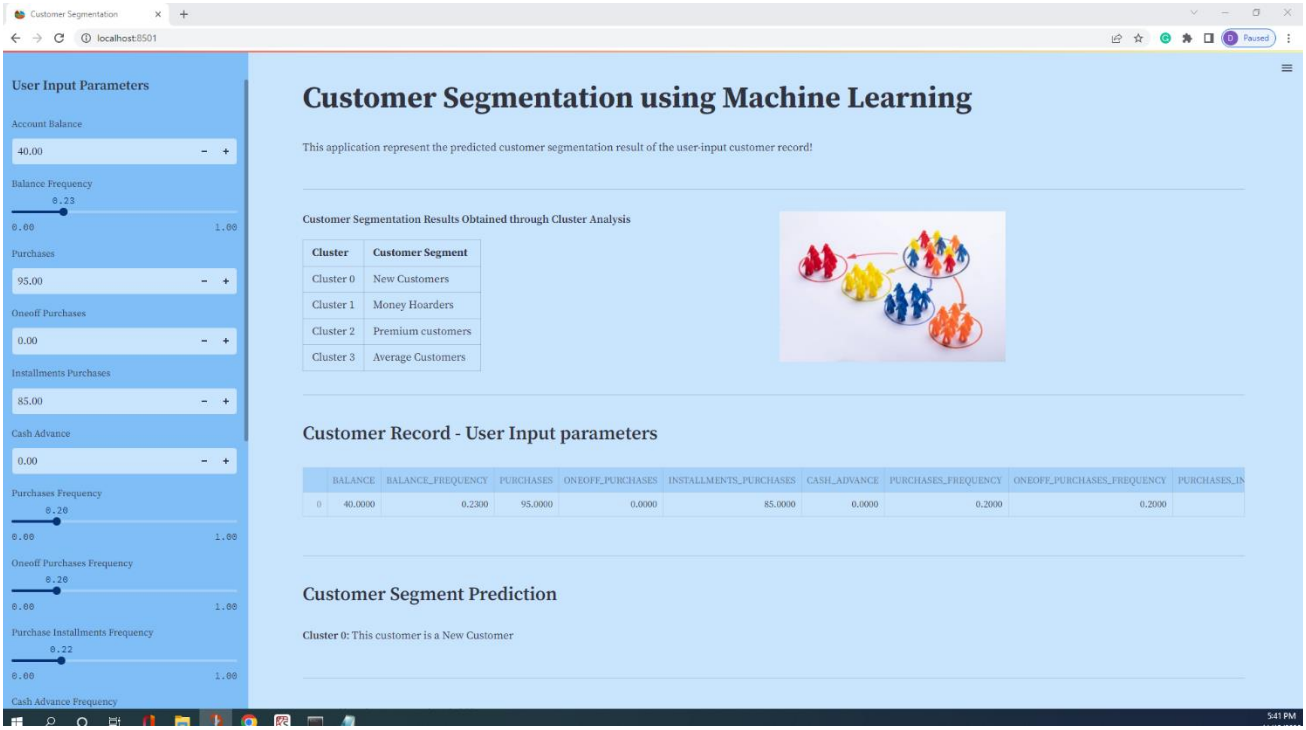

Figure 98: Customer Segmentation Prediction Web Application

The output of the final result is a web application, and this can be accessed through a browser. The figure 98 represents the implemented customer segmentation system. The future customers can be predicted through this prototype web application. The obtained four different clusters is the base of the prediction and treated as the target class labels to segment the customer to the relevant group according to the built ensemble prediction model. For the implementation of the prediction model a hybrid learning approach was utilized.

The machine learning model was implemented using python and the Streamlit python framework is used to create the web application after deployment of the model.

The end user can input the values for the parameters using the user input parameter side bar as shown in figure 99. Here, all the attributes of the customer record can be entered to obtain the customer segment of the input customer record.

Figure 99: User input parameters

Customer Segmentation Results Obtained through Cluster Analysis

Cluster	Customer Segment
Cluster 0	New Customers
Cluster 1	Money Hoarders
Cluster 2	Premium customers
Cluster 3	Average Customers

Figure 100: Customer segment labels

The customer segment labels as shown in figure 100 are the targeted classes for the prediction model and these segments are the resultant clusters obtained through training the clustering models for the customer dataset. The clustered dataset is the base for this model where the classification is performed on the clustered dataset targeting the above-mentioned cluster labels and the predictions are made according to the obtained estimator rules.

