Pattern Recognition and Customer Segmentation Using Feature Analysis on Postpaid Fixed Line Telecom Data

SIDRA UROOJ

01-241212-012

A thesis submitted in fulfillment of the
requirements for the award of the degree of
Master of Science (Software Engineering)

Department of Software Engineering

BAHRIA UNIVERSITY ISLAMABAD

MARCH 2024

APPROVAL FOR EXAMINATION

Scholar's Name: Sidra Urooj Registration No. 51783

Program of Study: MS (Software Engineering)

Thesis Title: Pattern Recognition and Customer Segmentation Using Feature Analysis on Postpaid Fixed Line Telecom Data

It is to certify that the above scholar's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for examination. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index 11% that is within the permissible limit set by the HEC for the MS degree thesis. I have also found the thesis in a format recognized by the BU for the MS thesis.

Principal Supervisor's

Signature:_____

Date: _____

Name:_____

AUTHOR'S DECLARATION

I, Sidra Urooj hereby state that my MS thesis titled "Pattern Recognition and Customer Segmentation Using Feature Analysis on Postpaid Fixed Line Telecom Data" is my own work and has not been submitted previously by me for taking any degree from this university Bahria University Islamabad or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my MS degree.

Name of scholar: Sidra Urooj (01-241212-012)

Date: _____

PLAGIARISM UNDERTAKING

I, Sidra Urooj, solemnly declare that the research work presented in the thesis titled "Pattern Recognition and Customer Segmentation Using Feature Analysis on Postpaid Fixed Line Telecom Data" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Bahria University towards plagiarism. Therefore, I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the university reserves the right to withdraw/revoke my MS degree and that HEC and the University has the right to publish my name on the HEC/University website on which names of scholars are placed who submitted plagiarized thesis.

Scholar / Author's Sign: _____

Name of the Scholar: Sidra Urooj (01-241212-012)

DEDICATION

This thesis is dedicated to my loving parents, who have always been an inspiration to me, who have never failed to show their love, support, and devotion, and who have always supported me morally, emotionally, mentally, and spiritually.

ACKNOWLEDGEMENT

First and foremost, I would like to thank Allah Almighty for the numerous blessings that have inspired me to take on the task of working on this thesis and completing it successfully.

I wish to express my sincere appreciation to my thesis supervisor, Dr. Tamim Ahmed Khan, for his support, guidance, and valuable feedback throughout this thesis. His expertise and encouragement have been instrumental in shaping my research and helping me to overcome the challenges that I faced.

I want to acknowledge my parents, who supported me throughout. This thesis work is dedicated to my parents, who have been a constant source of support during the challenges of life and prayed for my success. I am truly thankful for the support in every aspect whether that is financial, emotional, or mental.

Lastly, I would like to acknowledge the contributions of all the participants who took part in my study, without whom this research would not have been possible.

ABSTRACT

Businesses today need to know their customers to maintain the competitive margin and run effectively. Customer Segmentation helps companies to understand their customer base such as their behaviors, demographics, and psychographics etc. In addition, the Multi-Offer Recommendation model predicts purchasing patterns for non-bundled customers, if these customers will subscribe to one or more bundles. To understand customer segments, comparative analysis was conducted between DBSCAN and K-means clustering. While DBSCAN excelled in identifying noise points, its clustering performance was ineffective as it failed to capture discrete customer segments. On the contrary, K-means demonstrated satisfactory performance by clearly identifying heterogeneous clusters, facilitating insightful post-cluster analysis which helps understand different type of customers. Manual campaigns were conducted post cluster analysis and found to be effective overall in comparison with current manual campaigns. For the Multi-Offer Recommendation model, Random Forest Classifier and Random Forest Regressor with Multi-Output Classifier and Multi-Output Regressor were used. Post development marketing campaigns found that regressor based campaigns were more successful than classifier-based campaigns in terms of response rate, indicating their suitability for this dataset.

TABLE OF CONTENTS

CHAPTER TITLE PAGE

APPROVAL FOR EXAMINATION ii

AUTHOR'S DECLARATION iii

PLAGIARISM UNDERTAKING iv

DEDICATION v

ACKNOWLEDGEMENT vi

ABSTRACT vii

TABLE OF CONTENTS viii

LIST OF TABLES xi

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xiv

CHAPTER 1 1

INTRODUCTION 1

1.1. Motivation 2
1.2. Research Gap 3
1.3. Problem Statement 5
1.4. Research Questions 6
1.5. Research Objectives 6
1.6. Contribution of the study 7
1.7. Outline of this thesis 7

CHAPTER 2 8

LITERATURE REVIEW 8

CHAPTER 3 16

RESEARCH METHODOLOGY 16

3.1. Introduction 16
3.2. Research Methodology 16
3.3. Conceptual Framework 17
3.4. Data Collection 18
3.4.1. Dataset Features 18
3.4.2. Exploratory Data Analysis (EDA) 22

CHAPTER 4 31

RESULTS AND EVALUATION 31

4.1. DBSCAN 31
4.2. K-means Clustering 33
4.2.1. K-means on Two Principal Components 34
4.2.2. K-means on 60 principal components 35
4.3. Post Clustering Analysis 37
4.3.1. Average Profile 38
4.3.2. Monthly Usage Profile 39
4.3.3. Monthly Billing Profile 40
4.3.4. Usage Slabs Analysis 41
4.3.5. Bill Slabs Analysis 42
4.4. Cluster Profile Summary 44
4.5. Manual Campaign Results 45
4.6. Classification and Regression Model 48
4.6.1. Random Forest Classification 49
4.6.2. Random Forest Regression 52
4.6.3. Prominent Features 53
4.7. Multi-Offer Recommendation Model 54

CHAPTER 5 58

CONCLUSION 58

5.1. Conclusion 58

5.2. Future work 58

REFERENCES 60

LIST OF TABLES

TABLE NO. **TITLE** **PAGE**

Table 1 - Reviewed Research Work 12
Table 2 - Customer Profile features 19
Table 3 - Call Usage Features 20
Table 4 - Billing Features 21
Table 5 - Customer Complaints 22
Table 6 - Cluster Profile Summary 44
Table 7 - Cluster Profile and Results 46
Table 8 - Target Sample for each Cluster 47
Table 9 - Results of each Cluster 47
Table 10 - Campaign wise results 48
Table 11 - Classification Model Evaluation 51
Table 12 - Multi-Label Evaluation 51
Table 13 - MAE vs MSE 53
Table 14 - Regression Results 53
Table 15 - Multi-Offer Recommendation Model Prediction Results 55
Table 16 - Multi-Offer Recommendation Model - Campaign Results 55

LIST OF FIGURES

FIGURE NO. **TITLE** **PAGE**

Figure 1- Research Methodology 17
Figure 2- Conceptual Framework 18
Figure 3- Dataset Sample 22
Figure 4- Statistical Summary 23
Figure 5- Distribution of Customer Type 23
Figure 6- Bill Distribution by Customer Type 24
Figure 7- Distribution of Subscription Type 25
Figure 8 - Bill Distribution by Subscription Type 25
Figure 9- Density Plots for Billing Features 26
Figure 10- Correlation Analysis between Usage and Bill Features 27
Figure 11- One Hot Encoding of Categorical Variables 28
Figure 12- Two Components PCA and Elbow Method 29
Figure 13- Variance Ratio 30
Figure 14- Elbow Method 50 and 60 Components 30
Figure 15- DBSCAN - 2 Clusters 32
Figure 16- DBSCAN - 3 Clusters 33
Figure 17- 4 Clusters K-means 34
Figure 18- 5 Clusters K-means 35
Figure 19- t-SNE Visualization of 60 Components 36
Figure 20- Silhouette Score 37
Figure 21- Average Profile 39

Figure 22- Monthly Usage Profile 40
Figure 23- Monthly Billing Profile 41
Figure 24- On-net Usage Slabs 42
Figure 25- Mobile Usage Slabs 42
Figure 26- Total Bill Slabs 43
Figure 27- Mobile Bill Slabs 44
Figure 28- Base Model 49

LIST OF ABBREVIATIONS

RFM	-	Recency, Frequency, Monetary
CVM	-	Customer Value Management
CHAID	-	Chi-squared Automatic Interaction Detection
CLTV	-	Customer Lifetime Value
SVM	-	Support Vector Machine
ANN	-	Artificial Neural Network
XGBoost	-	Extreme Gradient Boosting
CPHM	-	Cox Proportional Hazard Model
MLP	-	Multi-layer Perceptron
PSO	-	Particle Swarm Optimization
PLS	-	Partial Least Squares
CDR	-	Call Detail Record
CNN	-	Convolutional Neural Network

CHAPTER 1

INTRODUCTION

We are living in the age of data and technology where huge amount of data is collected every day in different industries. However, this data is wasteful and useless if no useful information is derived from it. The analysis and patterns extracted from this data help us to understand the business and customers in various industries and can therefore enable us to devise effective business and marketing strategies and improve customer relationship management [1]. Machine learning and deep learning techniques help us in analyzing this data to extract useful information which leads to cost-effective business strategies. These business strategies are in terms of customer value management, customer retention and churn management. Customer segmentation allows us to meet above mentioned objectives by understanding each customer individually.

Customer Segmentation refers to the grouping of existing and potential customers of an organization based on the similarities between them. It helps organizations to know their customers better, to enhance their products, and to change and improve their marketing strategies. Many companies face difficulty in finding customer segments that can help them in their weekly or monthly campaigns which leads to failed CVM or digital campaigns and waste of resources used for marketing purposes [2]. Customer segmentation has been part of the business for a long time and the marketers used this to target their customers by offering what the customers want. This helps in marketing products and services to only those customers that show signs of propensity for buying the products. The propensity of buying is generally based on demographics, customer behavior, geography, psychographics, socio-cultural variables [2][3], needs, and customer value. Marketing is targeted to specific customer groups rather than to the whole customer base, which wastes company's money, time, efforts, and other useful resources. To divide the customers into various segments, it is pertinent to know who the customers are, what they do, what are their needs, and what they want from a company. Segmentation can be an effective tool for improving and maintaining long-term relationships with valued customers which helps in increasing the value given to and by the customer and increases the profitability [1].

Customer segmentation is common in industries such as telecommunication, banking sector, automobile industry, clothing, retail, and e-commerce sector. Pakistan is a country that has a strong competitive telecommunication market [4] and competitive market strategies are required to survive in this struggling economy. The telecom industry is one of the largest industries in Pakistan with 189 million cellular subscribers, 128 million mobile broadband subscribers, 3.0 million basic telephony subscribers, and 131 million broadband subscribers.

This research aims to perform feature analysis followed by pattern recognition and the formation of a customer segmentation model using telecommunication industry data of a well-reputed telecom company. We then aim to perform post cluster analysis and develop a Multi-Offer Recommendation model which predicts the propensity of a customer buying a specific product. The selected telecom company has a large customer base and a huge magnitude of usage and data which is difficult to analyze manually. Therefore, segmentation becomes a necessity to target specific customers according to their behavior and needs, to increase revenue, and to use the company's marketing resources efficiently and competently.

1.1. Motivation

Early identification of customer segments is a crucial task and can help save resources such as time and money used in campaign design and implementation. Identifying segments early during the process leads to specific campaign design, keeping in mind the cost that will be spent on these campaigns and what revenue uptake is expected from the relevant customer segment. It helps in finding out if the customer is a low-value customer, budgeted customer, the products the customer is likely to purchase, and helps in anticipating the customer lifetime value that the customer will be providing to the company. Moreover, an automated Machine learning based Multi-Offer Recommendation model will help to learn patterns of existing bundled customers and

apply it on non-bundled customers to increase product subscriptions while simultaneously reducing manual effort in the campaigns process.

1.2. Research Gap

Noticeable work has been done in the retail, e-commerce, and telecom sectors on customer segmentation. Logistic regression, cluster regression, CHAID (Chi-squared automatic interaction detection), RFM (Recency, Frequency, Monetary), and K-means clustering are commonly used techniques in the retail and e-commerce industry. Logistic regression, SVM, and two-step clustering are used in the Telecom sector.

RFM and k-means clustering are used in [2] and applied to sports retail store chain data in Turkey. They concluded that considering the large number of parameters in customer segmentation can provide reliable and satisfying results. It is also stated that loyalty card customers have a higher spending rate than customers without loyalty cards. The customer segmentation model using RFM, and K-means clustering can improve the purchase rate of customers as in [5] where the purchase rate in the SMS campaign was increased to 1 percent which was 0.1 percent in the old model. This helps the managers in targeting their customers according to their needs and behavior and offers them exciting vouchers. Classification algorithms such as multi-layer perceptron can also be used to make customer segments in the presence of target classes as in [1]. In this study, customer segmentation is done on demographic properties such as gender, age, and spending score which concluded that young males spend more than young females, a significant number of males and females lie in the ‘Elite’ category whereas only young males and female lie in ‘gold’ and ‘silver’ category. The customer segmentation model also helps in growing the active customer base and total purchase volume as evident in [6], the number of active customers increased by 529 and total purchase volume increased by 279% by using the RFM model and K-means clustering.

In the telecom sector, SPSS analysis is performed in [7] on China's two major telecom companies' data with customers' basic profile, SMS, call, cost, and other business-related information and concluded that call headers may have a positive impact on monthly expense.

Postpaid business-related work has been done in developed countries as evident in [8], a case study that uses data from a Telecom company and SVM to classify customer segments. It concludes that the results were not as high as expected possibly due to an inadequate feature set to make customers' profiles.

Prepaid-related work is done in Pakistan using data from a mobile telecommunication company as studied in [4]. Limited research has been found on fixed-line postpaid services in developing countries. Two-phase clustering algorithms using K means and RFM (Recency, Frequency, Monetary) by [9] are used for analysis for mobile telecom data analysis [4] however it is not used for postpaid fixed line segment in developing countries.

Limited work is found on offer recommendation systems in telecom. [10] research includes churn analysis due to increasing number of mobile operators where the customers can easily change switch to another telecom operator based on telecom services or pricing. To solve this problem, they developed a hybrid machine learning classifier to predict churning customers along with a rule-based model to recommend different plans to customers. Decision tree is used to suggest different plans to reduce customer's bill. [11] developed a package recommendation system based on CNN. They proposed a model called FGCIN that has the capability to capture the most prominent features which helps to generate new features through CNN and is effective for telecom package recommendation.

Our dataset comprises customer bundle profiles, with customers' subscriptions of multiple bundles. The distinct nature of our use case allows us to employ Multi-Label Classification/Regression, leading to the creation of a Multi-Offer Recommendation

Model. This model is designed to predict the likelihood of a customer, currently without any bundle subscriptions, subscribing to one or more bundles based on their usage and billing behavior. The aim is to provide the best model meeting customers' need while simultaneously increasing product subscriptions and revenue for the segment.

