Machine Learning based Multi-Variate MBB-User Growth Prediction and Worst-Cell Clustering in Cellular network

by

Tanjir Ahmed
20166032

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
BRAC University
February 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Tanjir Ahmed
Student ID:20166032

Approval

The thesis/project titled "Machine Learning based Multi-Variate MBB-User Growth Prediction and Worst-Cell Clustering in Cellular Network" submitted by

1. Tanjir ahmed (Student ID)

of Spring, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on April 09, 2024.

Examining Committee:

Supervisor:
(Member)

Mohammad Shamsul Arefin, Ph.D
Professor
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology

Program Coordinator:
(Member)

Md Khalilur Rhaman, Ph.D.
Associate Professor
Department of Computer Science and Engineering
Brac University

Supervisor:
(Member)

Md. Golam Robiul Alam, Ph.D.
Associate Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Amitabha Chakrabarty, Ph.D, Ph.D.
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, Ph.D
Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The thesis will analyze mean user number of cellular networks based on maximum user, data volume and cell availability from a KPI report generated from OSS server. Firstly, we will collect KPI data from OSS server of specific area where large number of cellular users taking service from network and then conduct data cleansing and data splitting. Finally, we will analyze the data with multiple variable regression and support vector machine (SVM) analysis to predict user number. We will also perform K-means clustering method to find the worst cells to improve the network performance and user experience. We will do the analysis with most popular and latest python which boasts the feature of pure object-oriented programming, platform independent, concise and very elegant language. So, we will call the corresponding library function to predict the user number and worst cells clustering which will help a telecom network operator to plan, design an effective and optimized network and also help to improve user experience as well. Moreover, this analysis will help to make cluster plan, and understand user behavior in a specific area. Keywords: —K-means, UE, data volume, cellular, network, prediction, clustering, regression

Keywords: Data Mining; Machine Learning; Cricket; T20; Prediction; Decision tree; Linear Regression Analysis

Acknowledgement

Firstly, all praise to the Great Allah for whom my thesis has been completed without any major interruption. Secondly, to my advisor Md. Golam Robiul Alam sir for his kind support and advice in my work. He helped me whenever I needed help. And finally, to my family without their throughout support it may not be possible. With their kind support and prayer, I am now on the verge of my graduation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Mobile networks have become an essential part of our lives, playing a crucial role in various aspects of our personal and professional spheres. They can provide support in medical sector, distance learning, information and access, navigation and location services and robustness when serving the end-user is the prerequisite. A mobile network is self-organizing, consisting of multiple embedded systems and those that support all the services underlying. For better network performance and coverage for end users, operators need to plan and optimize network capacity like voice traffic, data volume, throughput etc. Machine learning algorithms can help to plan and optimize mobile network efficiently

1.2 Research Problem

In a mobile network, there are two types of network traffic, one is voice traffic and another is data traffic. The importance of proper planning for data and voice traffic in mobile networks cannot be overstated. It's crucial for ensuring a seamless and reliable experience for users, optimizing network efficiency and cost, and even boosting network resilience. Proper and efficient network traffic prediction can ensure quality of service and network efficiency. Several types of research are ongoing based on MBB-User Growth Prediction and Worst-Cell Clustering and the types of machine learning are supervised machine learning, unsupervised machine learning, and semi-supervised learning

1.3 Research Objectives

With the increasing of growth of internet, mobile data usage increased a lot. If we see the statistics then we can understand how mobile data growth has increased than past years.4.88 billion peoples are now connected with internet in October 2021 that is almost 62In this work, we have taken a novel approach to predict MBB user growth and worst cell clustering in cellular network by machine learning based multivariate model. We have used linear regression model and support vector machine algorithm to predict user growth. And K-means and K-medoids to identify network worst cell.

For the both approaches, we have used cell availability, throughput and data volume as variables. The main contributions of the paper are summarized as follows:
1. We have introduced a unique contribution to predict user growth using linear regression and support vector algorithm along with worst cell clustering using K-means and k-medoids clustering approach in machine learning based multi-variant environment.
2. The linear regression model predicted the mean user almost 90 percent accurately and the support vector algorithm predicted 91 percent of accuracy. For further analysis, OLS (ordinary least method) is applied, which shows R-squared value 0.957, which is a acceptable output.
3. We have applied the Elbow method to find the number of clusters. From K-means we found 3 clusters, Cluster 0 has 2.77 UE count, Cluster 1 has 12.475 UE count and Cluster 2 has 7.39 UE count. And Cluster o is detected as worst cell. For proper interpretation we have also used LIME.
4. Finally, the Thesis shows the worst performing cell of a cellular network successfully with three different clusters. It is completely unique approach in the field of telecommunication.

1.4 Research Methodology

This Thesis is set to produce to [predict user growth using linear regression and support vector algorithm along with worst cell clustering using K-means and k-medoids clustering approach in machine learning based multi-variant environment. From the vast area of supervised and unsupervised machine learning few have been chosen and a comparative approach related to supervised and unsupervised learning has been proposed. The performance data feed from a real network. Main objective in this paper is to predict mobile user numbers in a cellular network by multiple variable linear regression model and worst cell clustering by K-means clustering algorithm. From our KPI data Max HSDPA UE, HSDPA Data Volume MB (MB) and Cell availability are the independent variables where Mean HSDPA UE (None) is only the dependent variable in our model. We will predict this Mean HSDPA UE based on the independent variables

Figure 1.1: Overall Methodology

1.5 Scope and Limitation

This thesis aims to predict mobile user number in UMTS network only. But in mobile network there are other technology like LTE, GSM which are not explored in this thesis. With increased number of Nodes in the network computational resource for model buildup need to increase.

