# Topic: A machine learning approach to predict User Number in cellular network and worst cell clustering by multiple variable linear regression and K-means clustering algorithm.

**Abstract: The paper will analyze mean user number of UMTS telecom network based on maximum HSDPA user , HSDPA 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 UMTS(3G) 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 programing, 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 a 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.**

**1 INTRODUCTION**

**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 62% of the total population of world. So, it is very important to collect, analyze and model a large number of data to conduct a valuable prediction in this field. Nowadays, many machine learning library widely used like Grap Lab, Mahout, Mad LINQ and so on. Under the background, python is mostly used more widely in the field of machine learning. Python is a very generalized programing language with strong functions and effective machine learning library. It very effectively fits for regression analysis as well as clustering analysis. The paper herein will mainly study on how to define the f relations between independent and dependent variables and then specify the parameters of the model and then finally go back to the assumed equation to predict the mean user number of a cluster area variation tendency of dependent variable based on the linear regression model and support vector machine of predicting the user number depends on data volume, max user number and cell availability of the network. Also, there is a worst cell clustering analysis with the help of K-means clustering method.**

**2 Python 3.6**

**Python 3.6 is the most popular and well received by learners in the field of machine learning. There are many features added in this version like formatting strings and literals, variable comment syntax,** **asynchronous generator, asynchronous derivation and so on. Python 3.6 attracted great attention from users for this easy learning and strong functions. Python 3.6 is a object oriented language that can widely use in the development of large scale software for its effective equations, functions, library. In Python, we can run codes and test immediately after being finished and which will help the develop the efficiency of the engineers.**

**3 Theoretical and Methodology**

**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=Bo+B1X1+……BnXn+e**

* **Y=the predicted value**
* **Bo=the y-intercept**
* **B1X1= the regression coefficient (B1) and the first independent variable(X1)**
* **BnXn= regression coefficient of last independent variable**
* **e= model error**

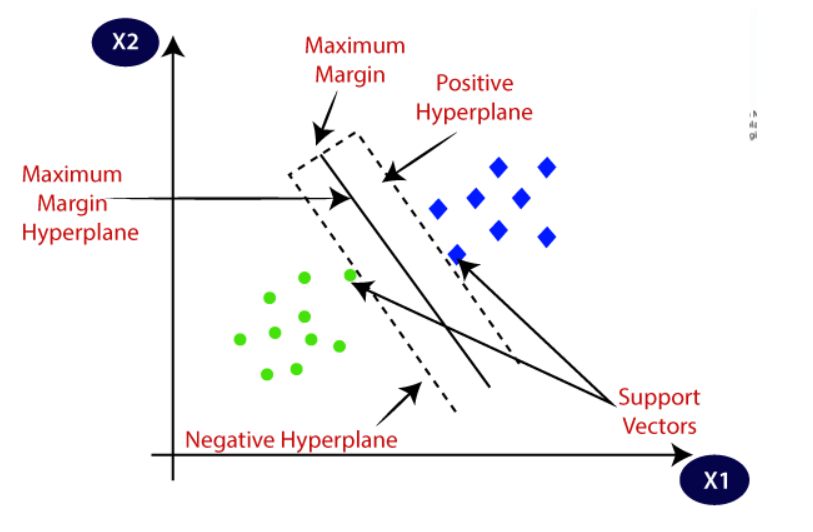
**To find the best fit line for independent variables, multiple variable linear regression mainly calculates following three things,**

* **the regression coefficients**
* **t-statistics of the model**
* **Associated p-value**

**Support vector Machine**

**Support vector machine or SVM is one the popular supervised learning algorithm which can be used for both classification and regression problems. The main objective is to create a best line or decision boundary that can precisely segregate the n-dimensional spaces between different classes so that we can put the new data in correct category in future. The decision boundary is named as hyperplane.**

**SVM normally choses the extreme points or we can call vectors to create the hyperplane. These points are called support vectors and the algorithm is named as support vector machine. Consider below diagram where two different categories are classified using hyperplane or decision boundary.**



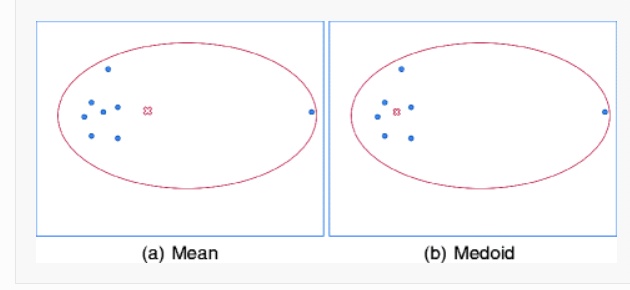
**Fig 1: Support Vector Machine**

**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.**

***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. In below figure we can see the difference between means and medoid method.**