To accomplish this, we conducted a thorough analysis, comparing clustering models like DBSCAN and K-means clustering. Additionally, we plan to investigate the application of a Multi-Offer Recommendation Model using a Random Forest Classifier/Regressor and Multi-Output Classifier/Regressor Wrappers. Our goal is to determine whether classification or regression is more suitable for real-time business environments when implementing the Multi-Offer Recommendation Model. This approach will enhance our understanding and prediction of customer behavior regarding bundle subscriptions and increasing fixed revenue monthly.

1.3. Problem Statement

Customer segmentation helps companies to know their customers better such as their behavioral trends, spending patterns, and the propensity of buying a product or service. It also helps companies retain their valued customers and implement effective marketing strategies based on customers' behaviors. Machine learning models can be effectively used for customer segmentation.

Marketing products to the whole customer base without knowing the behavior of the customers is an ineffective marketing strategy as it wastes companies' resources such as time, money, and effort and hence reduces the chances of customer retention. Machine learning models such as SVM, logistic regression, and k-means are used for customer segmentation.

We want to make marketing strategies effective to retain valued customers and to increase product subscriptions for the fixed-line postpaid segment of a well-reputed

telecom company through feature analysis, pattern recognition, customer segmentation model and Multi-Offer Recommendation Model. We are going to perform a comparative study for clustering to identify customer segments followed by classification/regression Multi-Offer Recommendation model for predictive purposes.

1.4. Research Questions

RQ 1: Which clustering algorithm performs best on this specific set of telecom data for customer segmentation?

RQ 2: Which performs more effectively in Multi-Offer Recommendation Model: Classifier or Regressor?

RQ 3: Which features played the most crucial role in predicting product propensity in the Multi-Offer Recommendation Model?

RQ 4: Do the proposed model and strategies work effectively in real time business environment?

1.5. Research Objectives

The objectives of this research are:

1. To find out which clustering algorithm performs best with our dataset.
2. To find out which of the two: classifier or regressor works best with Multi-Offer Recommendation model.
3. To find out features crucial in predicting product propensity.
4. To analyze if our proposed model and strategies work effectively in real time business environment.

1.6. Contribution of the study

This section of the thesis discusses the significant contributions that have been accumulated as outcome of this research:

- Application of exploratory data analysis (EDA) to understand the specific set of data acquired from a well renowned telecom operator in Pakistan.
- PCA using variance ratio and two principal components.
- Implementation of DBSCAN and K-means through hit and trial method and silhouette scores to find optimal number of clusters respectively.
- Comparison of DBSCAN and K-means clustering using evaluation metric such as silhouette score.
- Post cluster analysis was performed using K-means identified clusters which concluded that the clusters were rightly identified and aligned with the domain and business knowledge. This eventually implies that K-means is suitable for this set of telecom data.
- We applied Multi-Offer Recommendation Model using random forest as base classifier and base regressor. This was done to determine which of the two, classifiers and regressors, perform well with this data.

1.7. Outline of this thesis

The organization of this paper is as follows: **Chapter 1** "Introduction" section includes the introduction of the study. **Chapter 2** "Literature Review" section summarizes the previous work. **Chapter 3** "Research Methodology" section explains the methodology used to extract patterns in the data and implementation of the customer segmentation and recommendation model. **Chapter 4** "Results and Evaluation" section shows the inferences and results obtained. Finally, **Chapter 5** "Conclusion" section highlights our contributions, and plans to extend this work.

CHAPTER 2

LITERATURE REVIEW

This section gives an overview of the prior research work that has been done related to customer segmentation. Since customer segmentation, churn management, product proposition and retention management are very much inter-related, such research work focusing on these elements is done in retail, e-commerce, banking, and telecom sectors. Customer segmentation, if done precisely, cand help develop effective product propositions, reduce churn, and hence increase customer retention and customer happiness index. Machine learning models such as SVM, K-means clustering [2][6][12] [5][4][13], logistic regression [3][14], CHAID [3][14] and RFM [2][14][15][13] are commonly used for customer segmentation. Similarly, machine learning classifiers such as Random Forest, AdaBoost, Multi-layer perceptron, Bayesian network and Neural Network can be used for churn management[16][17][18][19][20].

[2] focuses on customer segmentation in retail industry using data of biggest sports retailer in Turkey and applying a RFM model and K-means clustering techniques. The dataset contains 715,328 registers of customers. They concluded that customer segmentation which is done by considering substantial number of features could enhance the reliability of the model. Likewise, [6] used RFM and K-means using PCA on real world online transactions data of 1,013 customers of a large enterprise in China for customer segmentation. The study finds that after the customer segmentation, an addition of total 529 customers was observed in the total active customers base with an increase of 279% in total volume purchase and increase of 101.97% in total consumption amount. [14] performed a comparative study of RFM, CHAID and logistic regression on two different datasets and found that CHAID tends to have better performance as compared to RFM however it is suitable in other circumstances.

In telecom sector, many telecom companies cluster their mobile users on billing data however [12] proposed the use of CDR for clustering purposes after thorough clustering on real CDR. They proposed marketing strategies to China Mobile, a mobile operator in China. Results show that new customers of China Mobile increased to 64% of the total new mobile customers ultimately proving the success of marketing strategies. Furthermore, [4] conducted a customer segmentation and analysis of a mobile telecom company in Pakistan by using two step clustering technique on nine-days call usage, SMS usage and recharge data of 5019 customers. The results argue that calls, SMS, and VAS revenue is highly in correlation with total revenue generated by customers which implies that calls, SMS, and VAS usage contributes to the total revenue. The more the usage, the more revenue generated by the customers. It was also observed that VAS usage was greater in number than SMS usage among all customer segments. Five clusters were identified in the study in which cluster 1 represented high revenue loyal customers of the company, cluster 2 represented customers who were making calls, cluster 3 identified customers who were calling and using VAS, cluster 4 targeted the customers contributing least to the total revenue and cluster 5 of customers that were generating VAS revenue.

Moreover, [8] conducted a Vodafone case study using postpaid business customers' data collected from Vodafone Teradata CDRs (Call Detail Records) with the help of SVM. In this research, customer segmentation was based upon incoming and outgoing call usage behavior of the customers. SVM makes it possible to classify the customer segment according to his profile and behavior in 80.35 of the cases with four segments whereas with six segments, 78.5% of the classification can be achieved according to Vodafone case study. In addition, [21] proposed a propensity to buy model for fixed line telecom customers using logistic regressions to identify customers with the highest propensity to buy the companies new products and services They concluded that parameters related to "Customer Experience" drive the purchase propensity of the customer.

Additionally, an integrated framework is proposed by [18] that combines both customer segmentation and churn prediction using three different datasets. The study claims that retaining existing customers is much easier than acquiring new customers[18][22], hence is a better strategy. The authors used machine learning classification techniques for model evaluation and churn prediction and k-means clustering for churn customer segmentation in which customer segments are identified to devise and adopt relevant retention strategies. AdaBoost, Random Forest and Multi-layer Perceptron algorithms are used for model evaluation on Accuracy and F1-score metrics. According to the evaluation results, AdaBoost had the highest accuracy and F1-score in dataset 1 with 77.19% accuracy and 63.11% F1 score. Random Forest classifier was the top performer with 93.6% accuracy and 77.20% F1-score for dataset 2. For dataset 3, Random Forest had the highest accuracy of 63.09%, on the other hand, multi-layer perceptron had the highest F1-score of 42.84%. This research helps understand the churn factor for each cluster and hence devise suitable retention strategies.

New customer acquisition costs are raising exponentially due to factors such as emerging competitors [23], innovative business ideas and optimized services [24]. These factors proved to be eye opening for customer-oriented companies which led them to realize the importance of customer retention [24][18] hence focusing on a customer leaving phenomenon called churn. Since churn prediction and customer segmentation together can bring value to the customer and the company it is very crucial to understand the relationship between the two. [23] claims that their proposed churn prediction model produced better results of churn classification by using the Random Forest Algorithm and K-means algorithm for customer profiling. [25] research suggests the use of boosting for churn prediction which helps in good partition of data and results in identification of customers on a high risk to churn. These high-risk customers can be prioritized for retention offers. [20] focused on the imbalance data distribution and proposed Improved Balanced Random Forests (IBRF) and how it can be used to predict churning behavior. IBRF technique was applied on real bank customers churning dataset. It can learn best features through iterative adjustments to the class distribution and by imposing high

penalty of misidentification of the minority class. It was observed that it enhanced prediction accuracy to a significant level when compared to alternative algorithms such as Decision Trees, ANN (Artificial Neural Network) and Class Weighted Core Support Vector Machines (CWC-SVM).

Multiple machine learning algorithms such as ANN, SVM, naïve bayes and XGBoost and CPHM are used in [22]. It is evident in the study that deep learning techniques perform better in more complex structures and can provide higher success rates upon improvement. [26] focuses on customer churn behaviors and retention based on PLS method. The datasets primarily used are obtained from Teradata Center, with 171 input features and 100,000 instances for original dataset. The model proposed in the study has greater potential than traditional models used for prediction in terms of performance and in identification of key input variables.

Furthermore, customer retention, customer loyalty and customer satisfaction are goals prioritized by the telecom operators as new customer acquisition is challenging due to saturated and matured market and presence of competition [27][28]. It is observed by [27] in a constructive and casual link study set in German telecom market that customer satisfaction has a substantial impact on customer loyalty. They claim that customer retention is impacted by causal relationship of customer satisfaction and customer loyalty where customer satisfaction drives customer loyalty which ultimately influences customer retention. It is recommended that service providers improve the customers' perception of the core services that they provide to them rather than relying on customer satisfaction and customer loyalty. Likewise, [28] focusing on Korean telecom market observed that service quality plays an essential role along with customer-oriented services to increase customer satisfaction which ultimately increases customer loyalty. Therefore, efforts should be made to improve the quality of the core services along with maximization of the switching barrier which primarily focuses on the difficulty of switching to another service provider. Factors such as call quality, customer care support,

and VAS (value added services) appeared to have greater significance in customer satisfaction.

To maintain customer retention, it is important to provide services to customers that meet their needs. For this, recommendation systems are used that predict different products relevant to customers based on their behavioral patterns. These systems are used in industries such as e-commerce, entertainment tourism and telecom etc. [10] focused on early churn prediction using a hybrid machine learning classifier based on logistic regression and voted perceptron. These predictions were based on the CDR parameters. The churn dataset used for training consisted of features such as voice mail plan, day calls, night calls, international calls, charges etc. Decision tree was used to suggest different plans for different kinds of usage such as calls, data, or messages. For those customers that are potential churners, this decision tree will suggest plans to reduce the customers' billing. [11] study focuses on package recommendation using CNN model. A click-through prediction model was built to model the interaction between the characteristics of users and packages. Their FGCIN model captures important feature interactions through CNN FGCNN and generates new features.

Table 1 - Reviewed Research Work

Ref.	**Year**	**Context**	**Key Findings**	**Limitations & Future Work**
[1]	2020	• Mall customers' data (200 records, 4 input variables) • Classification – WEKA, MLP, Regression, J48 and Naïve Bayes	MLP performed the best with 98.33% accuracy in comparison with Regression, J48 and Naïve Bayes.	• Dataset size could be increased. • K-means and other ensemble algorithms can be used in future. • Add customer related features.

		• Demographics based customer segmentation		
[6]	2020	• Enterprise data of shopping platform in China • 1,013 customers data with 10,248 rows • PCA applied. • RFM and K-means Clustering.	• Active customers increased by 529. • Purchase volume increased by 279% as result of segmentation. • Consumption amount increased by 101.97%.	• Suitable algorithms are required due to data updating every day. • How these algorithms can be integrated with CRM system.
[29]	2020	• 100,000 subscribers • 220 input features reduced to 20. • Customer information • PCA and Autoencoder Neural Network • K-means clustering	• PCA and novel approach Autoencoder Neural Network used to reduce dimensions. • Original and reduced dimensions analyzed using K-means clustering. • Autoencoder performed the best for the datatype used.	• Autoencoder can be optimized using PCA principles.

[30]	2021	• "Customer Churn Prediction 2020" telecom dataset. • 19 input features with 4250 records • Cross validation technique • WEKA, wrapper-based approach, PSO • Decision Tree, Naïve Bayes, KNN and Logistic Regression	• Decision Trees performed best with accuracy 94.56% with 8 features. • PSO proved to be working best with this algorithm for best features identification and performance improvement.	• This approach can be useful in any other industry for churn prediction.
[7]	2018	• SPSS Analysis of China's telco data • 4126 records • Customer information, call and cost behaviors	• Consumer habits vary from consumer to consumer. • Call variables have a positive relationship with monthly bill.	• The reliability of the model can be further analyzed by using real time data of other operators. • Repeated data testing is suggested to improve the model's effectiveness.
[4]	2013	• 5,019 customers	• Call data, SMS data and VAS	• Future models can focus on revenue

		• 9 days of Prepaid Mobile data (calls, SMS, recharge) • Two Step Clustering	revenue are highly correlated with the total final revenue. • SMS usage is less than VAS usage in all segments identified.	prediction based on call behavior, SMS, and VAS.

CHAPTER 3

RESEARCH METHODOLOGY

3.1. Introduction

This chapter outlines the research methodology and conceptual framework that is used in this research focusing on customer segmented clustering.

3.2. Research Methodology

This section describes the research methodology of this study including the conceptual framework, data collection, data preprocessing, Exploratory Data Analysis (EDA), Principal Component Analysis (PCA) and machine learning model development followed by post cluster analysis. Firstly, exploratory data analysis was performed to understand the characteristics, correlation between features and distribution of the dataset. Secondly, data preprocessing was performed on the dataset such as handling missing values using imputer, one hot encoding on categorical variables and feature scaling. Next, encoded categorical data and scaled numerical data was concatenated to be used in PCA which then led to the machine learning model development. DBSCAN and K-means clustering was used to identify which model performs best with our dataset. We then performed post cluster analysis to understand the characteristics of our clusters. Following the clustering analysis, we moved on to classification and regression models which ultimately helped us in creating a Multi-Offer Recommendation model. Manual and Model predicted campaigns were executed in real time environment to compare the results of the models.