1.6 Scope and Limitation

In chapter 2, some related work has been discussed. In chapter 3 has the details of mobile network traffic and capacity understanding such as Max mean HSDPA UE, Cell availability, HSDPA Meanchthroughput, HSDPA data volume- User Growth Prediction from Multi-variate Liner Regression Model is discussed in details in chapter 4. Worst Cell Clustering in Cellular Network from K-means algorithm details has been discussed in chapter 5. Research gap, algorithm understanding, methodology, data preparation and model implementation discussed in details in chapter 4 and 5.

Chapter 2

Related Work

2.1 Introduction

In this era, mobile network playing a very vital role towards digital transformation. There are many new technologies evolving and expanding in mobile network, so the challenges related with network planning and optimization is increasing day by day. Optimal coverage and capacity, enhanced network performance, cost effectiveness, Future proofing network, improved customer satisfaction is essential now to maintain for every mobile operator. mobile network planning is the backbone of a thriving mobile ecosystem. It ensures a positive user experience, efficient network operation, and paves the way for future advancements. It's the silent hero behind every seamless call, blazing-fast download, and lag-free online game. Machine learning can play a vital role in mobile network traffic growth and optimization. This chapter provides various challenges faced in traditional mobile network planning and explores various machine learning method for mobile network traffic growth prediction and clustering.

2.2 Multi linear regression model for mobile location estimation in GSM network

The authors in [1], tried to establish a regression model to improve the accuracy of location estimation. This paper introduced the first application of Multiple liner regression model for mobile location estimation in a GSM network without manipulating LDP (location dependent parameter), RSSI (Received signal strength indicator). The authors collected this data from drive test and was successful to show 67 percent of the calls, the positioning error is less than 64 m, and almost 95 of the calls in positioning error less than 115m and maximum error is 275m in urban area. The objective of this paper is Develop and evaluate a multi-linear regression (MLR) model for estimating the location of mobile subscribers in a GSM network based on received signal strength (RSS) measurements from surrounding cell towers. the MLR model used RSS values from three neighboring cell towers as independent variables to predict the mobile subscriber's location (coordinates) as the dependent variable. Data was collected from a real GSM network in Enugu, Nigeria, including RSS measurements and cell tower locations. The model was trained and tested using different training-testing data split ratios. Performance was evaluated based on: Mean squared error (MSE): Measuring the average squared difference between

predicted and actual locations. Cumulative distribution function (CDF): Showing the percentage of estimated locations within certain error radii

2.3 A regression approach to infer electricity consumption of legacy telecom equipment

The authors in [2], tried to introduce a method of statistically inferring electricity consumption of different kinds of device which is installed base of telecommunication equipment. After applying this in to a 3918 central offices of a major telecommunication provider, their analysis reveals that network-wide energy consumption of each major type of devices. In particular, they found that electricity consumption mainly dominated by Class-5 telephone switches, which carrying for 43 percent of aggregate total consumption. In order to manage efficiently electricity consumption, it is very important to understand where the electricity is mainly used. The authors used two regression analysis to partition the electricity consumption by equipment group. Their analysis has a high level of accuracy, estimated that they have identified the contribution of each equipment group to within two percentage. in order to significantly reduce total central office electricity consumption, they suggested to prioritizing efficiency measures and technology evolution for TDM switches

2.4 Regression in Wireless Sensor Networks

The authors in [3], the main purpose of this paper is to clustering of wireless nodes by multi-linear regression originates by combing the ideas of locating the nodes by regression and how to utilize nodes parameters in multi-linear regression formula. A wireless network (WSN) mainly contains lots of sensors which are capable of computing network data, sensing data and for communication with other devices or sensors. Mainly a sensor contains power unit, sensing unit, computational unit with a memory and transmitting unit. The basic function of these sensors is to collect signal from surrounding device's temperature, pressure, humidity, moving objects and share these observations with base stations. In this paper, the authors introduced linear regression model into WSN and showed how to use node values in multiple linear regression.

2.5 Hybrid prediction model for mobile data traffic: A cluster-level approach

The author in paper [4] proposes a hybrid prediction model for mobile data traffic based on a cluster-level approach. This approach aims to improve the accuracy of mobile data traffic prediction by considering the diverse traffic patterns observed in different geographical areas. The paper demonstrates that the proposed hybrid prediction model based on a cluster-level approach can be a promising approach for improving the accuracy of mobile data traffic prediction. This can be beneficial for network operators in terms of resource allocation, network planning, and service quality management. In this paper, author discussed, Mobile data traffic is rapidly growing and highly dynamic, making accurate prediction crucial for

network providers. Proposed solutions are Proposed Solution: K-means clustering: Groups mobile network base stations with similar traffic patterns into clusters. Long Short-Term Memory (LSTM) networks: A type of artificial neural network adept at handling time-series data like traffic patterns. The hybrid model demonstrates significantly improved accuracy compared to baseline models, with 10-15

2.6 A linear regression model for the analysis of life times.

The author in paper [5] proposes a linear regression model for analyzing life times (e.g., time to death in a medical study). The aim is to overcome limitations of the Cox proportional hazards model, particularly its vulnerability to inconsistencies when covariates change. Aalen's linear regression model presents a valuable alternative for analyzing life times, offering flexibility, robustness, and the ability to capture time-varying covariate effects. This paper, published in Statistics in Medicine (1989), proposes a novel linear regression model for analyzing survival data, offering an alternative to the popular Cox proportional hazards model. Traditional survival analysis methods, like the Cox model, focus on analyzing the hazard function (instantaneous risk of death at a given time point). However, they can be sensitive to model assumptions and prone to inconsistencies when covariates are omitted or redefined. Proposed Solutions are Aalen introduces a linear regression model directly focusing on the cumulative hazard function, the probability of an event (death) occurring by a certain time point. This approach has several advantages, less vulnerable to model assumptions: The model doesn't rely on the proportionality of hazards across covariate levels, making it more robust to model violations, Less sensitive to changes in covariate measurement: Redefining or omitting covariates has less impact on the analysis compared to the Cox model, allows for non-parametric estimation of regression functions: This enables visualizing how the effect of covariates on survival changes over time, offering richer insights.