Customer Record - User Input parameters

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_IN
0	20,000.0000	0.9100	10,000.0000	5,000.0000	85.0000	0.0000	1.0000	1.0000	

Figure 101: User input customer record

And the user input parameter is shown in the figure 101 and each parameter value is treated as a data frame for the prediction model and the prediction will be based on this input data of the customer record.

Finally, the prediction result will be shown under the customer segment prediction system and after analysis of the input customer record the resultant predicted customer segment is shown to the user as shown in figure 102.

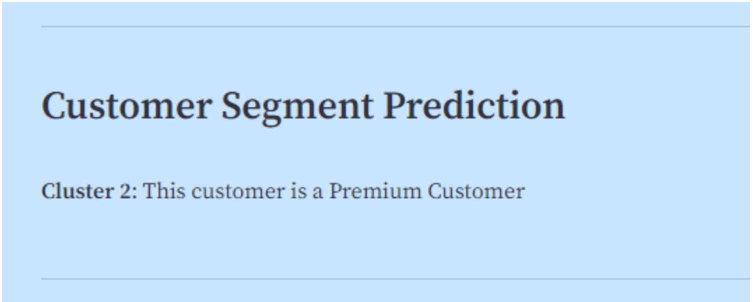

Figure 102: Customer segment prediction

The prediction probability of the predicted customer segment is shown under the prediction probability section as shown in the figure 103.

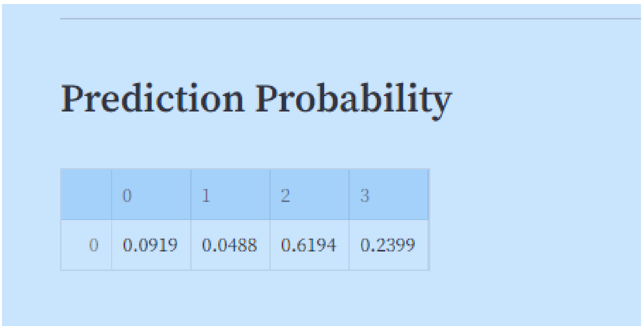

Figure 103: Prediction probability

This chapter described the evaluation and results of the research study where this chapter comprised the information about the methodologies applied to obtain the research objectives, presentation of cluster analysis and results, evaluation of the obtained results and further the designs of the system prototype. The next chapter is the final chapter of the thesis which documents the conclusion consisting of the limitations faced during this research study and further the future works of the research study.

CHAPTER 5: CONCLUSION AND FUTURE WORK

This chapter contains an overview of the research, a summarization of the findings of the entire research work, the limitations of the research study, and an outline of possible further improvements.

Customer segmentation is a significant method used for segmenting the customer base based on the similarities that they share with the aspect of any dimension. In the eye of the business world, the key objectives of customer segmentation are focusing on strategies for new products and services development, deciding the most appropriate market communication for the relevant customers, constructing the most appropriate customer servicing and also customer retention strategies, and increasing company profits and their customer retention rate.

This research study was focused on identifying the customer segmentation of the customer base by analyzing the customer purchasing pattern and gaining insight into the different customer profiles. The data science process applied to this research study is underlined with the CRISM-DM process. The customer dataset is preprocessed for a better result in clustering and prediction modeling. A hybrid approach was followed in this research study where the supervised learning model and unsupervised learning model were utilized for implementing the customer segmentation prediction system. The unsupervised machine learning was performed on the customer dataset to obtain the customer segmentation results whereas the outcome of the unsupervised learning model was the labeled dataset. The label of each customer record of the dataset identifies the cluster label to which that particular customer belongs in other words, the relevant customer group of the particular customer. Afterward, for the prediction model building supervised machine learning was performed on the clustered data set.

Clustering data is an intricated method where it involves the selection between several clustering methods and choosing the parameters and performance metrics to evaluate the cluster validity. To obtain the customer segmentation, clustering algorithms were performed on the customer dataset. For this research study six different clustering algorithms were selected and trained on the customer dataset. The selected algorithms are the k-means clustering, agglomerative clustering, spectral clustering, gaussian mixture model-based clustering, DBSCAN clustering and BIRCH clustering.

According to the obtained outcome of cluster analysis the cluster evaluation was performed and the internal cluster evaluation metrices were used to evaluate the clusters using the cohesion and separation measures. Internal measures evaluate the clustering results based on the cohesion and separation of the clusters. External measures could not be used for this study since the customer dataset does not include any ground-truth labels. According to the cluster result evaluation the spectral clustering model provided the best clustering performance with the optimal cohesion, separation, and customer distribution among identified clusters.