**Fig 2: K-means and K-Medoids Clustering**

**Lime Library**

**The Main purpose of LIME, Local Interpretable Model-agnostic Explanations is to explain what machine learning are doing. LIME can explain tabular models, test classifiers and also image classifiers. Lime normally explains what happens in our predictions when give variations of data in our model. This is kind of interpretable models that is used to explain the black box of machine learning models.**

**4.1 Algorithm implementation**

**The paper will use Python3.6 to set up linear regression analysis model targeting to predict the mean user of a telecom UMTS network of a mobile operator. We introduced the unified operating system interface function in Python3.6 Moreover, we will also introduced Pandas analysis package, Sklearn, Matplotlib to analyze and display the necessary plotting of our prediction model . The importing statements of Python3.6 will be showed as followed:**

**import pandas as pd**

**from sklearn.cluster import KMeans**

**from sklearn import linear\_model**

**from sklearn.model\_selection import train\_test\_split**

**import statsmodels.api as sm**

**from sklearn.metrics import accuracy\_score**

**from sklearn.metrics import classification\_report**

**from sklearn import svm**

**import seaborn as sns**

**from sklearn.svm import SVR**

**from sklearn.svm import SVC**

**import matplotlib.pyplot as plt**

**4.2 System Model**

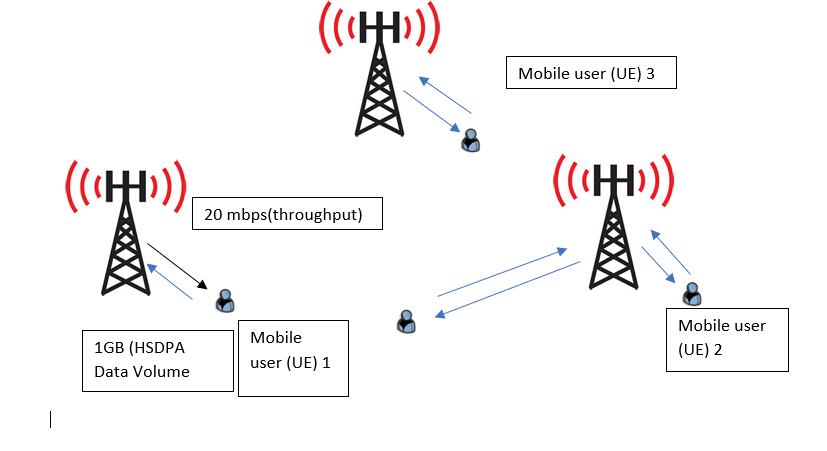
**In the system model, we consider KPI (key performance indicator) data of a UMTS (Universal Mobile Telecommunications System) network in a specific geographical area. In this KPI data we have some term which need to be cleared before we proceed to next step.**

**HSDPA\_Data\_Volume\_MB (MB): High speed data packet access is a packet based mobile telephony protocol mainly used in 3G UMTS network. To increase data speed and data capacity HSDPA is a popular protocol in 3G network. It evolved in WCDMA 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.**

**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). Nodeb is a telecommunication node which is serving in a telecommunication network where users are using its resource for voice and data service.**

**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.**

**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.**



**Fig 3: Mobile user under a telecom network**

**Our 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.**

**The procedure to predict the data volume in a cellular network from max, mean users and cell availability data is as flow chart in fig 4.**

Get the 3G network KPI data from Wireless OSS terminal

Plot the graph and show output result

Met the expectation

Predict the Data Volume

Analyze the KPI data by using Multiple carrier linear regression model and Support vector machine

Plot the graph and show output result

Predict the Mean HSDPA UE

Analyze the KPI data by using Multiple carrier linear regression model and Support vector machine