Chapter 3

Mobile Network traffic and Capacity Understanding

3.1 Mobile Network Traffic and Capacity

Mobile network traffic refers to the data that flows through mobile networks, while capacity defines the ability of those networks to handle that data flow. Understanding both is crucial for ensuring quality mobile experiences for users and efficient network management for operators. Mobile network traffic includes all data sent and received through mobile devices, like smartphones and tablets. Traffic types include voice calls, video streaming, web browsing, social media usage, gaming, and app downloads.
Mobile Network Capacity This measures the maximum amount of data a network can handle at a given time, often measured in megabits per second (Mbps) or gigabits per second (Gbps). Capacity depends on various factors like network infrastructure, spectrum availability, cell tower density, and technology limitations. Inadequate capacity leads to network congestion, resulting in slow speeds, buffering, dropped calls, and poor user experience.

3.2 Site and cell concept in Mobile network

site refers in telecom network to the actual structure (towers). Housing the equipment required to power and operate cells. Each cell site hosts one or more base stations that emit radio signals, creating a specific cell, an area where users can connect to the network. Cell sites house transmitters, receivers, antennas, power units, and other related electronics for signal amplification, routing, and network connectivity. Cells is Unlike a physical structure, a cell is a defined zone within the network coverage, often represented as a hexagon or circle on a map. Within a cell, users with compatible devices can connect to the network base station for calls, data transfers, and other services.

3.3 Max and Mean HSDPA UE

Max and Mean HSDPA UE means maximum and average number of users taking traffic or service in a certain telecom zone under a node (NodeB). This telecom

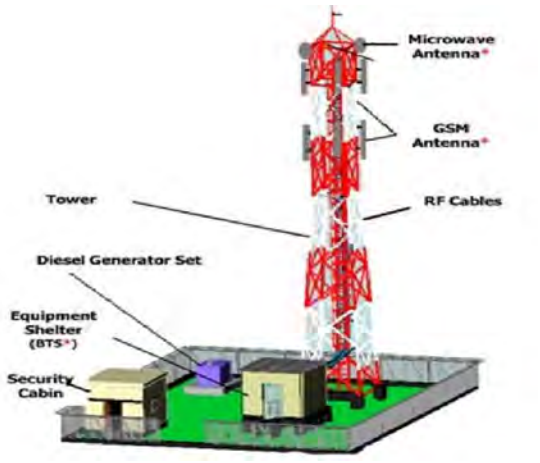

Figure 3.1: A telecom physical site

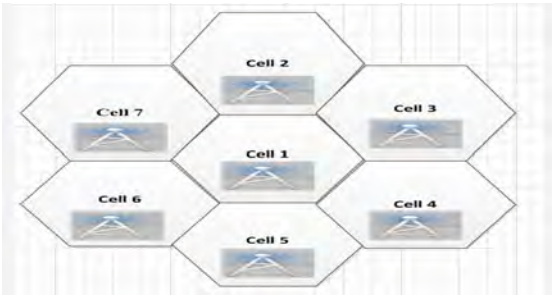

Figure 3.2: Cell in telecom network

zone can be a site or cell. Nodeb is a telecommunication node which is serving in a telecommunication network where users are using its resource for voice and data service. HSDPA (High-Speed Downlink Packet Access) is a technology used in 3G networks to significantly increase data speeds and capacity, particularly for downloads. It's essentially an enhancement built upon the existing 3G protocol, often referred to as 3.5G or 3G+. HSDPA technology Utilizes advanced coding techniques to pack more data into each radio signal. Employs shared channels more effectively, dynamically allocating resources based on demand. Focuses on download performance, with a separate technology called HSUPA (High-Speed Uplink Packet Access) handling uploads.

3.4 Cell Availability

Cell Availability is the success rate of a radio access network availability in specific zone or cluster. It helps the network optimization team to find various issue in telecom network during network access and roaming. It refers to the proportion of time a specific cell is accessible and functional for users to connect to and use. It essentially measures the reliability and uptime of a cell in providing service. High availability ensures reliable connectivity for users, which is crucial for seamless voice calls, uninterrupted data usage, and emergency services access. It translates to customer satisfaction and network operator trust. Low availability can lead to dropped calls, slow data speeds, and frustration for users, potentially impacting revenue and reputation. Overall, cell availability is a critical metric for measuring the perfor-

mance and quality of a cellular network. By monitoring and improving availability, network operators can ensure a positive user experience and maintain a competitive edge.

3.5 HSDPA Meanchthroughput

Throughput is the actual amount of data successfully sent or received between a link or between a user and a mobile tower. It is normally measured as kbps, mbps, gbps. throughput refers to the actual rate of data transfer achieved by users, taking into account various factors that impact the theoretical maximum bandwidth. It essentially measures the practical performance of the network in terms of delivering data. It highlights the effectiveness of the network in utilizing its resources to deliver data under actual conditions. Factors influencing throughput are Network technology, Cell configuration, Network congestion, User device capabilities, Environmental factors. It directly impacts user experience with mobile data applications like browsing, streaming, video calls, and online gaming. High throughput enables faster downloads, uploads, and more responsive applications, enhancing user satisfaction. Low throughput can lead to frustratingly slow data speeds, impacting productivity, entertainment, and overall network usability

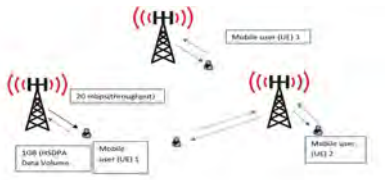

Figure 3.3: Mobile user under a Telecom network

3.6 HSDPA Data Volume

High speed data packet access is a packet based mobile telephony protocol mainly used in cellular network. To increase data speed and data capacity HSDPA is a popular protocol in 3G network. It evolved in cellular network which provides five-time faster data speed than previous version. So HSDPA data volume is simply the data consumed by a user in a telecom network by using HSDPA protocol. HSDPA data volume is the this is the entire amount of data transferred through the network in a given timeframe, measured in bytes, gigabytes, or petabytes.

