On The Analysis of Customer Engagements with A Telecommunication Company in Sokoto-North western Nigeria Using Machine Learning Techniques

Bello.A. Bodinga
Department of Computer Science,Usmanu Danfodiyo University, Sokoto, NIGERIA
A.O. Faramade
Undergraduate Student, Department of Computer Science,Usmanu Danfodiyo University, Sokoto, NIGERIA
Email: bellobodinga@gmail.com
Bello.A. Buhari
Department of Computer Science,Usmanu Danfodiyo University, Sokoto, NIGERIA
Muzzammil Mansur
Department of Computer Science, Waziri Ummaru Federal Polytechnic. Birnin-Kebbi, NIGERIA

ABSTRACT

This study was intended to analyse data mining techniques on the customer engagements with telecommunication companies in Nigeria. This study was guided by the following objectives; to provide an overview, on how prediction is being made in a telecommunication company using data mining. MTN Nigeria was chosen as a case study to identify fraud telecommunication companies in Nigeria; to identify the challenges of data mining faced by telecommunication companies in Nigeria. The study employed the descriptive and explanatory design; primary means were applied in order to collect data. Primary data sources were used and data was analyzed using orange data mining software. The study findings revealed that data mining significantly impacts on the performance of telecommunication industries. In this paper, we made an attempt in to the analysis of telecommunication company data to assess the impact of customer engagements.

Keywords**: MTN, Machine learning, K-Nearest Neighbor, K-Mean, Decision tree**

Date of Submission: Apr 19, 2022 Date of Acceptance: May 18, 2022

I. INTRODUCTION

The telecommunication industries generate and store a tremendous amount of data [4]. These data include call details data, which describes the calls that traverse the telecommunication networks, network data, which describes the state of the hardware and software components in the network, engagement data, which record the number of activities perform within a period of time and customer data, which describes the telecommunication customers and their engagements [5]. The amount of data is so great that manual analysis of the data is difficult, if not impossible. The need to handle such large volume of data led to the development of knowledge-based expert systems. These automated systems performed important functions such as identifying fraudulent phone calls and identifying network faults. The problem with this approach is that it is time consuming to obtain the knowledge from human experts (the “knowledge acquisition bottleneck”) and, in many cases; the experts do not have the requisite knowledge. The advent of data mining technology promised solutions to these problems and for this reason the telecommunications industry was an early adopter of data mining technology [5].

Telecommunication data pose several interesting issues for data mining. The first concerns scale, since telecommunication databases may contain billions of records and are amongst the largest in the world. A second issue is that the raw data is often not suitable for data mining. For example, both call detail and network data are time-series data that represent individual events. Before this data can be effectively mined, useful “summary” features must be identified and then the data must be summarized using these features. Because many data mining applications in the telecommunications industry involve predicting very rare events, such as the failure of a network element or an instance of telephone fraud, rarity is another issue that must be dealt with. The fourth and final data mining issue concerns real-time performance because many data mining applications, such as fraud detection, require that any learned model/rules be applied in real-time [3]. Several techniques have also been applied is tackling all these issues in telecommunication companies.

MTN Nigeria [11] is part of the MTN Group, Africa's leading cellular telecommunications company. On May 16, 2001, MTN became the first GSM network to make a call following the globally lauded Nigerian GSM auction conducted by the Nigerian Communications Commission earlier in the year. Thereafter the company launched full commercial operations beginning with Lagos, Abuja and Port Harcourt. MTN paid $285m for one of four GSM licenses in Nigeria in January 2001. To date, in excess of

US$1.8 billion has been invested building mobile telecommunications infrastructure in Nigeria.

Since its launch in August 2001, MTN has steadily deployed its services across Nigeria. It now provides services in 223 cities and towns, more than 10,000 villages and communities and a growing number of highways across the country, spanning the 36 states of the Nigeria and the Federal Capital Territory, Abuja. Many of these villages and communities are being connected to the world of telecommunications for the first time ever.

Fraud is a serious problem for telecommunication companies, leading to billions of dollars in lost revenue each year. Fraud can be divided into two categories: subscription fraud and superimposition fraud. Subscription fraud occurs when a customer opens an account with the intention of never paying for the account charges. Superimposition fraud involves a legitimate account with some legitimate activity, but also includes some "superimposed" illegitimate activity by a person other than the account holder. Superimposition fraud poses a bigger problem for the telecommunications industry and for this reason data mining technique is used for identifying this type of fraud. These applications should ideally operate in real-time using the call detail records and, once fraud is detected or suspected, should trigger some action. This action may be to immediately block the call and/or deactivate the account, or may involve launching an investigation, which will result in a call to the customer to verify the legitimacy of the account activity. However, this study will examine various data mining techniques of telecommunication companies in Nigeria.

II. RELATED WORK

Telecommunication companies, like other large businesses, may have millions of customers. By necessity, this means maintaining a database of information on these customers. This information will include name and address information and may include other information such as service plan and contract information, credit score, family income and payment history. This information may be supplemented with data from external sources, such as from credit reporting agencies. The customer data maintained by telecommunication companies does not substantially differ from that maintained in most other industries [4]. However, customer data is often used in conjunction with other data in order to improve results [6]. For example, customer data is typically used to supplement call detail data when trying to identify phone fraud.