Yes

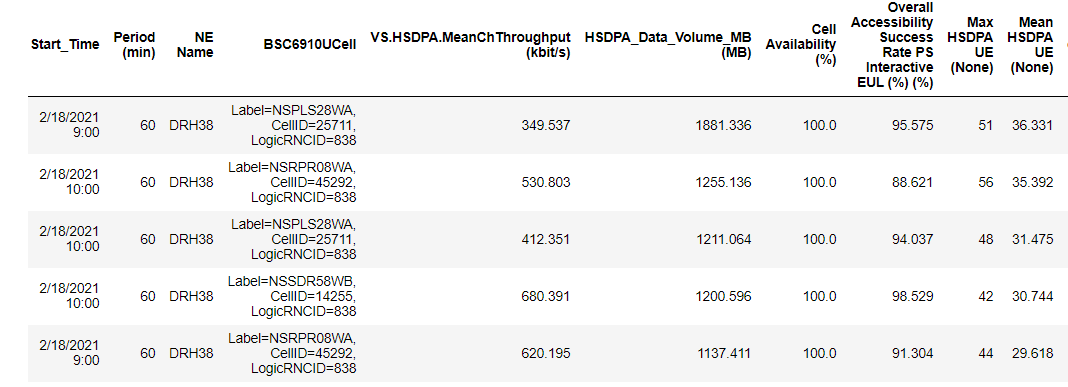
No

Met the expectation

**Fig 4: Flow chart for mean user number prediction**

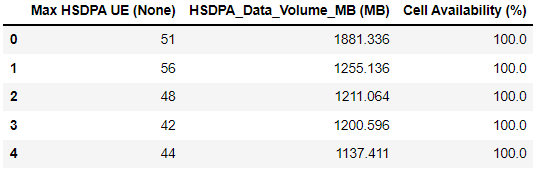
**5. Data Preparation and Result Analysis**

**Let’s Explain our data set. We have imported 24 hours KPI data of a UMTS 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,** **Mean HSDPA UE. We can see our dataset distribution in figure 5.**

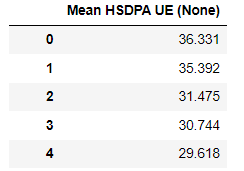


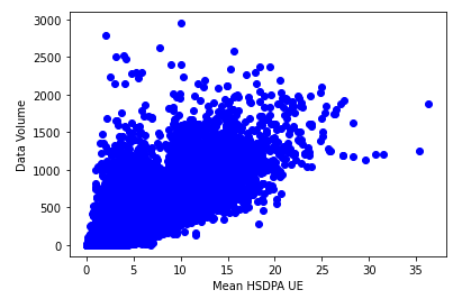
**First of all, we split the data set into training and test data set to fit in the model. We are considering 80% data for our training purpose and rest 30% for our test data. The length of independent and dependent data set is 10938 and test data set is 2735.After data split is completed, we will now fit our data into a multiple variable linear regression model and predict the output. In the same way we will fit our data into a SVM (support vector machine) for comparison purpose. So, after analyzing both model we observed following output from our both models.**

**Independent variables**



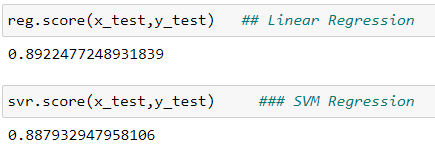
**Dependent Variable**



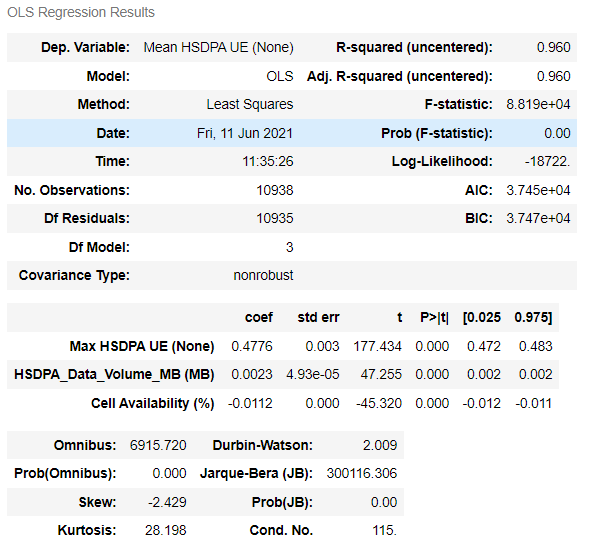


**Fig 5: Data set distribution**

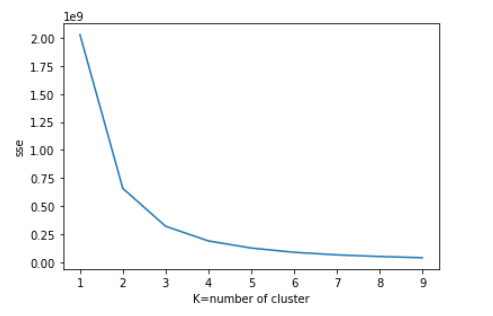
**Model Performance**



**So, here we can see our model accuracy is almost 90% for both model, that means our model can predict Mean user 90% 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), P-value, t-value, number of observations. Normally, OLS is statistical method where it can analyze the relationship between one or more independent variables.**