Chapter 4

MBB- User Growth Prediction from Multi-variate Linear Regression Model

4.1 Research Gap Details

There are two types of machine learning supervised machine learning and unsupervised machine learning. In this chapter it is demonstrated how supervised machine learning is used to predict MBB- User (Mobile broadband user) from Multi variate Linear Regression model. The base data set is the UMTS network data which includes HSDPA Meanchthroughput. HSDPA data volume, Cell availability and Max and Mean user and the key contribution of this chapter is as follows
1. Telecom Technology is upgrading day by day like from 2G, 3G,4G and now we are heading towards 5G and 6G. No such standard research is available to predict network growth from machine learning algorithm. In this chapter, we have used multi-variate Linear Regression and support vector machine model to predict Mean user in telecom network from Data volume and cell availability.
2. Mobile network runs on proper planning and optimization. From this research we have introduced a efficient machine learning technique which will help mobile operators for good planning and optimization which is not available in current research. A good prediction of network growth and capacity always help a planner to design a efficient network which is the main focus of this thesis work.
The following sections of this chapter will have the details of Linear Regression model, Multivariate linear regression model, support vector machine, data preparation, methodology, model implantation and validation.

4.2 Supervised Machine Learning

Supervised machine learning is a fascinating and powerful branch of artificial intelligence that allows computers to learn from labeled data, meaning data where each input has a corresponding desired output. It's like having a teacher guide the computer through examples, enabling it to make predictions or classifications for new, unseen data. In this chapter we will discuss two supervised machine learning technique which are Linear regression model and support vector machine

4.3 Linear Regression model

Linear regression model is a supervised learning technique in machine learning algorithm. Normally, regression model is used to predict some continues values depending on some independent values, as example forecasting a future value depending on past values. In statistics, linear regression is often used to find the correlation between input and output variables. Linear regression model is very easy to represent in machine learning field where a distinct set of input numbers(X) is used to predict the output numbers(Y), where both are numerical. Linear regression analysis can be separated into two parts, simple linear regression model and multiple variable linear regression model. This paper will mainly analyze the multiple carrier linear regression model, where there will be two or more independent variables and one dependent variable. The formula of linear regression model is,

$$Y = B_0 + B_1X_1 + \ldots + B_nX_n + e \quad (1) \tag{4.1}$$

where:

- Y = the predicted value
- B_0 = the y-intercept
- B_1X_1 = the regression coefficient (B_1) and the first independent variable (X_1)
- B_nX_n = regression coefficient of the last independent variable
- e = model error

To find the best fit line for independent variables, multiple variable linear regression mainly calculates following three things,
1.the regression coefficients
2.t-statistics of the model
3. Associated p-value
Intercept: The value of the dependent variable when all independent variables are zero. It's represented by the symbol 0.
Slope: The coefficient that represents the change in the dependent variable for a unit change in an independent variable. It's denoted by $\beta_1, \beta_2, and so on for multiple independent variables.$
Line of best fit: The straight line that minimizes the differences between the observed data points and the predicted values from the model.

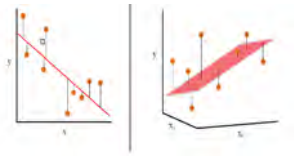

Figure 4.1: Simple Linear Regression and Multiple variable Linear Regression

How it works:
1. Data collection: Gather a dataset with observations for both the dependent variable and independent variables.

2. Model fitting: Use statistical techniques to estimate the intercept and slopes that best fit the data, typically using the least squares method (minimizing the squared differences between observed and predicted values).
3. Evaluation: Assess how well the model fits the data using measures like R-squared (how much of the variability in the dependent variable is explained by the model) and residual analysis (examining the patterns of errors in predictions).
4. Prediction: Once satisfied with the model's fit, use it to predict the dependent variable for new values of the independent variables.

4.4 Support Vector Machine

Support vector machine (SVM) is a type of supervised machine learning which is also popular in the classification problem. The main objective of SVM is to create the best decision boundary in the data set into classes so that any new data can also be classified easily. The term for the decision boundary is known as hyperplane. As name suggested SVM used points/vectors to create the hyperplane. SVM is widely used for face detection, image detection or text categorization. There are two types of SVM
1. Linear SVM: If the data set is completely classified with a straight line, then it is called the linear SVM.
2. Non-Linear SVM: If there is non-linearity in the dataset then the dataset is not classified by a straight line. For such cased non-linear SVM is used. There are a set of tuning parameters which will determine the best hyperplane. Those are as follows,
Regularization: This parameter signifies how much you want to misclassify the training data. For large values regularization the optimization will choose a smaller margin hyperplane and for low regularization value it will consider larger-margin separating hyperplane.
Gamma: The parameter gamma signifies the with how many features dataset the hyperplane is created. If gamma value is low it takes the far data and if the gamma value is high, it only takes the closer data set to build a hyperplane.
Margin: A good margin is one where this separation is larger for both the classes.