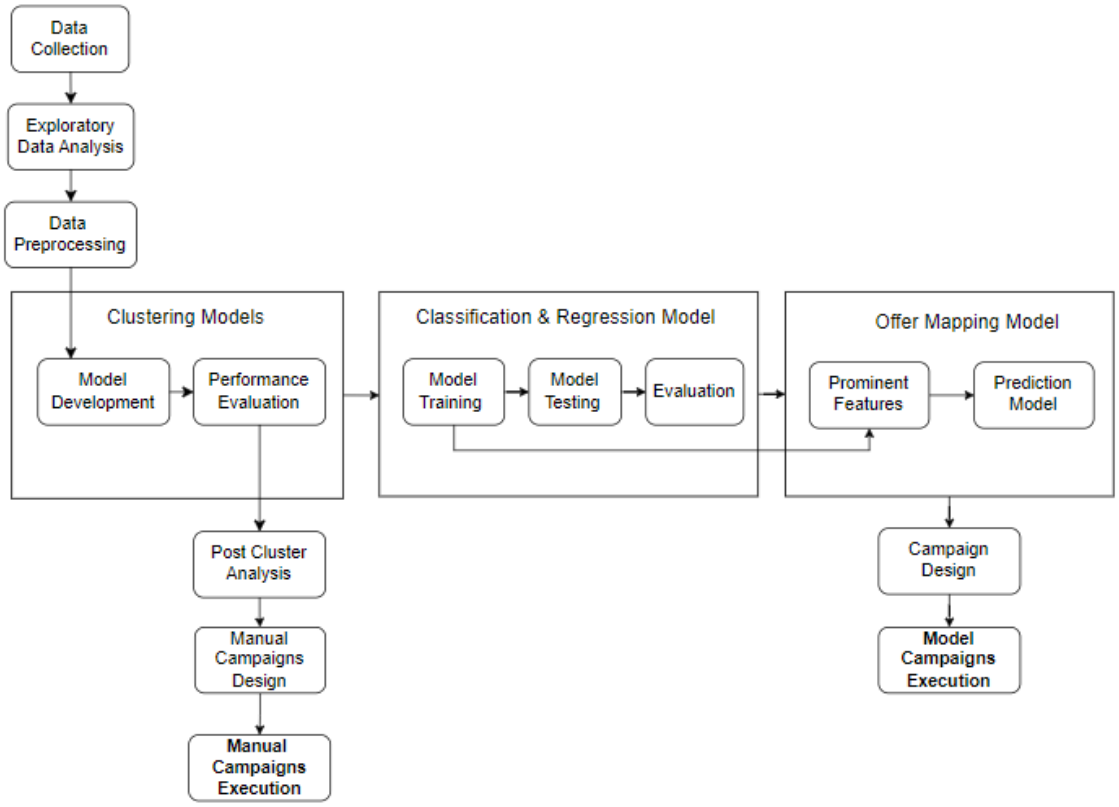

Figure 1- Research Methodology

3.3. Conceptual Framework

To develop the conceptual framework for our research, we studied various papers by esteemed authors to understand the theoretical basis of this area of research. Call usage data (CDRs) [4][8][7], SMS data, VAS data and recharge information is used to identify customer segments [4]. [7] performed SPSS analysis on 4,126 data samples containing basic information, SMS, call, cost, and other business information from China's major telecom companies.

Based on the above referred literature, we identified some important factors such as customer profile, usage data and revenue/billing data of customers and combined them into one which shall act as a conceptual framework of our research and shall help us in

identifying distinct customer segments. These factors are further divided into different variables which are mentioned in the following section. The conceptual/theoretical framework of customer segmentation is presented in the figure below:

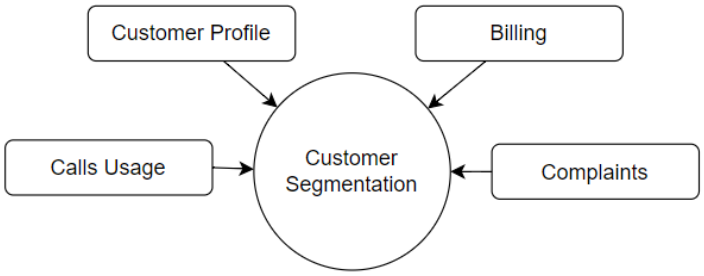

Figure 2- Conceptual Framework

These three categories consist of several variables that determine their scope. Next section shows the list of variables for customer profile, calls usage, billing and complaints that are used to identify customer segments. We distinguish our customer segments based on the three months profile of voice customers which allows us to devise a customized strategy for that customer segment.

3.4. Data Collection

This section describes the population sampling technique and dataset features used in the study.

3.4.1. Dataset Features

The target population in our study is the total fixed line postpaid customer base of the chosen company. We are targeting fixed line customers only in terms of voice usage, specifically outgoing traffic, to be able to devise strategies for each segment. We want to

utilize our technical knowledge with business knowledge to propose the best strategic way to improve performance in this segment as it is a declining product naturally due to the presence of OTT (Over the Top) platforms such as WhatsApp and IMO. For this purpose, we have chosen a sample of 4000 random fixed line postpaid customers. Three months data of customers is used for better understanding of the patterns and trends.

The data consists of three months call usage, billing profiles and complaints of 4000 random fixed line postpaid customers with 70 features. The data consists of four distinct categories that are:

3.4.1.1. Customer Profile

This includes the customer's basic profile such as cust_id, bundle, and age of customer. The following table shows basic profile variables as well as data types and description on each variable.

Table 2 - Customer Profile features

Customer Profile	Data Type	Description
CUST_ID	Integer	Masked Customer ID
NETWORK_AGE	Numerical	The network age of customer
BUNDLE	String	The bundles customer has subscribed
CATEGORY	String	If the customer is using the services for Home or Business
BUNDLED/NONBUNDLED	Boolean	If the customer has subscribed to new commercially available packages

3.4.1.2. Calls Usage Features

CDR data of individual customers indicates if the customer is willing to use the services of a service provider. If the customer is a high usage customer, they may be

targeted with high minutes bundles to increase customer engagement and voice traffic. Call usage data includes outgoing minutes usage of the customer such as on-net minutes, mobile minutes, and off-net minutes. Usage data includes Month 1 (M1), Month 2 (M2) and Month 3 (M3) features for each variable below.

Table 3 - Call Usage Features

Call Usage	Datatype	Description
ONNET_MINS	Numerical	On-net minutes generated by customer
MOBILE_MINS	Numerical	Mobile minutes generated by customer
OFFNET_MINS	Numerical	Other landline minutes generated by customer
AVG_ONNET_MINS	Numerical	Three months average of on-net minutes generated by customer
AVG_MOB_MINS	Numerical	Three months average of mobile minutes generated by customer
AVG_OFFNET_MINS	Numerical	Three months average of other landline minutes generated by customer

3.4.1.3. Billing Features

Billing data plays a crucial role in segmentation as it allows us to anticipate customers' payment habits and eventually helps us to identify low value and high value customers. The billing data of a customer includes the bill customer is paying to the company on monthly basis. This includes bundle charges (of any subscribed bundle), mobile charges, offnet charges, line rent, and monthly bill of the customer. Billing data includes Month 1 (M1), Month 2 (M2) and Month 3 (M3) features for each variable in table below.

Table 4 - Billing Features

Billing	Data Type	Description
LOCAL_ONNET_BILL	Numerical	Standard local on-net calling rates charged to customer
NATIONWIDE_ONNET_BILL	Numerical	Standard nationwide on-net calling rates charged to customer
LOCAL_MOBILE_BILL	Numerical	Standard local mobile calling rates charged to customer
NATIONWIDE_MOBILE_BILL	Numerical	Standard nationwide mobile calling rates charged to customer
LINE_RENTAL	Numerical	Line rent charged to customer monthly
BUNDLE_BILL	Numerical	Bundle charges charged to customer
OFFNET_BILL	Numerical	Other landline calling rates charged to customer
TOTAL_MOBILE_BILL	Numerical	Total mobile charges charged to customer
TOTAL_ONNET_BILL	Numerical	Total on-net charges charged to customer
TOTAL_BILL	Numerical	Total bill charged to customer
AVG_ONNET_BILL	Numerical	Three months average of on-net charges charged to customer
AVG_MOBILE_BILL	Numerical	Three months average of mobile charges charged to customer
AVG_OFFNET_BILL	Numerical	Three months average of other landline charges charged to customer
AVG_LINE_RENTAL	Numerical	Three months average of line rental charged to customer
AVG_BUNDLE_BILL	Numerical	Three months average of bundle charges charged to customer
AVG_TOTAL_BILL	Numerical	Three months average of total bill charged to customer

3.4.1.4. Complaints

Customer complaints helps us understand the probability of churn. If the complaints are higher for any segment, they are more likely to leave the services.

Table 5 - Customer Complaints

Complaints	Data Type	Description
COMPLAINTS	Numerical	Number of complaints generated by customer in a month

3.4.1.5. Dataset Sample

The following table shows a sample from the dataset. Categorical features are encoded using one hot encoding whereas feature scaling is performed on numerical features using Standard Scaler after exploratory data analysis.

CUST_ID	NETWORK_AGE	ONNET_MINS01	ONNET_MINS02	ONNET_MINS03	MOBILE_MINS01	MOBILE_MINS02	MOBILE_MINS03	LOCAL_ONNET_BILL01
1	5675.0	1.0	4.0	32.0	1.0	3.0	1.0	0.0
2	5522.0	255.0	242.0	197.0	95.0	84.0	59.0	0.0
3	5540.0	9.0	1.0	4.0	8.0	14.0	10.0	13.5
4	6516.0	28.0	5.0	32.0	42.0	53.5	1.0	0.0
5	565.0	37.0	34.0	36.0	104.0	215.0	152.0	0.0

Figure 3- Dataset Sample

3.4.2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed before preprocessing to understand the characteristics and distribution of the dataset. Performing analysis before preprocessing gives insights into the data such as identification of outliers, anomalies, relationship between variables and data distribution. Firstly, we performed description check to understand basic statistical values of the dataset such as count, mean and percent quartiles as shown in the figure below. From this, we can deduce that mobile minutes' consumption

is higher than on-net minutes consumption as the max value of mobile minutes is greater than on-net minutes and the mean is comparatively higher.

	CUST_ID	NETWORK_AGE	ONNET_MINS01	ONNET_MINS02	ONNET_MINS03	MOBILE_MINS01	MOBILE_MINS02	MOBILE_MINS03
count	4000.000000	3998.000000	2863.000000	2919.000000	2751.000000	2846.000000	2922.000000	2830.000000
mean	2000.500000	5756.415958	80.107579	106.001028	101.294075	137.763879	180.991786	166.008481
std	1154.844867	3553.555788	161.085567	252.987096	247.275848	373.273436	435.469191	405.401848
min	1.000000	3.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	1000.750000	2765.000000	7.000000	9.000000	9.000000	9.000000	10.000000	10.000000
50%	2000.500000	5479.500000	28.000000	34.000000	32.000000	42.000000	53.500000	50.500000
75%	3000.250000	8716.750000	87.000000	112.000000	102.000000	148.750000	201.750000	180.000000
max	4000.000000	17989.000000	2803.000000	7854.000000	7558.000000	10636.000000	10557.000000	10402.000000

Figure 4- Statistical Summary

3.4.2.1. Data Distribution

Furthermore, our data set includes a higher number of instances of Home customers than Business customers. It is generally assumed that Business customers are high value customers compared to Home customers that are mostly average billed customers. The assumption is proved as evident in the bill distribution plot below where business customers pay a higher billing amount on average than home customers, amount of Rs. 1004.24 and Rs. 940.94 respectively. This may give us an opportunity to design campaigns for Business customers to increase Average Revenue Per User (ARPU) of the customer base.

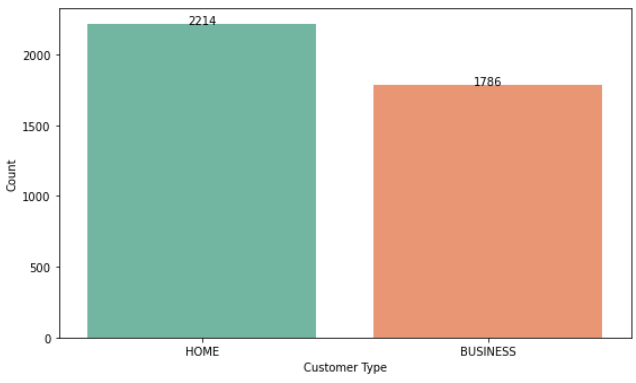

Figure 5- Distribution of Customer Type

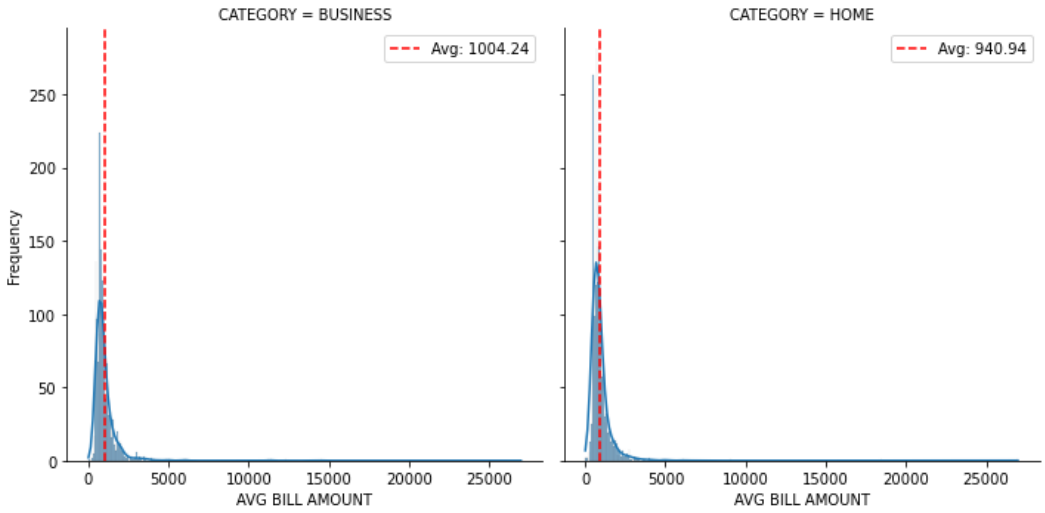

Figure 6- Bill Distribution by Customer Type

The dataset includes a higher number of instances for Bundled customers which means that the customer has subscribed to a calling package, while on the other hand, Non_Bundled implies that the customer is a Pay As You Go (PAYG) customer where standard calling rates are charged to the customer generating PAYG revenue. Bundled users enjoy several benefits as compared to Non_Bundled such as discounted prices and enhanced value features. It helps build stronger relationships with customers by providing value for money leading to high Customer Lifetime Value (CLTV) and thus higher retention rates.