The highest silhouette score which is 0.3055 and the lowest Davies-Bouldin index which is 1.3062 and a higher Calinski and Harabasz score which is 3690.4564 reported on the spectral clustering evaluation and further according to the data distribution among the clusters this model resulted in a better distribution of the customer data among the identified four different clusters.

Therefore, the cluster results obtained through training the spectral model was utilized for the prediction modelling and the three different classification algorithms were trained on the clustered dataset to evaluate the model accuracy. The Random Forest ensemble learning model was reported the best accuracy of 0.97 hence, the RF model was used to build the customer segmentation prediction model.

5.1 Limitations in the Research Study

However, there are some limitations identified in this research study. The proposed customer segmentation algorithms require diverse and various customer data in order for better validation. Here, we had a customer dataset of only one New York City bank and the dataset has data collected only for a period of six months about the customers and their purchasing history. And there were many missing values reported on the dataset and those had to be filled in before applying the machine learning models.

In the machine learning application, only six unsupervised learning clustering algorithms and three different supervised learning algorithms were modeled and analyzed on the customer data. But there might be better clustering methodologies and prediction methodologies that exist for this customer segmentation prediction study.

5.2 Future Work

Considering the formerly mentioned limitations of the research study for further work of the research to improve the validity of the customer segmentation different datasets can be integrated to perform the segmentation work. Moreover, adding more features to the dataset, or appropriate further improvements to the available features might help to increase the overall performance of the machine learning model.

Further comprehensive knowledge about the customer base might help to simplify and also to improve the labeling process of customer segmentation. Furthermore, other suitable clustering models and multiclass classification models can be analyzed for improving the accuracy of the segmentation and predictive performance.

APPENDICES

Appendix A: Customer dataset
https://drive.google.com/drive/folders/1NPop3t13FCx0hwkd_7Bu-NOrqHGaXfgr?usp=share_link

Appendix B: clustered dataset with labels obtained after cluster analysis
https://drive.google.com/drive/folders/1O7pLXpg8cj61m3XAw_YpEfKM-4Ptv23i?usp=share_link

Appendix C: URL for source code
https://drive.google.com/drive/folders/1Yah4OpHawWGggqIr2WW7a9e-BYM3Ssjm?usp=share_link

REFERENCES:

Anthony D. Fontanini and Joana Abreu, 2018. A Data-Driven BIRCH Clustering Method for Extracting Typical Load Profiles for Big Data.

Arora P, Deepali, Varshney S, 2016. Analysis of K-means and K-medoids algorithm for big data. Procedia Comput Sci 78:507–512.

Babak Sohrabi, Amir Khanlari, 2007. Customer Lifetime Value (CLV) Measurement Based on RFM Model, Iranian Accounting & Auditing Review, Spring 2007,Vol. 14 No. 47, pp 7- 20.

Bhade K., Gulalkari V., Harwani N., Sudhir N. Dhage, 2018. A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization. 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE.

Blanchard T., Behera D., & Bhatnagar P., 2019. Data Science for Marketing Analytics: Achieve Your Marketing Goals with the Data Analytics Power of Python.

Chan H.C., Chang C.H., Chen P.A., Lee J.T., 2019. Using multinomial logistic regression for prediction of soil depth in an area of complex topography in Taiwan

Cheng Li, 2008. Research on Segmentation implementation process of air cargo Customer based on Data Mining.

Dolnicar S., Grün B. & Leisch F., 2018. Market Segmentation Analysis, Management for Professionals.

Goller S., Hogg A., & Kalafatis S. P., 2002. A new research agenda for business segmentation. European Journal of Marketing, 36(1/2), 252–271.

Halkidi M. and M. Vazirgiannis, 2001. Clustering validity assessment: finding optimal portioning of a dataset.In proceeding of ICDM Conference, California, USA.

Hammouda K. M., 2001. Web Mining: Clustering Web Documents A Preliminary Review. Available at: http://watnow.uwaterloo.ca/pub/hammouda/review-document-clustering.pdf. (Accessed: 18 May 2022)

Hiziroglu, A., 2013. Soft computing applications in customer segmentation: State-of-art review and critique’, Expert Systems with Applications, 40(16), pp. 6491–6507.

Hyunseok Hwang, Taesoo Jung, Euiho Suh, 2004. An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry Expert systems with applications, Vol. 26, pp. 181-188.

Ina Maryani, D. Riana, Rachmawati Darma Astuti, Ahmad Ishaq, Sutrisno, Eva Argarini Pratama, 2018. Customer Segmentation based on RFM model and Clustering Techniques with K-Means Algorithm. STMIK Nusa Mandiri Jakarta.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2017. An introduction to statistical learning: with applications in R. Springer: New York.