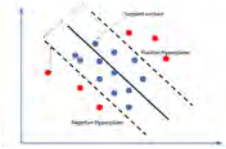

Figure 4.2: Simple Linear Regression and Multiple variable Linear Regression

4.5 Methodology of MBB-User growth prediction from Multi-variate Linear Regression model

so, our overall methodology is to divide the dataset between dependent and independent variable and the process the input data. In data processing we labeled the data set and split the data set between train data and test data. Then we input the

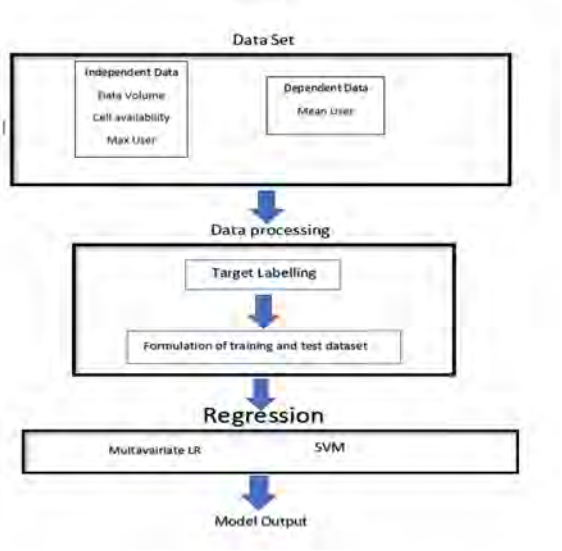

Figure 4.3: Methodology

train data set in the regression model and get our desired output data. Finally test the model with test data to check the accuracy.

4.6 Data Preparation

Let's explain our data set. We have imported 24 hours KPI data of a cellular network where we have Time periods, start time, cell name, HSDPA.MeanChThroughput, HSDPA Data Volume MB, Cell Availability, Overall Accessibility Success Rate PS Interactive EUL, Max HSDPA UE, and Mean HSDPA UE. Data set distribution shows in Fig 4.4

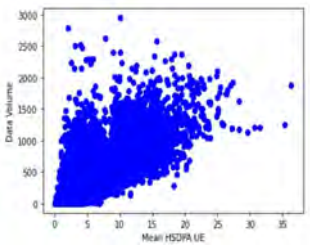

Figure 4.4: Data set Distribution

4.7 Data Details and Feature Description

From data 4.1 we can see our data set sample which we will fit in model. Table 4.2 shows the independent variable and 4.3 dependent variable.The model will take independent variables as input and predict the dependent variable which mean user. From the output operators can have a prediction of mean users from data volume, max user and cell availability.

p

Table 4.1: Sample Raw Data

Date Time	***Data Volume***	***Cell Avaiability ()***	***Max User***	***Mean User***
2/18/2021 0:00	1598.071	100	20	10.736
2/18/2021 0:00	312.347	100	9	4.294
2/18/2021 0:00	510.195	100	10	5.174

p

Table 4.2: Dependent variables with RAW data

Mean User
10.736
4.294
5.174
10.426
8.847

p

Table 4.3: Independent variables with RAW data

Data Volume	***Cell Avaiability (%)***	***Max User***
1598.071	100	20
312.347	100	9
510.195	100	10
1022.205	100	19
1465.025	100	15

4.8 Model Implementation and Validation

From our dataset we kept MAX HSDPA UE, HSDPA data volume and cell availability as a independent variable and MEAN HSDPA UE as dependent variable. That means, Operator can predict Mean User from other three variable like max hsdpa , data volume and cell availability. For Model validation we applied the OLS (ordinary least square method) for better understanding our model performance.From this analysis, we will able to find coefficients, R-squared values how well the data points fit the regression line, number of observations

4.9 Result Discussion of MBB-User growth prediction from Multi-variate Linear Regression model

From Figure 4.4 we can see our model accuracy is almost 90 percent for both model that means our model can predict Mean user 90 percent correctly from given data sets. For more analysis, we applied the OLS (ordinary least square method) for better understanding our model performance. From this analysis, we will able to find coefficients, R-squared values (how well the data points fit the regression line), number of observations. Normally, OLS is statistical method where it can analyze the relationship between one or more independent variables. Form this analysis, we can observe, or R-squared value is .957 which a better output.

Model	Score
Multi-Variable Liner Regression model	0.908153513
Support Vector Machine	0.910176985

Method	Least Squares
R-squared	0.957
No. Observations:	10938

Figure 4.5: Performance matrix

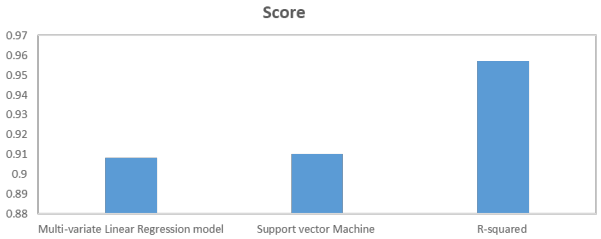

Figure 4.6: Performance of the Model (Linear Regression)

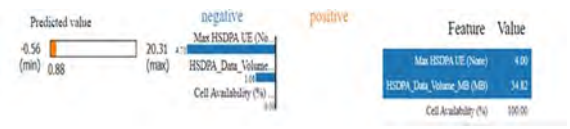

Figure 4.7: Model Interpretation with Lime Library

We have applied lime library in our model for a better model interpretation. As if we don't know what is going inside our model, we will not be able to improve it, that's why model interpretation is very useful method in machine leaning models. So, our analysis shows the prediction as per Fig. 4.5 From our analysis we can see, the most positive impact from our independent variables is max HSDPA UE rather than Cell Availability and Data Volume to determine the Mean User in a cellular network.