As seen below, the average bill of Bundled Customer is higher than Non_Bundled customer since customer is giving fixed bundled charges every month, whereas Non_Bundled customer's usage varies month to month due to high calling rates. It is crucial and highly prioritized by telecom companies to convert Non_Bundled customer to Bundled so that PAYG revenue can be secured as fixed on monthly basis and fixed revenue provides a predictable cash flow and revenue in total. On the other hand, PAYG revenue fluctuates monthly with changes in demand, economic factors, and regional events such as protests and holidays.

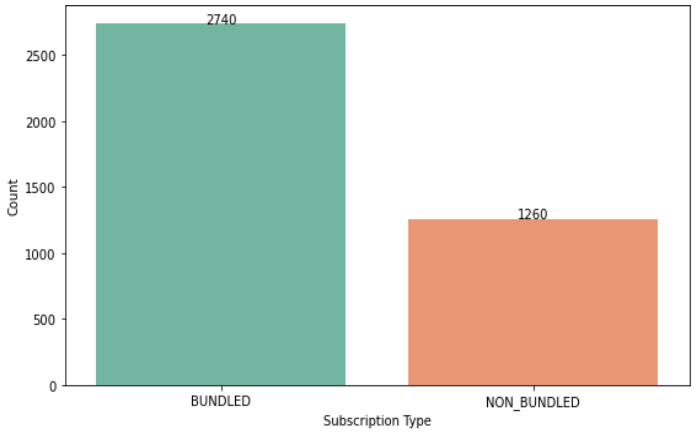

Figure 7- Distribution of Subscription Type

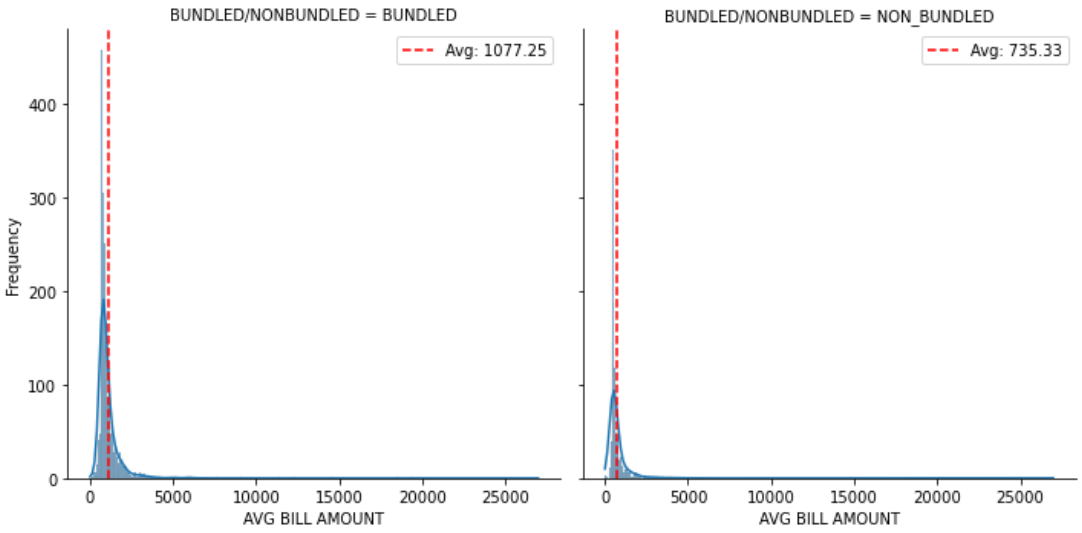

Figure 8 - Bill Distribution by Subscription Type

Density plots display the distribution of continuous variables. Below density plots of billing features show how data is distributed in terms of billing. As we have seen, data is rightly skewed which implies that most of the data points are concentrated towards the left side of the graph however there exist few larger values that are trailing off to the right. This is due to the presence of high value outliers in the data.

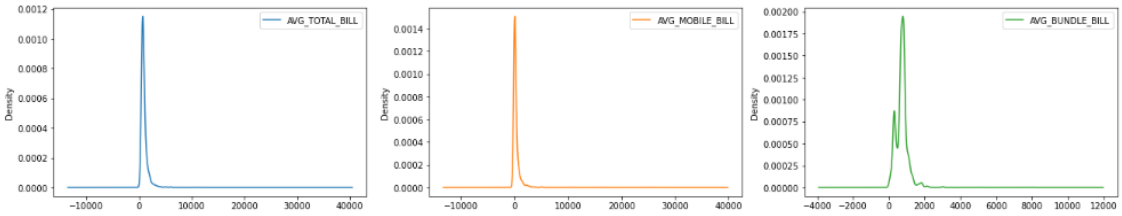

Figure 9- Density Plots for Billing Features

3.4.2.2. Correlation Analysis

Correlation analysis is a statistical method which is used to measure the strength and direction of the relationship between two variables. It involves a correlation coefficient ranging from -1 to 1.

- The correlation between the two variables is strongly positive if the value of the coefficient is 1 or close to 1. This indicates that increase in one variable is directly proportional to the other variable.
- The correlation between the two variables is strongly negative if the value of the coefficient is -1 or close to -1. This indicates that as the value of one variable increases, the other variable decreases.
- If the value of the coefficient is around 0, it indicates a weak relationship between the two variables.

Correlation analysis was performed to find out features that are highly correlated with the total bill that the customer must pay monthly. AVG_MOBILE_BILL feature is highly correlated with AVG_TOTAL_BILL with 0.88 correlation followed by AVG_MOBILE_MINS and AVG_BUNDLE_BILL with 0.83 and 0.31 correlation respectively. AVG_MOB_MINS is also mildly correlated with AVG_BUNDLE_BILL implying that bundled customers generate outgoing mobile minutes as they have subscribed to a bundle and get limited resources. AVG_ONNET_BILL has a correlation of 0.23 with AVG_TOTAL_BILL which is significantly lower than

AVG_MOBILE_BILL (0.88) indicating that customers tend to generate mobile minutes as compared to on-net minutes.

	AVG_ONNET_MINS	AVG_MOB_MINS	AVG_ONNET_BILL	AVG_MOBILE_BILL	AVG_LINE_RENTAL	AVG_BUNDLE_BILL	AVG_TOTAL_BILL
AVG_ONNET_MINS	1.000000	0.206858	0.653104	0.208004	-0.010215	0.106807	0.219811
AVG_MOB_MINS	0.206858	1.000000	0.116631	0.863085	-0.019530	0.389464	0.832785
AVG_ONNET_BILL	0.653104	0.116631	1.000000	0.156607	0.029764	-0.103190	0.226950
AVG_MOBILE_BILL	0.208004	0.863085	0.156607	1.000000	-0.022945	0.120910	0.881663
AVG_LINE_RENTAL	-0.010215	-0.019530	0.029764	-0.022945	1.000000	0.017557	0.017416
AVG_BUNDLE_BILL	0.106807	0.389464	-0.103190	0.120910	0.017557	1.000000	0.316165
AVG_TOTAL_BILL	0.219811	0.832785	0.226950	0.881663	0.017416	0.316165	1.000000

Figure 10- Correlation Analysis between Usage and Bill Features

3.4.2.3. Data Preprocessing

Data preprocessing is an essential step performed to clean, transform and prepare raw data for analysis. This helps in improving quality of the dataset by handling data inconsistencies, missing values, removing duplicates, and normalizing numerical features. Following data preprocessing steps were performed to achieve a dataset suitable for PCA and Machine learning models.

a) Categorical Variables

Categorical variables or non- numeric data requires to be converted to numeric to be fed into machine learning model. Categorical variables in our data were separated followed by one hot encoding since there is no ordinal relationship between categories. This led to the creation of separate binary columns for each category with value of ‘0’ and ‘1’ where ‘0’ is absence of the categorical value and ‘1’ is represents presence.

MOBILE_MINS_SLABS_ZERO	CATEGORY_BUSINESS	CATEGORY_HOME	BUNDLED/NONBUNDLED_BUNDLED	BUNDLED/NONBUNDLED_NON_BUNDLED
0	1	0	1	0
0	0	1	1	0
0	0	1	0	1
0	1	0	0	1
0	0	1	1	0

Figure 11- One Hot Encoding of Categorical Variables

b) Numerical Variables

Handling numerical data leads to an optimized machine learning model. We performed following steps to handle numerical data:

- Handling missing and special characters such as "?", blank values, "#Value!" and "#DIV/0"
- Conversion of other types of data such as "object" to "numeric"
- Handling null and infinite values by imputing median values as the data is rightly skewed as evident in distribution plots.

After handling both categorical and numerical data, we concatenated both types of data and normalized the data using Standard Scalar to transform the data to having a mean of 0 and Standard deviation of 1.

$$z = \frac{x - u}{6}$$

3.4.2.4. Principal Component Analysis (PCA)

Principal Component Analysis helps in reduction of dimensions while simultaneously preserving essential information in the data, ultimately helping in data analysis and machine learning model development.

a) Two Principal Components

Firstly, we applied two components PCA to reduce data complexity while preserving the variance in the dataset as well as to visualize the dataset in two dimensions to be able to understand the data and identify clusters. Our dataset is concentrated towards the left and is not spread out which implies that the clusters might not be distinguished easily and will be overlapped. We then used, elbow method to find the optimal number of clusters to be found in K-means clustering which results in four optimal clusters.

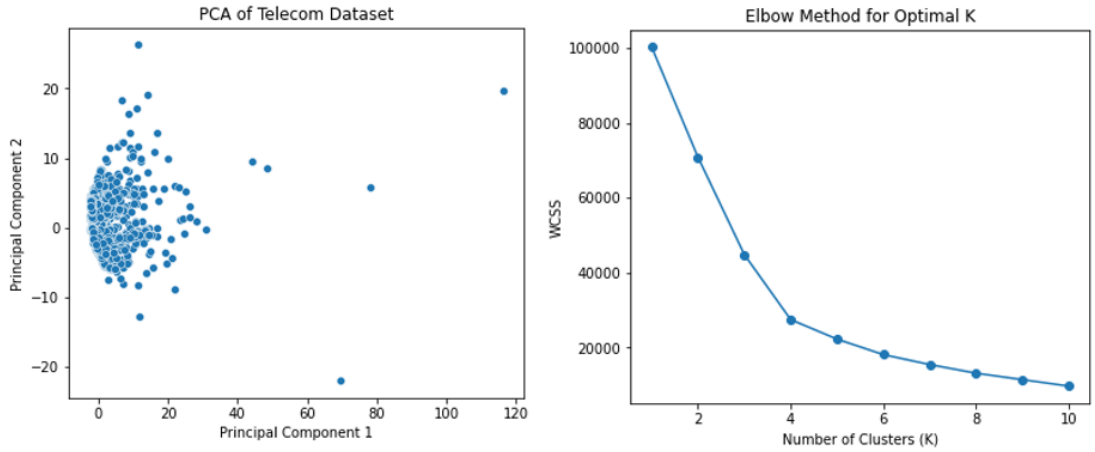

Figure 12- Two Components PCA and Elbow Method

b) Principal Components Identification using Variance Ratio

Variance ratio is a common approach that is used to decide the number of principal components. It indicates the proportion of variance of each principal component relative to total variance in the dataset. We applied variance ratio to get optimal number of principal components having variance ratio greater than 80% which implies that more than 50 principal components should be considered for optimal clusters as seen in the figures below. The elbow method was then applied on 50 and 60 principal components for comparison of optimal clusters which resulted in 4 to 5 clusters.

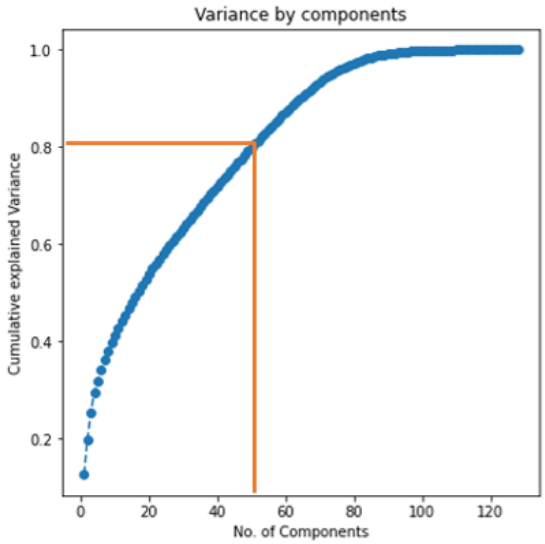

Figure 13- Variance Ratio

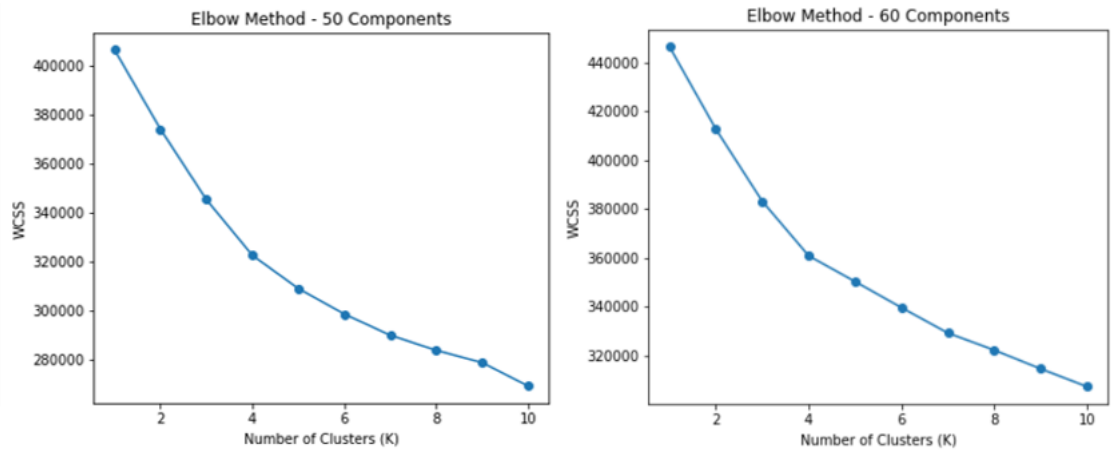

Figure 14- Elbow Method 50 and 60 Components

CHAPTER 4

RESULTS AND EVALUATION

Following an in-depth Exploratory Data Analysis (EDA) and careful data preparation, our focus shifted towards applying clustering algorithms to uncover the most suitable approach for our dataset. DBSCAN, which is known for its ability to identify density-based regions and noise in different shapes and sizes was used to detect clusters based on density connectivity. Simultaneously, K-means clustering, a partitioning method which aims to group data into K distinct clusters, was applied for comparative analysis. Through this comparative process, we sought to determine the algorithm that optimally aligns with our dataset's internal properties and characteristics, assisting a deeper understanding of its underlying patterns and relationships.