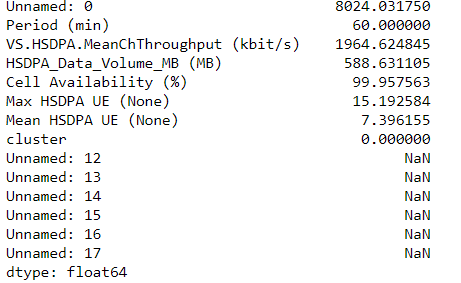
**Form this analysis, we can observe, or R-squared value is .960 which a better output. Also, we can see the coefficients, standard error and t-values from our analysis. “Max HSDPA UE” is covering 4% and HSDPA\_DATA Volume .23% coefficient data to find the user number in our dataset. The next step is to find the number of clusters to determine the worst cells from our KPI data. We applied an Elbow method to find the number of clusters.**



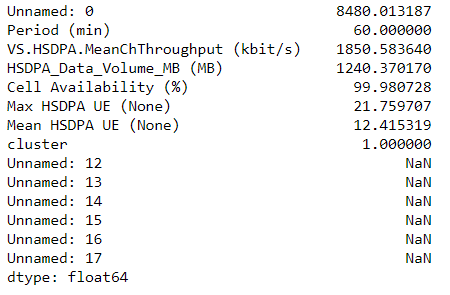
**Fig 6: Elbow method to find the number of clusters**

**From the result, we can fix three clusters for our analysis. Now, we will apply K-Means clustering algorithm to find the worst cells from our dataset. After analyzing, we have found three different clustering output from our model.**

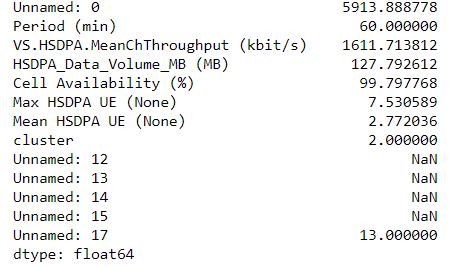
**Cluster 0:**



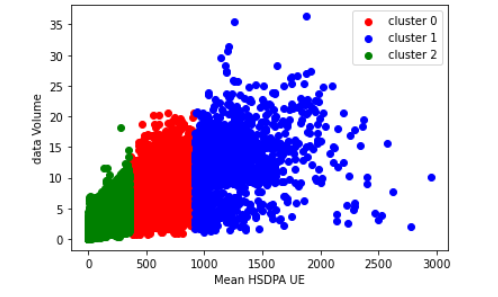
**Cluster 1:**



**Cluster 2:**

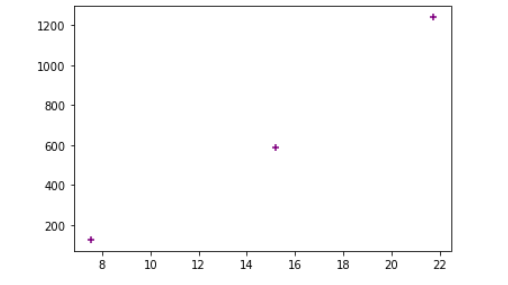


**So, we have divided our data set into three clusters where the minimum data users found in cluster 2 where mean HSDPA UE is only 2.77. In cluster 1, mean HSDPA UE is 12.415 which is better than cluster 0 where mean HSDPA user is only 7.39. So, if we plot the data, we can have a clear idea of our worst cells clustering from our analysis.**



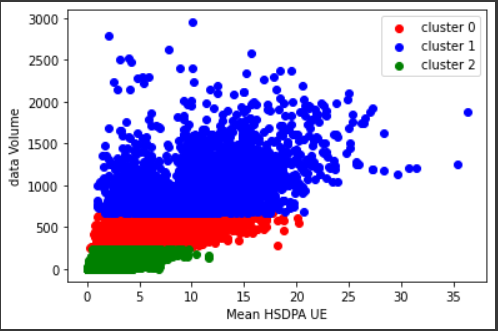
**Fig 7: Worst cell clustering by K-means algorithm**

**From below figure we can observe the centroids points determined by our k-means clustering algorithm.**



**Fig 8: Centroids points in K means clustering**

**We have also applied a K-Medoids and found almost a similar output.**



**Fig 9: Worst cell clustering in K-Medoids algorith**

# 5.1 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 below prediction

# 

# Fig 10: Model Interpretation by Lime Library.

# 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.

# 6 Related Work

# In this section, we discuss some of the significant related works and challenges,

# 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% 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 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% 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.

# 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 multilinear 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.

# 7 Conclusion

T**he Paper 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**

# 7 References

# [1] Longinus S. Ezema\* and Cosmas I. Ani, ”Multi Linear Regression Model for Mobile Location Estimation in GSM Network”

# February 2016

# [2] Steven Phillips, Sheryl L. Woodward, Mark D. Feuer “A regression approach to infer electricity consumption of legacy telecom equipment”

# [3] Muhammad Kashif Ghumman and Tauseef Jamal,” Regression in Wireless Sensor Networks”