Form the Linear Regression model, we have got a better output and our model performed very well. From this model, we can now predict Mean user for future prediction and can have a clear idea how much mean user we can achieve from data volume, cell availability. It will help us for traffic forecasting, analyzing and planning new sites and cells. The potential applications of our linear regression model for predicting users in a telecom network depend largely on the specific outcome we predicted. Here are some general ideas based on different potential outcomes. Predicting customer turnover, resource allocation like prioritize customer service resources to focus on high-risk users, improving overall customer satisfaction and potentially reducing churn, Personalized Data Plans like can offer agreed data plans tailored to individual user's predicted usage patterns, optimizing revenue and subscriber satisfaction, Network capacity optimization, Proactive maintenance like identify areas with potentially degraded network performance based on user predictions, allowing for targeted maintenance and infrastructure upgrades. Traffic management like optimize traffic flow and congestion control based on predicted user distribution and activity, providing smoother user experience, Investment Prioritization like allocate resources and investments to areas with high predicted user concentration, we can do monitor and refine model performance, combine model with other data sources,

Chapter 5

Worst Cell Clustering in Cellular Network from K-means algorithm

5.1 Research Gap Details

There are two types of machine learning supervised machine learning and unsupervised machine learning. In this chapter it is demonstrated how unsupervised machine learning is used for Worst Cell Clustering in Cellular Network from K-means algorithm. The base data set is the UMTS network data which includes Mean User and Data volume and the key contribution of this chapter is as follows
1. Telecom Technology is upgrading day by day like from 2G, 3G,4G and now we are heading towards 5G and 6G. No such standard research is available for worst cell clustering through machine learning technique. In this chapter, we have used K-Means clustering and K-medoids Clustering for worst cell clustering in cellular network
2. Mobile network runs on proper planning and optimization. From this research we have introduced an efficient machine learning technique which will help mobile operators for good planning and optimization which is not available in current research. A good prediction of network growth and capacity always help a planner to design a efficient network which is the main focus of this thesis work.
The following sections of this chapter will have the details of Unsupervised machine learning, K-means clustering, K-medoids clustering, Elbow method, data preparation, methodology, model implantation and validation.

5.2 Unsupervised Machine Learning

Unsupervised machine learning is a type of machine learning in which does have a pre-requisite of label information. This converts the overall data into multiple clusters based on the data characteristics and by unsupervised machine learning model parameters. In this chapter two unsupervised machine learning technique will be discussed, which are K-Means clustering and K-medoids clustering. This clustering algorithm has its model parameter to cluster the data.

5.3 K-means Clustering

K-Means clustering is one of the simple machine learning algorithms to solve any clustering problems.it divides the dataset into different clusters (assume K clusters) . The main idea of this model is defining the K-center of each cluster. These centers should be identified in a intelligent way because of different locations cause different results. So, it's better to place the centers as far away from each other. The next step is to collect each point belonging a given data and place it to the nearest center. After covering all points, the first step is done and we need find the new centroids and new calculation has to be done with the data set like first step. So, in this way a loop has been created and k-centers change their location step by step until no changes are done

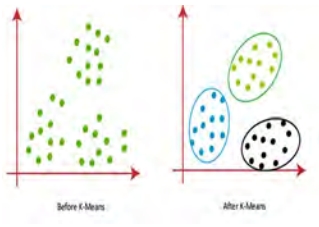

Figure 5.1: K-Means Clustering

The working method of the K-Means algorithm is explained in the below steps: Step -1: Number of K selection to decide the number of cluster Step-2: Random K points or centroid selection Step 3: Data point assignment to the closest centroid Step 4: Variance calculation and new centroid selection of each cluster Step 5: Repeat 3rd step 6: If any reassignment occurs, then go to step-4 else go to FINISH Step 7: The model is ready.

5.4 K-medoids Clustering

K-medoids clustering is a variant of K-means clustering which is more robust to noises and outliers. In this clustering method it doesn't use mean points like k-means clustering rather than it uses actual points in the cluster to represent it. Medoid is the most centrally located object of the cluster with very minimum sum of distance compare to other points. K-medoids clustering is a partitioning algorithm similar to K-means but with some key differences. It also aims to group data points into k clusters based on their similarity, but instead of using centroids (means of the cluster), it uses medoids (actual data points within the cluster) as representatives.

5.5 Elbow Method

The Elbow Method is a popular way to find the optimal number of clusters. The Method uses the value of WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster.The elbow method is a commonly used technique to help you figure out the optimal number of clusters (k) in K-means clustering. It's essentially a visual approach that relies on plotting a graph and identifying a specific point on it.

$$WCSS = \sum_{i=1}^{K} \sum_{j=1}^{|C_i|} (x_j - \mu_i)^2 \tag{5.1}$$

Where:

- i is the index of the cluster
- j is the index of the data point in the cluster
- $|C_i|$ is the number of data points in cluster i
- xj is the jth data point in cluster i
- μ_i is the centroid of cluster i

Here's how it works:
Calculate WCSS: For different values of k (typically ranging from 1 to a certain number), you run the K-means algorithm and calculate the "within-cluster sum of squares" (WCSS) for each k. WCSS measures the total distance between data points and their respective cluster centroids within each cluster. Plot the graph: Plot the WCSS values on the y-axis and the corresponding k values on the x-axis. Identify the elbow: Look for a point on the graph where the decrease in WCSS starts to slow down significantly. This point resembles an "elbow" shape, hence the name. Choose the k at the elbow: The k value at the elbow is generally considered the optimal number of clusters for your data. This is because adding more clusters after this point doesn't significantly improve the clustering quality, while it increases the complexity of the model.

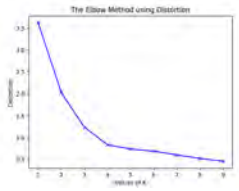

Figure 5.2: Elbow Method

5.6 Methodology of Worst Cell Clustering in Cellular Network from K-means algorithm

so, our overall methodology is to divide the dataset between dependent and independent variable and the process the input data. In data processing we labeled the data set and split the data set between train data and test data. Then we input the train data set in the regression model and get our desired output data.
In this chapter, unsupervised learning performs over the performance data set. There is two parts of the overall work. In the first part, we have used Elbow method to find the number of clusters to fit in the k-means clustering algorithm and second part we fit the data set in the model to find the worst cell in a cellular network.