4.1. DBSCAN

We first implemented DBSCAN (Density-Based Spatial Clustering of Applications with Noise) for comparative analysis. DBSCAN clusters data using density of the data points in the data space. It defines clusters as areas of high density and defines noise as outliers. Hit and trial method and silhouette scores were employed to find out optimal epsilon value and minimum samples. Silhouette score is a unit of measure of the quality of clusters that are created by clustering algorithms. It ranges from -1 to 1, where:

- 1 indicates that the clusters are better defined with significant separation.
- 0 indicates that the samples are located closer to the decision boundary between two clusters neighbor to each other.
- -1 indicates that the samples are assigned to the wrong clusters.

Keeping the value of epsilon = 2 and minimum samples of 1000 provided us with two clusters and noise of 492 points. We then calculated the silhouette score to find out the quality of the clusters which resulted in a score of 0.50. Nevertheless, if we increase the number of clusters in DBSCAN, clusters are identified in an irregular manner, not suitable for post clustering analysis.

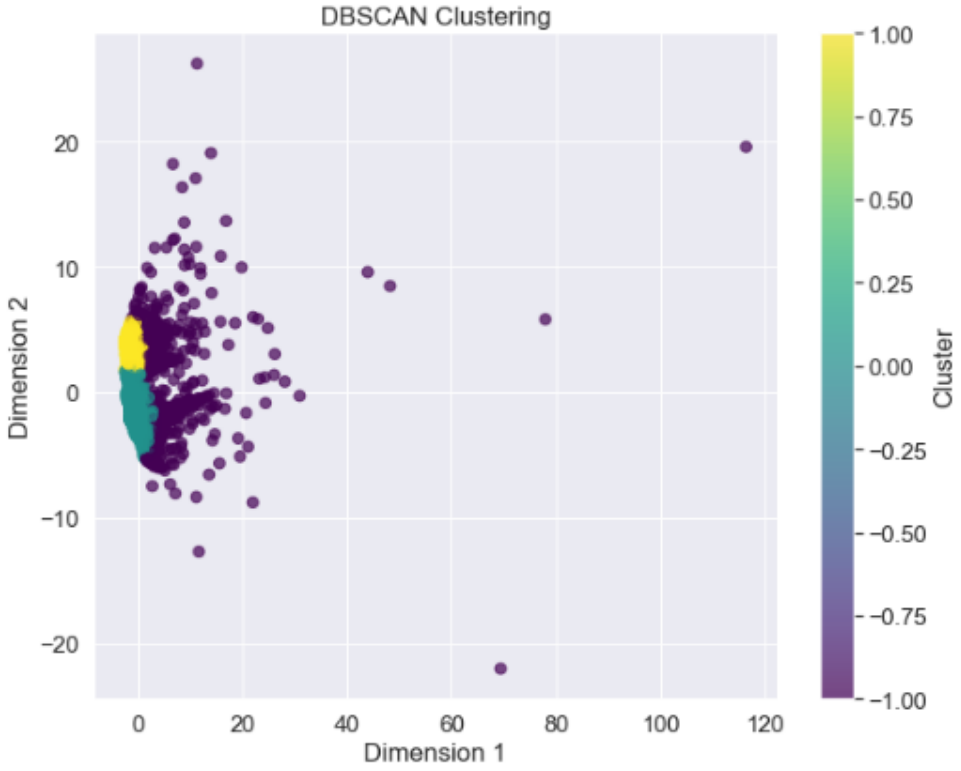

Figure 15- DBSCAN - 2 Clusters

If we increase the number of clusters, we get the following figure with one cluster having majority of data points and two small clusters along with noise in the data space. Even though these clusters have the highest silhouette score of 0.77, they are not suitable for analysis as all the datapoints are associated with one cluster instead of being spread out to other clusters. The noise in the data is also considered as Cluster 1 and 2 implying that DBSCAN, despite its effectiveness in identifying dense regions, is having difficulty delineating distinct clusters from our dataset. This could stem from the dataset's internal

characteristics such as high dimensionality making it challenging in defining clearly visible distinct clusters and simultaneously making it less suitable for our specific dataset.

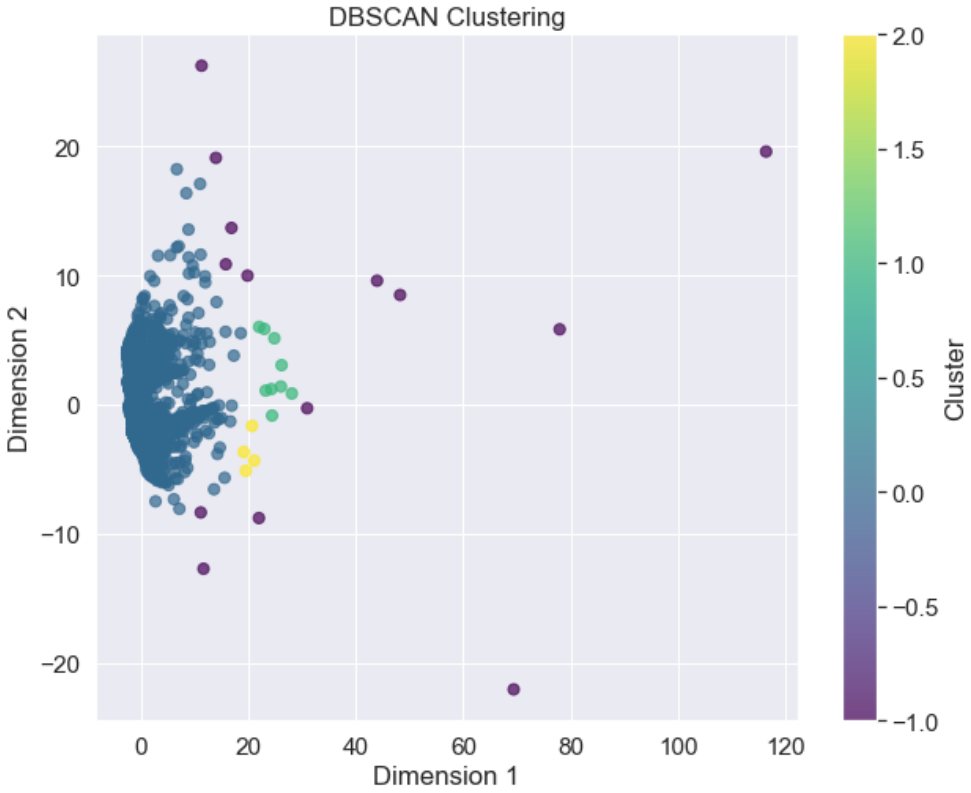

Figure 16- DBSCAN - 3 Clusters

4.2. K-means Clustering

After DBSCAN's inability to identify distinct clusters, we used alternate algorithm, k-means clustering model to fulfill the same purpose. K-means clustering is an unsupervised machine learning algorithm which is used to find distinct and non-overlapping clusters. Clusters are formed by grouping similar data points based on the calculated distance between them along with a cluster centroid. The data point is assigned to the cluster whose centroid is closest to the data point.

We first applied k-means on two principal components with 4 and 5 clusters followed by k-means clustering on 60 principal components using variance ratio. We used k-means++ algorithm for smart initialization of the centroids. K-means clustering picks a centroid on random basis and then other centroids are selected randomly from the data points like first centroid. However, k-means++ chooses the first centroid randomly from the dataset, then other points farthest away from existing centroid are chosen as centroids. This initialization process produces centroids that are spread out and thus reducing the chance of wrong and poor clustering. In short, k-means++ picks initial centroids that are spread out across the data leading to accurate identification of clusters.

4.2.1. K-means on Two Principal Components

We applied k-means clustering with k=4 and k=5 using two principal components for visualization of different clusters in two dimensions. For k=4, we can see that the clusters are concentrated towards the left and there exist some outliers forming a distant cluster. This also suggests that most of the data points have lower values. Cluster 1 includes a higher number of instances followed by Cluster 2.

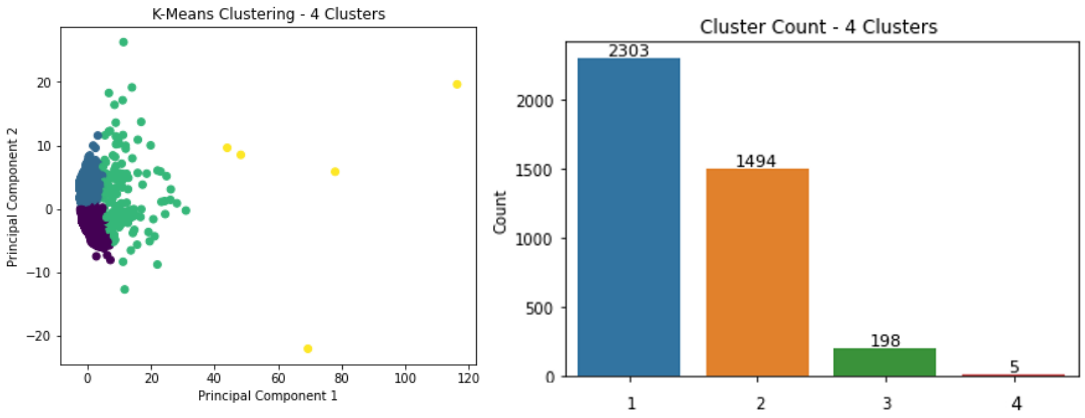

Figure 17- 4 Clusters K-means

We then applied five clusters k-means to analyze the effect of outliers on the data. We can see that extreme outliers are now separated into a different cluster. Cluster 1 count reduced from 2303 to 1792 and Cluster 2 from 1494 to 1442 and Cluster 3 count increased from 198 to 616.

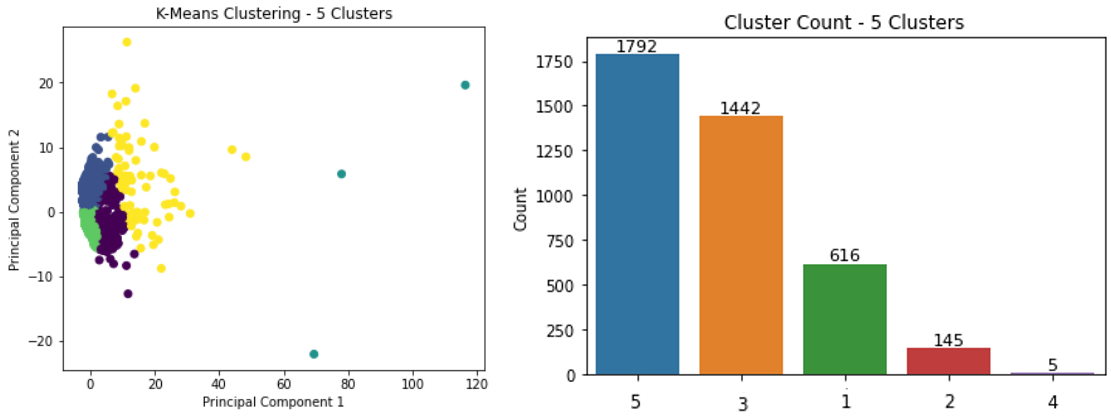

Figure 18- 5 Clusters K-means

4.2.2. K-means on 60 principal components

We then applied k-means on 60 principal components achieved through principal component analysis using variance ratio. 60 components were chosen to preserve more than 80 percent variance in the data as described in exploratory data analysis. We used t-SNE (t-Distributed Stochastic Neighbor Embedding) technique which is common for visualizing high dimensional data by reducing it two or three-dimensional spaces which makes interpretation relatively easier. The algorithm attempts to preserve similarity between the neighboring points hence the clustered points in t-SNE plots are representative of the points closer to each other in original high-dimensional space. As seen in the figures below, Cluster 0.0 seems to be the biggest cluster with multiple closely related data points followed by Cluster 2.0.

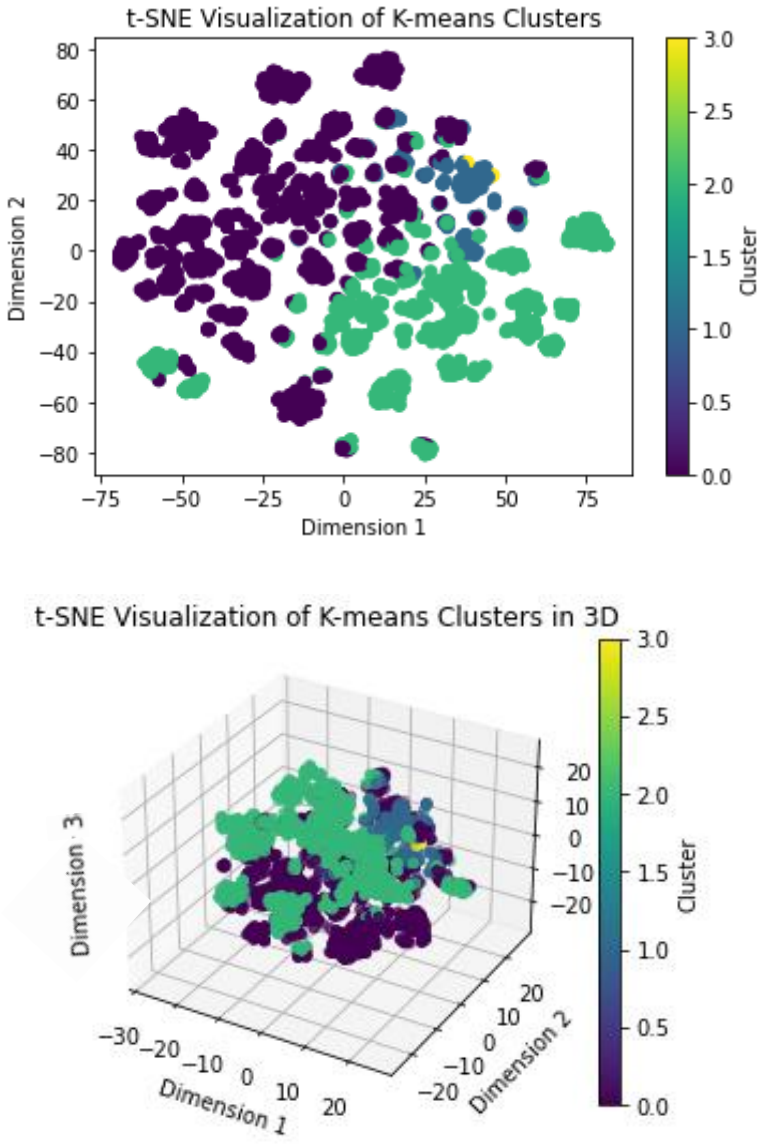

Figure 19- t-SNE Visualization of 60 Components

4.2.2.1. Silhouette Score

We then calculated silhouette score for clusters founded by using 60 principal components. As described above, silhouette score is a unit of measure of the quality of clusters that are created by clustering algorithms. Hence, the highest the score, the better the quality of the clusters. The silhouette score of clusters of our dataset lies close to zero,

that is 0.118 and 0.1217 precisely for four and five clusters respectively. This implies that the score is reasonable, however it is relatively low and there is room for improvement. This is possibly due to the small sample size of the dataset and may be improved on large dataset.