Jan Panuš, Hana Jonášová, Kateřina Kantorová, Martina Doležalová, Kateřina Horáčková, 2016. Customer segmentation utilization for differentiated approach. The International Conference on Information and Digital Technologies.

Kansal Tushar, Bahuguna Suraj, Vishal Singh, Tanupriya Choudhury, 2018. Customer Segmentation using K-means Clustering. International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), IEEE.

Kelleher, J.D., Namee, B.M., Arcy A.D., 2015. Fundamentals of Machine Learning for Predictive Data Analytics Algorithms, Worked Examples, and Case Studies. The MIT Press, Cambridge, Massachusetts, USA.

K.A. Sawicz, C. Kelleher, T. Wagener, P. Troch, M. Sivapalan, G. Carrillo, 2014. Characterizing hydrologic change through catchment classification Hydrol. Earth Syst. Sci., 18, pp. 273-285, 10.5194/hess-18-273-2014

Kotler, P., and Armstrong G., 2010. Principles of marketing. Pearson education.

L. Chang and X. Bai, 2010. Data Mining: A Clustering Application.

Lindholm A., Wahlström N., Lindsten F., Schön T., 2021. Machine Learning - A First Course for Engineers and Scientists.

Marcus, C., 1998. A practical yet meaningful approach to customer segmentation approach to customer segmentation, Journal of Consumer Marketing 15, 494-504.

M. Inaba, N. Katoh and H. Imai, 1994. Applications of weighted Voronoi diagrams and randomization to variance-based kclustering, in Proceedings of the tenth annual symposium on Computational geometry, New York, pp. 332-339.

Pavel Berkhin, 2012. A Survey of Clustering Data Mining Techniques.

Prabha Dhandayudam, Dr. Ilango Krishnamurthi, 2012. ‘An Improved Clustering Algorithm for Customer Segmentation’. International Journal of Engineering Science and Technology (IJEST) vol. 4 no.02. (Accessed: 15 February 2022)

Pranay Modukuru. ‘Customer Segmentation and Acquisi tion using Machine Learning’, Towards Data Science[online]. Available at: https://towardsdatascience.com/customer-segmentation-and-acquisition-using-machine-learning-a219ce0ec139 (Accessed: 20 May 2022)

R. Xu, S. Member, and D.W. Ii, 2005. Survey of Clustering Algorithms.IEEE Transactions on neural networks, vol.16.

Sabbir Hossain Shihab, Shyla Afroge, Sadia Zaman Mishu, 2019. RFM Based Market Segmentation Approach Using Advanced K-means and Agglomerative Clustering:A Comparative Study. International Conference on Electrical, Computer and Communication Engineering (ECCE).

Salvador, Stan, and Philip Chan, 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. 16th IEEE international conference on tools with artificial intelligence. IEEE.

Sami Madani, 2009. Mining changes in customer purchasing behavior : a data mining approach, Luleå University of Technology, Department of Business Administration and Social Sciences, Division of Industrial marketing and e-commerce

Shaikh R., 2018. Feature Selection Techniques in Machine Learning with Python [WWW Document]. Medium. Available at: https://towardsdatascience.com/feature-selection-techniques- in-machine-learning-with-python-f24e7da3f36e (Accessed: 15 June 2022).

Su-Yeon Kim, Tae-Soo Jung, Eui-Ho Suh and Hyun-Seok Hwang (2006), Customer segmentation and strategy development based on customer lifetime value: A case study, Elsevier Conference on Expert Systems with Applications Volume 31, pp. 101–107.

Sunitha Cheriyan, 2019. Intelligent Sales Prediction Using Machine Learning Technique. IT Department Higher College of Technology, ResearchGate.

Tan P., Steinbach M., Karpatne A. & Kumar V., 2014. Introduction to data mining. 1st Ed, Pearson: Harlow.

Tan Steinbach Kumar, 2005. Introduction to Data Mining. Addison Wesley press.

Tripathi, S., A. Bhardwaj, and E. Poovammal, 2018. Approaches to clustering in customer segmentation. International Journal of Engineering & Technology 7.3.12: 802-807.

Vasilis Aggelis, 2005. Customer Clustering using RFM analysis.

You-Shyang Chen, Ching-Hsue Cheng, Chien-Jung Lai, Cheng-Yi Hsu, Han-Jhou Syu, 2012. Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospitalbased assessment, Computers in Biology and Medicine, vol. 42, no. 2, pp. 213-221.

Xu, Huajie, and Guohui Li.,2008. Density-based probabilistic clustering of uncertain data. International Conference on Computer Science and Software Engineering. Vol. 4. IEEE.