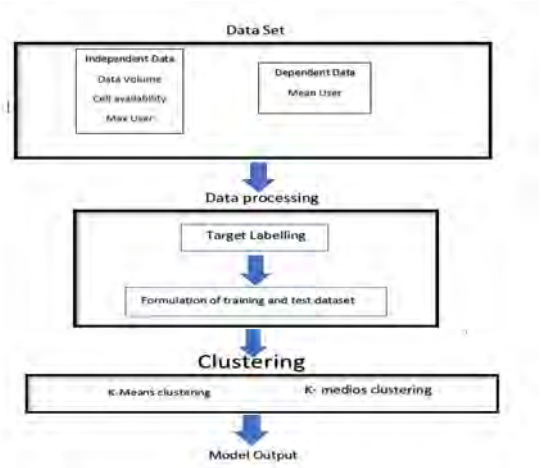

Figure 5.3: Methodology K-means clustering

5.7 Data Preparation

Let's explain our data set. We have imported 24 hours KPI data of a cellular network where we have Time periods, start time, cell name, HSDPA.MeanChThroughput, HSDPA Data Volume MB, Cell Availability, Overall Accessibility Success Rate PS Interactive EUL, Max HSDPA UE, and Mean HSDPA UE. Max HSDPA UE, HSDPA Data Volume and cell availability used as a independent variable and MEAN HSDPA UE used as dependent variable.

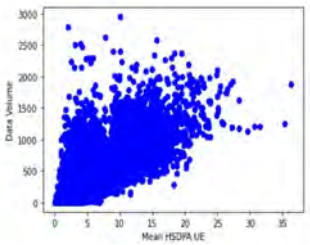

Figure 5.4: Data set distribution

5.8 Model Implementation and Validation

Before fitting our dataset in our model, we first found the Number of clusters by Elbow method. From our analysis, we found 3 number of clusters from our data which will be used to train the model.

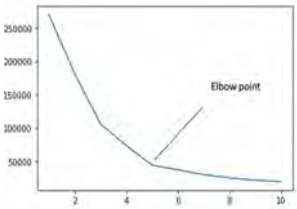

Figure 5.5: Number of Clusters by Elbow Method

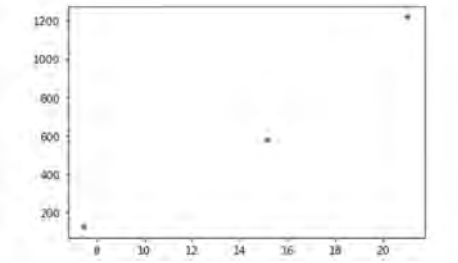

Figure 5.6: Centroids of K-Means clustering

Now, we have got the number of clusters and our data set is also ready. So we fitted our dataset in K-Means algorithm where K=3 (Number of cluster).

5.9 Data Details and Feature Description

From data 5.7 we can see our data set sample which we will fit in model. Table 5.8 shows the independent variable and 5.9 dependent variable predict the dependent variable which mean user. From the output operators can have a prediction of mean users from data volume, max user and cell availability.

Start Time	VS.HSDPA.MeanChThroughput (kbit/s)	HSDPA_Data_Volume_MB (MB)	Cell Availability (%)	Max HSDPA UE (None)	Mean HSDPA UE (None)
2/18/2021 0:00	2363.843	1598.071	100	20	10.736
2/18/2021 0:00	1450.319	312.347	100	9	4.294
2/18/2021 0:00	1907.42	510.195	100	10	5.174
2/18/2021 0:00	1450.726	1022.205	100	19	10.426
2/18/2021 0:00	1860.788	1465.025	100	15	8.847
2/18/2021 0:00	1776.941	1066.484	100	15	9.667

Figure 5.7: Sample Data set

Data Volume	*Cell Availability (%)*	*Max User*
1598.071	100	20
312.347	100	9
510.195	100	10
1022.205	100	19
1465.025	100	15

Figure 5.8: Independent variable with Raw data

The model will analyze the data set and provide worst cells in three clusters based on HSDPA data volume, cell availability, Max HSDPA UE and Mean HSDPA UE.

5.10 Result Discussion

If we plot the data, we can have a clear idea our worst cells clustering from our analysis. From our output we need to focus on cluster 0 where Mean user, Data Volume and Max UE are less only 2.74 compared to other two cluster, 7.3 and 11.8 respectively. We need to optimize, enhance capacity, physical tuning the worst cluster cells for better improvement.
From Table 5.14 we can see all the cells are clusterized in three from where we can easily Catogorized in three clusters. So K-means clustering model is conceptually

Mean User
10.736
4.294
5.174
10.426

Figure 5.9: Dependent variable with raw data

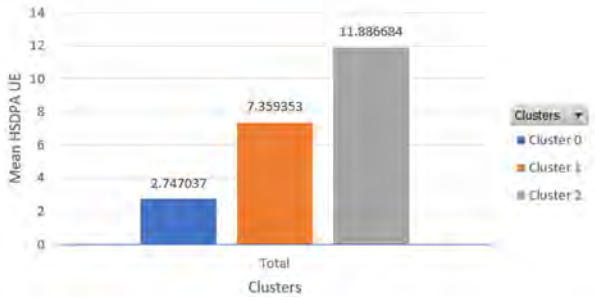

Figure 5.10: Worst cell clustering based on Mean HSDPA UE

straightforward and easy to understand, though we have worked with 24 hours data, but this model can handle large dataset and big data application. From our model the visualization is very clear to understand and helps to have a clear idea between data points. From our model output we can can following actions with the worst performing cells. We can perform some antenna optimization, power adjustment as cell optimization. We can implement more cells and cell site relocation for capacity enhancement. Deploying newer technologies like Upgrade cell sites to newer technologies like 4G or 5G, which can offer better coverage and capacity, even in low-traffic areas, potentially attracting more users and increasing data usage. Network virtualization like Implement network virtualization technologies like Network Function Virtualization (NFV) to optimize network resources and dynamically allocate capacity based on traffic demand, potentially reducing costs and improving efficiency in low-traffic areas. We can offer Offer special pricing or discounts for users in this area to incentivize data usage. We can Collaborate with other operators to share infrastructure or allow users to roam onto their network in the underutilized area. For above actions, there are some additional consideration like Cost, technical feasibility,regularity enviorment, long term strategy Considering whether the area is expected to remain underutilized or see future growth. The best course of action will depend on your specific context and goals. Analyzing data usage patterns, potential future growth in the area, and the cost-effectiveness of each option will help us make an informed decision.