	Number of Clusters	Silhouette Score
0	2	0.118238
1	3	0.121417
2	4	0.118448
3	5	0.121718
4	6	0.130539
5	7	0.106021
6	8	0.056157
7	9	0.054680
8	10	0.071343

Figure 20- Silhouette Score

This silhouette score of 0.12 is significantly lower than DBSCAN clustering which was 0.50 and 0.77 however the number of clusters in DBSCAN are lower than k-means clustering, 2 and 4 clusters respectively.

4.3. Post Clustering Analysis

After applying the k-means clustering model, we performed post clustering analysis on the scaled data using k-means clusters. Post clustering analysis is the interpretation and evaluation of the clustering results performed on the following:

1. **Average Profile:** Customer profile, average billing, and usage
2. **Monthly Usage Profile:** Month1, 2 and 3 Usage profile and complaints
3. **Monthly Billing Profile:** Month 1, 2 and 3 billing profile

4.3.1. Average Profile

We first visualized the average customer profile using radar plots. Radar plots, also commonly known as spider plots, are a visualization tool used to display multivariate data in two dimensions. These plots help in comparing identified cluster profiles based on multiple variables. Following are some of the observations that can be deduced from these radar plots:

1. Cluster 1 includes a high number of bundled customers which means they have subscribed to a bundle and pay a fixed bundle amount monthly. Since bundles are subscribed, customers are generating outgoing calling minutes. These customers are apparently new customers as their network age is below average (0).
2. Cluster 2 mainly compromises the non_bundled customers with no subscribed bundle and high network age referring to customers being old and loyal. Home and business customers are equally distributed in this cluster. As avg_total_bill is lower than 0, it implies that the cluster is a low value cluster.
3. Cluster 3 is a high value cluster as billing amount, minutes and variable bill is higher in this group. Most of the customers in this cluster are non_bundled and new sales.
4. Cluster 4 is a cluster with extreme values and is separated for the same reason. These customers are non_bundled customers with extremely high mobile and offnet usage, variable mobile billing, and total billing.

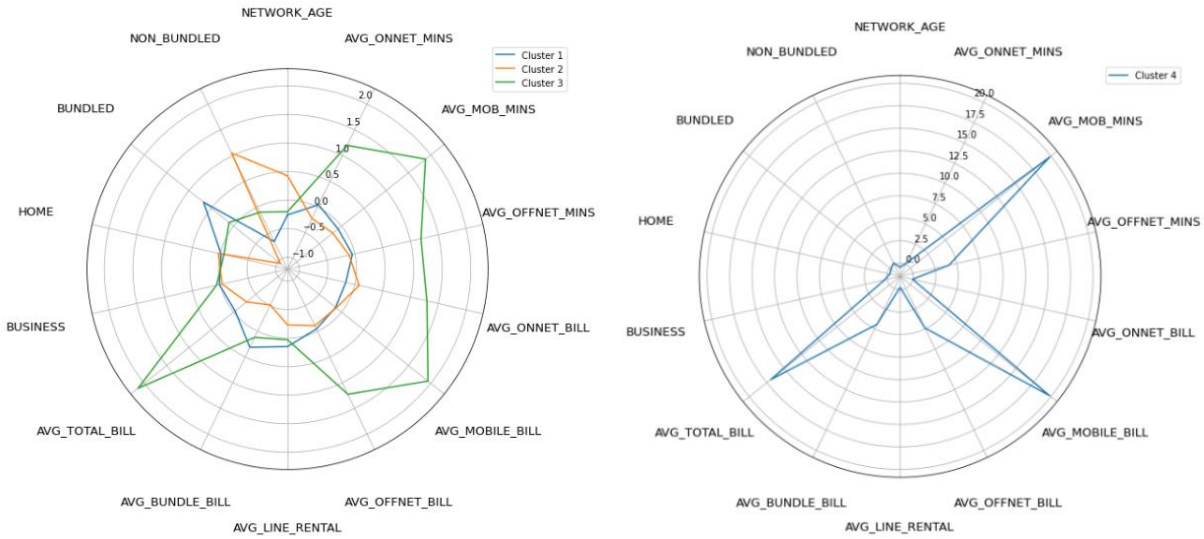

Figure 21- Average Profile

4.3.2. Monthly Usage Profile

Secondly, we made radar plots for customers' monthly usage and complaint trends. We analyzed three months data to discover if the trends are continued or one time. We observe the following from usage profile:

1. Cluster 1 generates on-net minutes a little above average, however mobile and offnet usage is lower than average. This cluster does not file complaints presumably due to low usage.
2. Cluster 2 is apparently a non-user customer group that generates minimum or no usage apparently due to being charged standard calling rates as they are non_bundled customers seen in above average profile.
3. Cluster 3 is a high voice user. The share of mobile usage is higher in this segment followed by on-net minutes and offnet minutes. This segment also tends to file customer complaints such as line issues, connectivity issues or billing issues etc.
4. Cluster 4 consists of extreme users. These customers have extremely high mobile usage followed by offnet minutes.

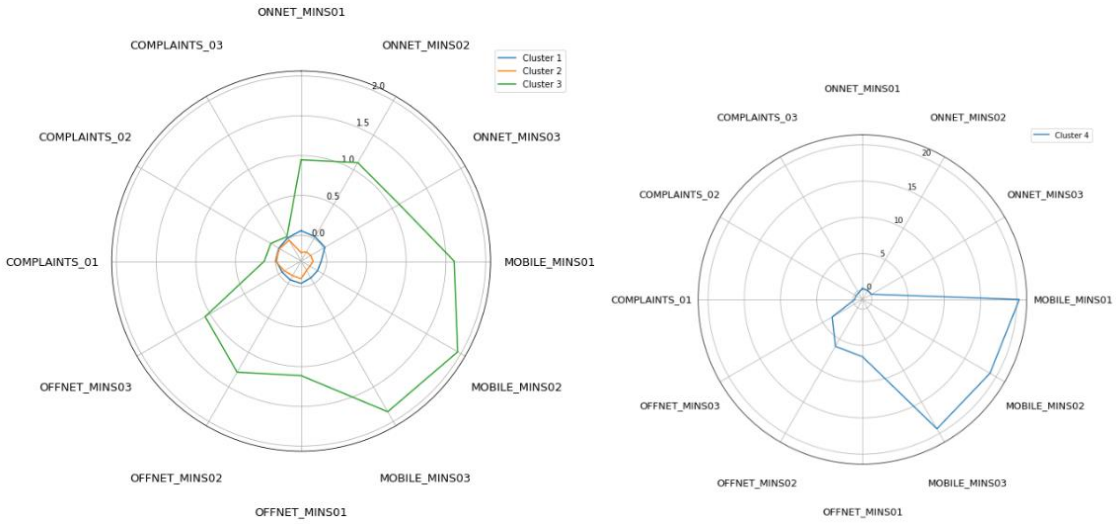

Figure 22- Monthly Usage Profile

4.3.3. Monthly Billing Profile

Furthermore, we also analyzed customers' billing profiles for month 1, month 2, and month 3. We make our observations on total bill, bundle bill, line rental bill, variable mobile and on-net bill charged to customer. Following points are concluded from cluster comparison on billing variables:

1. Cluster 1 has a high bundle bill for three consecutive months as they are bundled customers identified in average profile above. Their line rental bill is significantly below the average as they are not charged line rental. They pay a decent amount of total bill monthly due to bundle subscription.
2. Cluster 2 consists of high rental non_bundled customers. They are low value customers with the lowest total bill and on-net charges.
3. Cluster 3, as stated above, is a high value non_bundled customer with high variable mobile bill, and total bill.
4. Cluster 4 has an extremely high mobile variable bill and total bill.

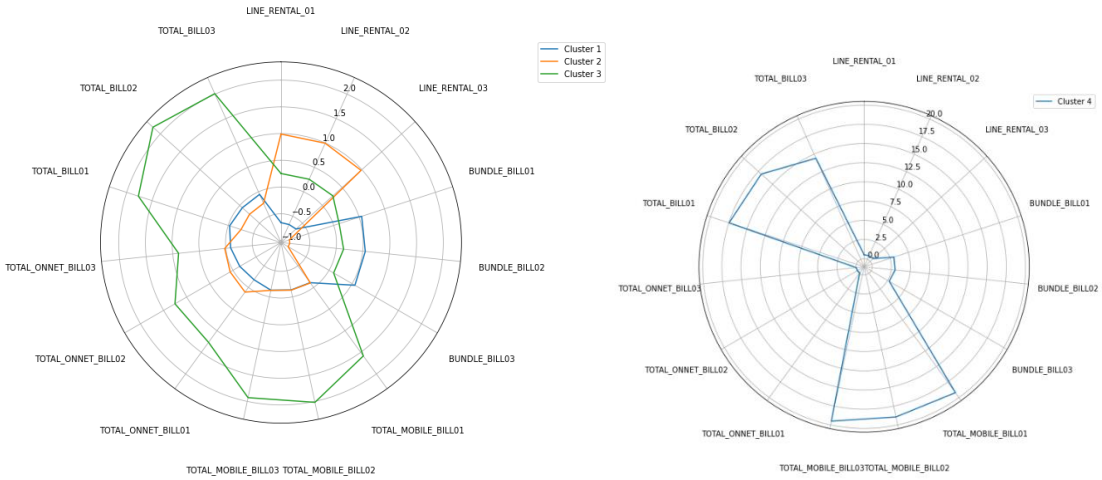

Figure 23- Monthly Billing Profile

4.3.4. Usage Slabs Analysis

We performed usage slab analysis on on-net minutes usage and mobile minutes usage as they are highly prominent in radar plots.

4.3.4.1. On-net Usage Slabs

As we can see in the count plots below, the majority of Cluster 1 and 2 customers fall under <50 on-net usage slab suggesting they are low value users. Cluster 1 and 2 also has customers with usage varying from 50 to 300. These customers can be offered on-net bundles depending on their variable on-net usage. A fair number of customers fall under the Zero usage slab indicating they are three months non -users. These customers can be given free minutes as a trial to enable customer experience.

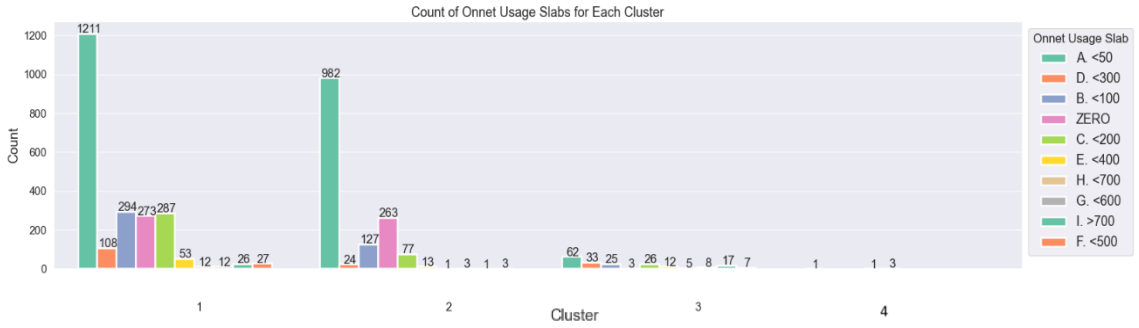

Figure 24- On-net Usage Slabs

4.3.4.2. Mobile Usage Slabs

Similarly, it is to note that mobile usage behavior is in correlation with on-net usage behavior. Most of the customers in Cluster 1 and 2 fall usage 50 minutes slab followed by 50 to 200 minutes slabs. Cluster 1 and 2 also has a noticeable number of non-users with zero usage that can be offered mobile minutes on trial to enable usage. Cluster 3 has high value users that consume more than 500 mobile minutes on average each month. These customers can be offered a mobile minute's bundle for securing revenue.

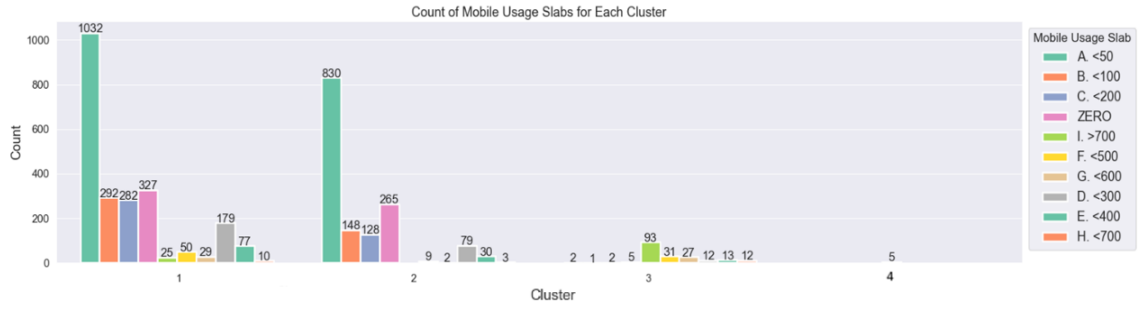

Figure 25- Mobile Usage Slabs

4.3.5. Bill Slabs Analysis

We performed slab analysis on total bill and mobile bill of the customer for further understanding of customer's payment behavior.

4.3.5.1. Total Bill Slabs

As we can see, around 484 customers in Cluster 1 pay a billing amount of less than PKR. 700, followed by 382 customers paying a billing amount greater than PKR. 1200 making them high value customers in voice segment. Next in line are 255 customers with a billing amount ranging from PKR. 800 to PKR. 850. Cluster 2 consists of low value customers with billing amounts ranging from PKR 0 to PKR 550. These are budgeted customers that need to be given special attention as they have a high propensity to churn due to low usage and billing. Cluster 3 consists mainly of high value customers with total bull exceeding PKR. 1200.

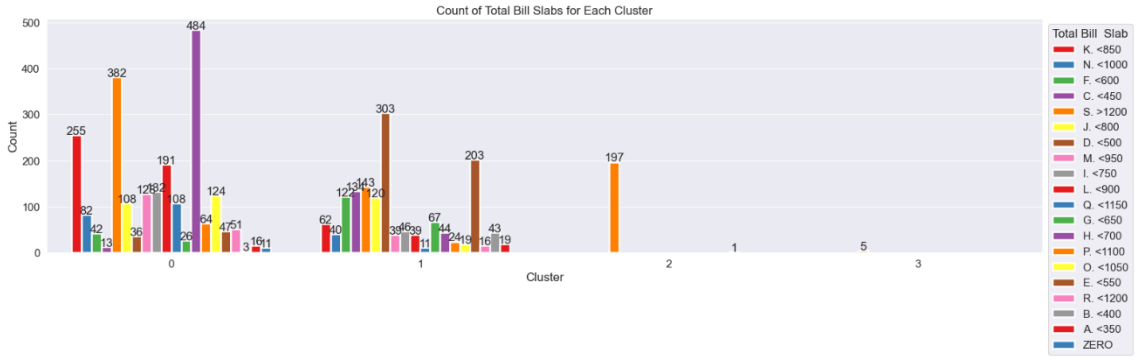

Figure 26- Total Bill Slabs

4.3.5.2. Mobile Bill Slabs

Moreover, further analysis was performed on the mobile bill slabs, a subset of total bill. It is evident that most of the customers in Cluster 1 have zero mobile charges since they are Bundled users and get minutes resources at discounted prices as part of their subscribed bundle. However, 782 customers are paying a bill amount less than PKR. 350, which is possible due to excessive mobile usage, also called overage. In Cluster 2, 975 customers pay mobile charges ranging from PKR 0 to PKR 350 like Cluster 1 nonetheless, the number of non-paying customers is lower to 374.

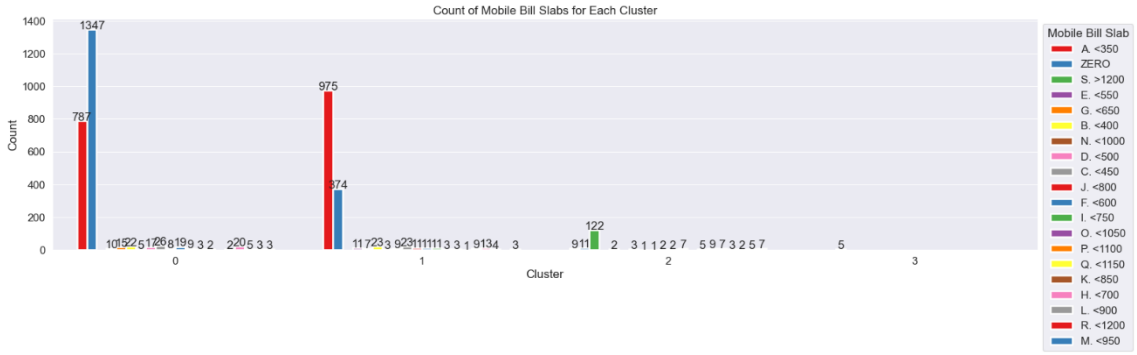

Figure 27- Mobile Bill Slabs

4.4. Cluster Profile Summary

Below table shows the summary of the clusters identified by K-means clustering. The scale with Low, Mid, High, and Extreme values are used for this purpose with following definitions:

1. Low indicates that the values analyzed by the model are below average (Value < 0.0)
2. Mid indicates that the values are on average (Value = 0.0)
3. High implies that the values are above average and below 2.0 (0.0 < Value < 2.0)
4. Extreme indicates that the values are greater than 2.0 categorizing them as extreme values (Value > 2.0)

Table 6 - Cluster Profile Summary

Cluster	**1**	**2**	**3**	**4**
Bundled	Yes	No	Yes	No
Home/Business	Both	Home	Business	Business
Total Bill	Mid	Low	High	Extreme
Variable Mobile Bill	Low	Low	High	Extreme
Variable On-net Bill	Low	Mid	High	Mid
Variable Off-net Bill	Low	Low	High	High

Mobile Minutes	Low	Low	High	Extreme
On-net Minutes	Mid	Low	High	Mid
Off-net Minutes	Low	Low	High	High
Complaints	Mid	Mid	High	Low
Total Customers	2,303	1,494	198	5
Segment Category	**Fixed Mid Value**	**Low Value**	**High Value**	**Extremer**

4.5. Manual Campaign Results

Campaigns were designed according to the cluster segment to evaluate the effectiveness of the model. We chose cluster 1, 2 and 3 for testing of the model as cluster 4 has only a few extreme values not suitable for campaigns. Our campaigns were designed into four different categories:

1. **Awareness Campaigns:** For bundled customers with minutes package as part of their voice bundle.
2. **Low Value Mobile:** Customers with low bills and low mobile usage to test if they will subscribe to a low value offer.
3. **High Value Mobile:** For Bundled customers with high bill and high mobile usage.
4. **Hybrid:** Multiple offerings for customers with no bundle however paying bill amount in the same range.

Our sample consisted of 4,000 unique telecom customers, however only 3,457 customers were active at the time of campaigns execution, leaving the active base at 86%. The campaigns were executed on 3,030 customers from these 3,457 active customers (88%), this is since 427 customers had more than one bundled subscribed, making them ineligible for the campaigns.

Since cluster 1 has customers who are bundled and paying a decent amount of bill each month, we pitched them mobile minutes. Awareness campaigns were executed on customers who have already subscribed to a bundle however are not utilizing monthly minutes they are being given. Cluster 2 was offered hybrid offers with multiple minutes resources to encourage them to make voice usage. Cluster 3 was offered high value bundles according to their billing.

Overall, the campaigns were effective and produced better results than the existing model. Cluster 1 secured 23.7% with 382 people opting for offered bundle or becoming a user, Cluster 2 secured 5.6% with 74 customers opting for promoted offer whereas Cluster 3 response rate was only 5% due to expensive offering and small data set.

Table 7 - Cluster Profile and Results

Cluster	**1**	**2**	**3**
Segment Category	Fixed Mid Value	Low Value	High Value
Campaigns	Awareness, High Value Mobile, Low Value Mobile	Hybrid, High Value Mobile, Low Value Mobile	Hybrid
Total Customers	2,303	1,494	198
Active Customers	1957	1319	177
Targeted Customers	1611	1319	100
Results	**382**	**74**	**5**
Results (%)	**23.7%**	**5.6%**	**5.0%**

The following table shows the target sample for each cluster. As we can see that High value target sample is highest in cluster 1 (890) as they are already bundled and have the tendency to opt for mobile addons. Awareness campaigns have the second highest sample of 604 customers. The aim was to convert bundled non-users (with 0 minutes usage) to users through making them aware of the bundle resources they are not

utilizing, however paying for on monthly basis. This campaign was highly successful and will be replicated for all bundled non-users. For cluster 2, the Hybrid campaign had the highest target sample of 1000 customers, to convert non-bundled customers into bundled, followed by a Low Value Mobile campaign of 198 customers. Sample for cluster 3 was the lowest due to a smaller number of customers in the cluster.

Table 8 - Target Sample for each Cluster

Target	1	2	3
Awareness	604	-	-
Low Value Mobile	117	198	-
High Value Mobile	890	121	-
Hybrid	-	1000	100
Total	**1611**	**1319**	**100**

We can see in the cluster wise results that awareness campaign was the most effective among all the campaigns, followed by Hybrid campaign in cluster 2 of non-bundled users. Low value mobile campaign was not much successful as expected due to small dataset and non-bundled/non- usage customers.

Table 9 - Results of each Cluster

Results	1	2	3
Awareness (Non-User to User)	343 (56.8%)	-	-
Low Value Mobile	-	4 (2.0%)	-
High Value Mobile	39 (4.4%)	6 (5.0%)	-
Hybrid	-	64 (6.4%)	5 (5.0%)
Total	**382**	**74**	**5**

The following table shows the results campaign-wise. Awareness campaigns were the most effective as stated above in converting non-user to user with 56.8% response rate, followed by Hybrid and High Value Mobile campaigns which were 6.3% and 4.5% successful respectively. The Low Value Mobile campaign was the least successful, possible reasons are low usage/ non-bundled users and small dataset. In total, the campaigns were 15.2% effective. Excluding the awareness campaigns, the response rate is 4.9%, which is almost double the existing campaigns being executed currently. Thus, the model can be deployed in a real environment for real application and results.

Table 10 - Campaign wise results

Campaign Category	Target	Results	%
Awareness (Non- User to User)	604	343	**56.8%**
Low Value Mobile	315	4	**1.3%**
High Value Mobile	1,011	45	**4.5%**
Hybrid	1,100	69	**6.3%**
Total	**3,030**	**461**	**15.2%**
Excluding Awareness	**2,426**	**118**	**4.9%**

4.6. Classification and Regression Model

The aim is to predict if a non-bundled customer would subscribe to one or more bundles. For this purpose, we applied classification and regression models to train our model on actual bundled customers data in real business environment and compare which model will perform best on telecom data. This is to enable the model to learn the behavioral patterns of the real time bundled customers and utilize and apply this knowledge when presented with real time non-bundled customers data to predict product propensity for each bundle.

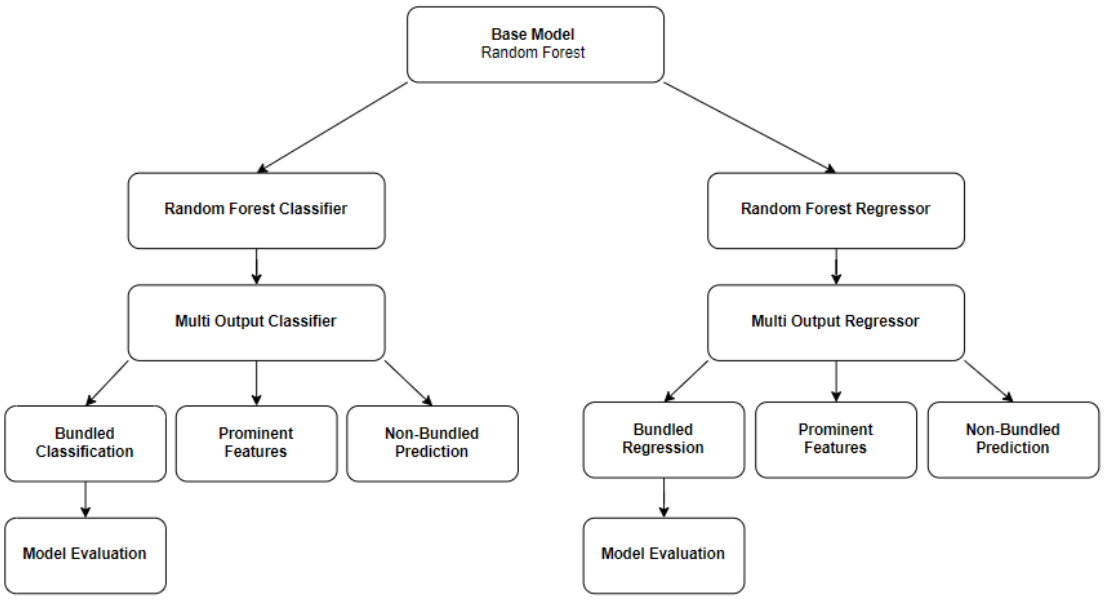

Figure 28- Base Model

4.6.1. Random Forest Classification

We first selected Random Forest as the base model. Then we selected the data of real time bundled customers and trained the model integrated with Multi Output Classifier. These classifiers hold the capability to predict multiple target variables simultaneously. Since customers can subscribe to more than one bundle at a time, it is crucial to forecast this possibility. The model was evaluated with performance metrics such as accuracy, precision, recall, F1-score, Bias, and Variance. Performance metrics are measures that allow us to evaluate the effectiveness of different machine learning classification algorithms:

Accuracy: The number of predictions correctly predicted by the model, out of all the predictions made by the model. It is the sum of true positive (TP) and true negative (TN) divided by the total predictions.

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

Precision: The total number of predictions that were true positive (positive, predicted positive), divided by the overall number of positive predictions.

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

Recall: The ratio of Positive samples that were correctly identified as Positive to all Positive samples. The recall measures how well the model can identify Positive samples. The more positive samples that are identified, the higher the recall.

$$\frac{TP}{TP + FN}$$

F1 Score: A machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model.

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

Bias: A metric to measure the systematic error in predictions due to incorrect assumptions in the model learning process. Higher values of bias suggest that model is underfitting, which implies that the model is unable to capture the hidden patterns in the training data. The lower the bias, the better the performance of the model and its ability to learn the underlying patterns.

Variance: Measures the variability of predictions made by model across different sets of datasets based on model learning. High variance implies that the model is overfitting the data and tends to capture noise in the data. The lower the variance, the better the performance and its ability to make predictions without being influenced by noise in the data.

In an ideal scenario, a model should have lower bias and lower variance. This implies that the model can learn the underlying patterns and make accurate predictions

and shows consistent performance on different datasets. It is essential to maintain the balance between the two metrics.

Bundled Classification Results

The model performed significantly well in terms of testing accuracy, bias, and variance. The model was trained on bundled customers and has a testing accuracy of 0.92. The model has a lower variance of 0.06 and bias of 0.24, considered satisfactory for this type of data.

Table 11 - Classification Model Evaluation

Metric	Values
Training Accuracy	0.99
Testing Accuracy	0.92
Bias	0.24
Variance	0.06

Since we used Multi-Output Classifier which predicts outcome of multiple labels simultaneously, we calculated precision, recall and F1-score of each bundle/label. A total of 10 labels were chosen for this exercise. As it is evident in the results below, the F1-score of Bundle 1 and Bundle 6 is the highest (0.95) implying that the model is making 95% correct predictions. Bundles 4 and 7 have lower F1- score as compared to others indicating improvements in the model.