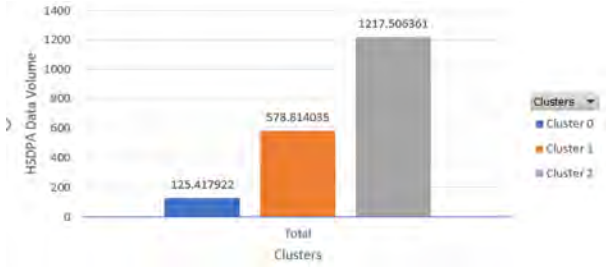

Figure 5.11: Worst cell clustering based on HSDPA Data Volume

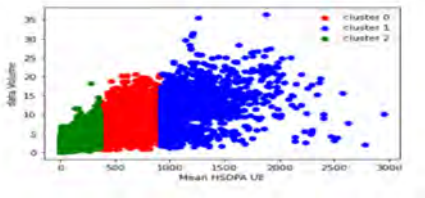

Figure 5.12: Worst cell clustering by K-means Clustering

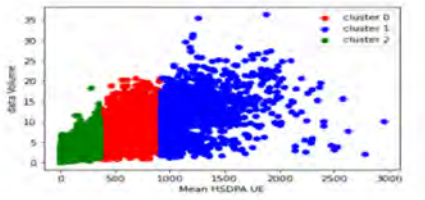

Figure 5.13: Worst cell clustering in K-Medoids algorithm

Cell Name	VS.HSDPA.MeanChThroughput (kbit/s)	HSDPA_Data_Volume_MB (MB)	Cell Availability (%)	Max HSDPA UE (None)	Mean HSDPA UE (None)	Cluster
Cell A	2363.843	1598.071	100	20	10.736	0
Cell B	1450.319	312.347	100	9	4.294	2
Cell C	1907.42	510.195	100	10	5.174	1
Cell D	1450.726	1022.205	100	19	10.426	1
Cell E	1860.788	1465.025	100	15	8.847	1
Cell F	1776.941	1066.484	100	15	9.667	1
Cell G	1701.682	1388.079	100	17	11.389	1
Cell H	2144.942	1908.918	100	28	15.211	0

Figure 5.14: Cluster wise worst cell distribution

Chapter 6

Conclusion

The thesis herein, introduces the algorithm and the model of machine learning where Multiple variable linear regression model, support vector machine, K-Means clustering method is used to predict the mean user number depending of some other variables. This analysis will help a market operation and planning team of a telecom network, to design and optimize the network and achieve maximum users under a telecom network. The worst cells which we obtained from the clustering methods will help them to work with the worst cells and solve network issues or capacity issues to get more subscriber to improve their profit. This analysis will help to plan and design cluster by cluster which is also very important for a telecom operator. The final result correctly leads the company to predict the user number of a network, which definitely provides great commercial value and help to build a good and customer centric mobile network

Bibliography

[1] Ezema, L.S. and Ani, C.I., 2016. Multi linear regression model for mobile location estimation in GSM network. Indian Journal of Science and Technology, 9(6), pp.1-6.

[2] Phillips, S., Woodward, S.L., Feuer, M.D. and Magill, P.D., 2011. A regression approach to infer electricity consumption of legacy telecom equipment. ACM SIGMETRICS Performance Evaluation Review, 38(3), pp.61-65.

[3] Jamal, Tauseef, and Muhammad Kashif Ghumman. Regression in Wireless Sensor Networks. PiEAS, 2018.

[4] Shawel, Bethelhem S., et al. "Hybrid prediction model for mobile data traffic: A cluster-level approach." 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.

[5] Aalen, Odd O. "A linear regression model for the analysis of life times." Statistics in medicine 8.8 (1989): 907-925.

[6] Poole, Michael A., and Patrick N. O'Farrell. "The assumptions of the linear regression model." Transactions of the Institute of British Geographers (1971): 145-158.

[7] Noble, William S. "What is a support vector machine?." Nature biotechnology 24.12 (2006): 1565-1567.

[8] Suthaharan, Shan, and Shan Suthaharan. "Support vector machine." Machine learning models and algorithms for big data classification: thinking with examples for effective learning (2016): 207-235.

[9] Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. "The global k-means clustering algorithm." Pattern recognition 36.2 (2003): 451-461.

[10] Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering." International Journal 1.6 (2013): 90-95.

[11] Park, Hae-Sang, and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering." Expert systems with applications 36.2 (2009): 3336-3341.

[12] Madhulatha, Tagaram Soni. "Comparison between k-means and k-medoids clustering algorithms." International Conference on Advances in Computing and Information Technology. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[13] Liu, Fan, and Yong Deng. "Determine the number of unknown targets in open world based on elbow method." IEEE Transactions on Fuzzy Systems 29.5 (2020): 986-995.

[14] Cui, Mengyao. "Introduction to the k-means clustering algorithm based on the elbow method." Accounting, Auditing and Finance 1.1 (2020): 5-8.

[15] Dieber, Jürgen, and Sabrina Kirrane. "Why model why? Assessing the strengths and limitations of LIME." arXiv preprint arXiv:2012.00093 (2020).