Table 12 - Multi-Label Evaluation

	Precision	Recall	F1-Score	Support
Bundle 1	0.96	0.95	0.95	23,311
Bundle 2	0.89	0.76	0.82	4,089
Bundle 3	0.91	0.91	0.91	264
Bundle 4	0.93	0.71	0.80	3,340
Bundle 5	0.89	0.87	0.88	4,699
Bundle 6	0.97	0.93	0.95	2,999

Bundle 7	0.86	0.76	0.81	1,761
Bundle 8	0.93	0.84	0.88	5,293
Bundle 9	-	-	-	-
Bundle 10	0.95	0.91	0.93	4,499
micro avg	0.94	0.89	0.91	50,255
macro avg	0.83	0.76	0.79	50,255
weighted avg	0.94	0.89	0.91	50,255
samples avg	0.51	0.50	0.50	50,255

4.6.2. Random Forest Regression

In addition to the classification model, we trained a Random Forest Regression Model on the same bundled customers dataset. As a regression model, we assessed its performance using the following evaluation metrics:

Mean Absolute Error (MAE):

MAE is the average of absolute differences between the predicted values and the true values.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} | \, y_i - \hat{y_i} \, |$$

Where:

- n is the number of observations in the dataset.
- y_i is the true value of the outcome for the i-th sample.
- $\hat{y_i}$ is the predicted value of the outcome variable for the i-th sample.

Mean Squared Error (MSE):

MSE is the average squared difference between the predicted values and the true values. It can penalize larger values of error due to the involvement of square function. The lower the values of MSE, the better the performance of the model.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

Table 13 - MAE vs MSE

MAE	MSE
Measures average absolute difference	Measures average squared difference
Less sensitive to outliers	More sensitive to outliers
Does not penalize large errors	Penalizes large errors

The model's MAE of 0.07 and MSE of 0.17 indicate satisfactory performance, as these metrics are closer to zero. Their proximity to zero suggests that the model's predictions are relatively close to the actual values. Overall, the model demonstrates favorable accuracy, given the low values of these evaluation metrics.

Table 14 - Regression Results

Metric	Values
Mean Absolute Error (MAE)	0.07
Mean Squared Error (MSE)	0.17

4.6.3. Prominent Features

The models were trained on 137 features, and it is essential to find out the top 10 features that were prominent in model learning and prediction. Below table compares the features that were the most helpful in making predictions on unseen data. Even though the importance values are lower, they give us a rough idea about models' learnings. AVG_BUNDLE_BILL is the most important feature with the importance of 0.15 and

0.51 in classifier and regressor respectively. TOTAL_BILL and AVG_MOB_MINS are also included in the top 10 features, though their significance is much lower.

Classification Model	
Feature	**Importance**
AVG_BUNDLE_BILL	0.15
AVG_TOTAL_BILL	0.08
BUNDLE_BILL01	0.07
BUNDLE_BILL02	0.07
BUNDLE_BILL03	0.07
TOTAL_BILL01	0.06
TOTAL_BILL02	0.05
TOTAL_BILL03	0.05
AVG_MOB_MINS	0.03
NATIONWIDE_MOBILE _BILL01	0.02

Regression Model	
Feature	**Importance**
AVG_BUNDLE_BILL	0.51
BUNDLE_BILL01	0.10
TOTAL_BILL03	0.08
AVG_LINE_RENTAL	0.05
BUNDLE_BILL03	0.04
BUNDLE_BILL02	0.02
AVG_MOB_MINS	0.02
TOTAL_BILL02	0.02
AVG_TOTAL_BILL	0.02
MOBILE_MINS03	0.01

4.7. Multi-Offer Recommendation Model

We implemented Multi-Offer Recommendation Model with both Classifier and Regressor Algorithm. These models were first trained on bundled customers as explained in the above section. The models were then used to make predictions on non-bundled customers to predict the probability if the customers will subscribe to one or more bundles based on their underlying billing and usage patterns. Comparing the performance on training data, both models exhibited satisfactory performance which can be validated through marketing campaigns.

The models were fed with data of 100k non-bundled customers for offer model predictions, 50K for classification model and 50k for regression model. Looking at the results, regressor performed better in terms of bundle prediction on unseen non-bundled customers data as it made 87% distinct predictions on 50k dataset. On the other hand,

Classifier was able to predict a bundle for only 9% of the target data which is significantly lower than regression model.

Table 15 - Multi-Offer Recommendation Model Prediction Results

	Classifier	Regressor
Total Data for Prediction	50,000	50,000
Distinct Prediction	4,663	43,687
Predicted (0)	45,337	6,313

Following the predictions, data was prepared for campaigns execution. Additional filters were implemented to ensure that the campaign data is thoroughly cleaned i.e., no duplicates, invalid data, so that it does not affect the execution of campaigns. Among the 4.6k customers predicted by the classifier, only 2k were chosen for campaigns. 2k out of 12k customers were chosen for regressor based campaign.

Table 16 - Multi-Offer Recommendation Model - Campaign Results

	Classifier	Regressor
Final Campaigns Target	2,000	2,000
Response	3	47
Response rate (%)	0.2%	2.4%

Overall, the campaign derived from regression model performed well due to a higher response rate 2.4% than classifier model with 0.2% response rate. This implies random forest regression model is effective in this type of postpaid telecom customers data.

Research questions

RQ 1: Which clustering algorithm performs best on this specific set of telecom data for customer segmentation?

In the context of our telecom dataset, which exhibits a unique nature of postpaid segment with a small scale, K-means clustering outperforms the DBSCAN algorithm in terms of essentially identifying distinct clusters aligned with domain knowledge. The challenge with DBSCAN lies in its difficulty to differentiate between cluster points and outliers, making it less suitable for this specific dataset. However, it is worth considering that DBSCAN might exhibit better performance on larger datasets, where its characteristics could potentially be more beneficial.

RQ2: Which performs more effectively in Multi-Offer Recommendation Model– Classifier or Regressor?

We introduced a Multi-Offer Recommendation Model using both Random Forest Classifier and Regressor to forecast the likelihood of customers subscribing to one or more bundles. Each algorithm exhibited distinct predictions for individual bundles. To assess their effectiveness, campaigns were conducted on a dataset of ~4k real-time customers in a live environment. The dataset was randomly split into two sets for each campaign – one for the classifier and another for the regression. Our findings reveal that regression model outperforms classification model by attracting a higher number of opt-ins. This indicates that regression model proves more effective for the Multi-Offer Recommendation Model and will generate convincing results on higher target set.

RQ 3: Which features played the most crucial role in predicting product propensity in the Multi-Offer Recommendation Model?

Out of the 137 features that the models were trained on, AVG_BUNDLE_BILL is the most important feature with importance of 0.15 and 0.51 in classifier and regressor respectively. This implies that the fixed amount that the customer must pay monthly plays a significant role in the decision to subscribe to a bundle. TOTAL_BILL and AVG_MOB_MINS are among the top 10 features; however, their significance is considerably lower compared to other features.

RQ 4: Do the proposed model and strategies work effectively in real time business environment?

The manual campaigns we executed performed better than existing manual campaigns in the real time environment. We managed to convert 56.8% of non-callers into callers by running awareness campaigns. This is significantly important for voice penetration. Excluding awareness campaigns, we managed to achieve a 4.9% response rate, which is almost double that of existing manual campaigns. As for the Multi-Offer Recommendation Model Campaigns, regression model performed better than classification model in terms of opt-ins.

CHAPTER 5

CONCLUSION

5.1. Conclusion

This study focused on Customer Segmentation and Multi-Offer Recommendation model. The Customer Segmentation model helps to understand the customers and their behavioral patterns whereas Multi-Offer Recommendation model learns the behaviors of bundled customers and predicts if non-bundled customers will buy one or more bundles.

We understood customer segments through comparative analysis of DBCAN and K-means clustering. While DBSCAN performs better in identification of noise points using density-based clustering, it did not perform well clustering analysis. K-means on the other hand outperformed DBSCAN with clear identification of heterogeneous clusters leading to meaningful post cluster analysis. It is pertinent to note that this analysis is based on a small data set and therefore is limited to interpretation.

For the Multi-Offer Recommendation model, we chose a random forest classifier and random forest regressor with Multi -Output classifier and multi-output regressor to find the effectiveness of these models. Regressor model performed significantly well than classification model in terms of package opt-ins hence is more suitable on this set of data.

5.2. Future work

Since our segmentation analysis is based on a small data set of 4000 customers, the results of the segmentation can be further enhanced by using a large dataset. Moreover, this study can be advanced by automating the whole process of data collection, data preparation, model deployment and post cluster analysis in a real business

environment. The model deployed in real time business environment can benefit the telecom company in terms of improved businesses processes, optimized CVM campaigns and quick campaign results. The Multi-Offer Recommendation model can be trained on large dataset and pilot campaigns to:

1. Prove the effectiveness of model on huge dataset.
2. Minimize the effort of manual campaigns and improve CVM processes.

REFERENCES

[1] N. Singh, P. Singh, K. K. Singh, and A. Singh, ‘Machine learning based classification and segmentation techniques for CRM: a customer analytics’, *International Journal of Business Forecasting and Marketing Intelligence*, vol. 6, no. 2, pp. 99–117, Jan. 2020, doi: 10.1504/IJBFMI.2020.109878.

[2] O. Dogan, E. Ayçin, and Z. A. Bulut, ‘Customer segmentation by using RFM model and clustering methods: a case study in retail industry’, *International Journal of Contemporary Economics and Administrative Sciences*, vol. 8, no. 1, pp. 1–19, 2018.

[3] B. Cooil, L. Aksoy, and T. L. Keiningham, ‘Approaches to Customer Segmentation’, *Journal of Relationship Marketing*, vol. 6, no. 3–4, pp. 9–39, Jan. 2008, doi: 10.1300/J366v06n03_02.

[4] S. Masood, M. Ali, F. Arshad, A. M. Qamar, A. Kamal, and A. Rehman, ‘Customer segmentation and analysis of a mobile telecommunication company of Pakistan using two phase clustering algorithm’, in *Eighth International Conference on Digital Information Management (ICDIM 2013)*, 2013, pp. 137–142. doi: 10.1109/ICDIM.2013.6693978.

[5] M. Tavakoli, M. Molavi, V. Masoumi, M. Mobini, S. Etemad, and R. Rahmani, ‘Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining Techniques: A Case Study’, in *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*, 2018, pp. 119–126. doi: 10.1109/ICEBE.2018.00027.

[6] J. Wu *et al.*, ‘An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and *K*-Means Algorithm’, *Math Probl Eng*, vol. 2020, p. 8884227, 2020, doi: 10.1155/2020/8884227.

[7] T. Zhang, ‘Telecom customer segmentation and precise package design by using data mining (Dissertação de mestrado, Iscte-Instituto Universitário de Lisboa). Repositório do Iscte 2018’.

[8] S. M. H. Jansen, ‘Customer segmentation and customer profiling for a mobile telecommunications company based on usage behavior’, *A Vodafone case study*, vol. 66, 2007.

[9] M. Namvar, M. R. Gholamian, and S. KhakAbi, ‘A Two Phase Clustering Method for Intelligent Customer Segmentation’, in *2010 International Conference on Intelligent Systems, Modelling and Simulation*, 2010, pp. 215–219. doi: 10.1109/ISMS.2010.48.

[10] A. Wali and R. S. Sunitha, ‘Churn analysis and plan recommendation for telecom operators’, *Journal for Research*, vol. 2, no. 3, 2016.

[11] X. Xia and G. Zhao, ‘Telecom package recommendation model based on convolutional neural network’, in *Second International Symposium on Computer Applications and Information Systems (ISCAIS 2023)*, SPIE, 2023, pp. 73–79.

[12] Q. Lin and Y. Wan, ‘Mobile customer clustering based on call detail records for marketing campaigns’, in *2009 International Conference on Management and Service Science*, IEEE, 2009, pp. 1–4.

[13] C.-H. Cheng and Y.-S. Chen, ‘Classifying the segmentation of customer value via RFM model and RS theory’, *Expert Syst Appl*, vol. 36, no. 3, pp. 4176–4184, 2009.

[14] J. A. McCarty and M. Hastak, ‘Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression’, *J Bus Res*, vol. 60, no. 6, pp. 656–662, 2007.

[15] M. Namvar, M. R. Gholamian, and S. KhakAbi, ‘A two phase clustering method for intelligent customer segmentation’, in *2010 International conference on intelligent systems, modelling and simulation*, IEEE, 2010, pp. 215–219.

[16] D. AL-Najjar, N. Al-Rousan, and H. AL-Najjar, ‘Machine learning to develop credit card customer churn prediction’, *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 4, pp. 1529–1542, 2022.

[17] D. Wadikar, ‘Customer Churn Prediction’, 2020.

[18] S. Wu, W.-C. Yau, T.-S. Ong, and S.-C. Chong, ‘Integrated churn prediction and customer segmentation framework for telco business’, *IEEE Access*, vol. 9, pp. 62118–62136, 2021.

[19] A. K. Ahmad, A. Jafar, and K. Aljoumaa, ‘Customer churn prediction in telecom using machine learning in big data platform’, *J Big Data*, vol. 6, no. 1, pp. 1–24, 2019.

[20] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, ‘Customer churn prediction using improved balanced random forests’, *Expert Syst Appl*, vol. 36, no. 3, pp. 5445–5449, 2009.

[21] A. AGRAWAL and Y. CHOUHAN, ‘Propensity to Buy Model for Fixed Line Telecom Customers’, 2015.

[22] O. Çelik and U. O. Osmanoglu, ‘Comparing to techniques used in customer churn analysis’, *Journal of Multidisciplinary Developments*, vol. 4, no. 1, pp. 30–38, 2019.

[23] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, 'A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector', *IEEE access*, vol. 7, pp. 60134–60149, 2019.

[24] A. Saran Kumar and D. Chandrakala, 'A survey on customer churn prediction using machine learning techniques', *Int J Comput Appl*, vol. 975, p. 8887, 2016.

[25] N. Lu, H. Lin, J. Lu, and G. Zhang, 'A customer churn prediction model in telecom industry using boosting', *IEEE Trans Industr Inform*, vol. 10, no. 2, pp. 1659–1665, 2012.

[26] H. Lee, Y. Lee, H. Cho, K. Im, and Y. S. Kim, 'Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model', *Decis Support Syst*, vol. 52, no. 1, pp. 207–216, 2011.

[27] T. J. Gerpott, W. Rams, and A. Schindler, 'Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market', *Telecomm Policy*, vol. 25, no. 4, pp. 249–269, 2001.

[28] M.-K. Kim, M.-C. Park, and D.-H. Jeong, 'The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services', *Telecomm Policy*, vol. 28, no. 2, pp. 145–159, 2004.

[29] M. Alkhayrat, M. Aljnidi, and K. Aljoumaa, 'A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA', *J Big Data*, vol. 7, no. 1, p. 9, 2020, doi: 10.1186/s40537-020-0286-0.

[30] A. Yaseen, 'Next-wave of E-commerce: Mobile customers churn prediction using machine learning', *Lahore Garrison University Research Journal of*

Computer Science and Information Technology, vol. 5, no. 2, pp. 62–72, 2021.