Several researches have been made in the field of customer attrition and retention analysis in banking sector [7]. Some studies reveal that the most important variables influencing customer choice are effective and efficient customer services, speed and quality services, variety of services offered and low e-service charges, online banking facilities, safety of funds and the availability of technology based service(s), low interest rate on loan, convenient branch location, image of the bank, well management, and overall bank environment. On the other hand, customer is the core of their operation, so nurturing and retaining them are important for their success. Many researches [2,4,5,6] were held on customer retention as well as customer attrition analysis Lift is used as a proper measure for attrition analysis and compare the lift of data mining models of decision tree, boosted naive Bayesian network, selective Bayesian network, neural network and the ensemble of classifiers of the above methods. Their main focuses were on attrition analysis using lift. Lift can be calculated by looking at the cumulative targets captured up to p% as a percentage of all targets and dividing by p%. A churn model [9,10] with a higher predictive performance in a newspaper subscription context was constructed support vector machines. They showed that support vector machines show good generalization performance when applied to noisy marketing data. The model outperforms a logistic regression only when the appropriate parameter-selection technique is applied and SVMs are surpassed by the random forests. A software using Clementine was used to analyze 300 records of customers Iran Insurance Company in the city of Anzali, Iran. They used demographic variables to determine the optimal number of clusters in K-means clustering and evaluated binary classification methods (decision tree QUEST, decision tree C5.0, decision tree CHAID, decision trees CART, Bayesian networks, Neural networks) to predict customers churn used Decision trees and Neural Networks to develop model to predict churn. Models generated are evaluated using ROC curves and AUC values. They also adopted cost sensitive learning strategies to address imbalanced class labels and unequal misclassification costs issues discussed commercial bank customer churn prediction based on SVM model, and used random sampling method to improve SVM model, considering the imbalance characteristics of customer data sets. A study investigated determinants of customer churn in the Korean mobile telecommunications service market based on customer transaction and billing data. Their study defines changes in a customer's status from active use ton on-use or suspended as partial defection and from active use to churn as total defection. Results indicate that a customer's status change explains the relationship between churn determinants and the probability of Churn [10]. A neural network (NN) based approach to predict customer churn in subscription of cellular wireless services. Their results of experiments indicate that neural network based approach can predict customer churn with accuracy more than 92%. Anacademic database of literature between the periods of 2000–2006 covering 24 journals andproposes a classification scheme to classify the articles. Nine hundred articles were identified and reviewed for their direct relevance to applying data mining techniques to CRM. They found that the research area of customer retention received most research attention; and classification and association models are the two commonly used models for data mining in CRM A critique on the concept of Data mining and Customer Relationship Management in organized Banking and Retail industries was also discussed. Most of these paper used existing customer's data from a single database. Some of them used only demographic data. But in our system, we used data from different branches of a bank and merge these into a single database. We have analyzed borrower's transactional data. We focused on predicting prospective business sectors to disburse loan in retailing commercial bank.

III. THE DATA SET

The data set used for this research was collected at MTN office in Sokoto- Nigeria, it contains customer engagement with the company from the month of May 2021 to the month of August 2021. It has four rows, eighty-seven columns in comma separated values (csv). The data feature the date which is ignored as meta, Subscriber Identification Module (SIM) registered, SIM upgraded which is the target value and SIM replaced, the file size is about 2.48 KB (2,548 bytes).

IV. TOOLS AND METHODS

ORANGE, data mining software which was originally developed by scientists at the University of Ljubljana in 1997 using the Python, C ++ and C programming languages is used. The software's graphical environment and interfaces have been developed using the Python and Qt3 libraries [12]. This software, with the latest version presented on March 6, 2017 with ORANGE 3.4.0 has a simple interface on which users create a data analysis workflow by placing graphical components (widgets).

V. TECHNIQUE AND ALGORITHM

1. **Classification techniques**: is a technique where we categorize data into a given number of classes. The main goal of classification is to identify the category or class to which a data will fall under. In this research work, we have used two classifiers namely decision tree, k-nearest neighbour classifier, which are briefly explained as follows.

   a. **Decision Tree (DT):** In this algorithm, the input features are used to construct a tree. A set of rules representing the different classes is then derived from the tree. These rules are used to forecast the class of a new instance with an unknown class [1].

   b. K-Nearest Neighbours Classifier (KNN): This is a supervised learning method where a new coming sample is classified based on the closest training samples present in feature field. When the test data is given, it is mapped to the class that is most common among the k neighbours [1].

2. **Clustering techniques**: is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other group (clusters).

a. K-Mean algorithm: is an unsupervised learning algorithm, which group the unlabelled dataset into different cluster.

b. Density based spatial clustering of application (DBSCAN): this is a popular learning method utilize in model building and machine learning algorithms. This is use in machine learning to separate clusters of high density to clusters of low density.

VI. EXPERIMENTAL SETUP

This experiment was performed with Microsoft Windows 10, with physical memory (RAM) of 4GB, processor: intel(R) Celeron(R) CPU N2920 @1.86GHz, 1863Mhz, 4core(s) and system type: 64-bit operating system x64-based processor and with hard disk of 194GB. The data collected was edited in Microsoft Excel Spreadsheet Software version 2016, it was save as comma separated value(csv file), and orange data mining software of version 3.29 is used to analyze the data collected. Orange software contains a graphical component called Widget, widget is place into a blank space called Canvas and the result is also display in Canvas.

VII. RESULTS AND DISCUSSIONS

This section is devoted to presentation, analysis and interpretation of data used in research, the data are based on the number of customer that went to MTN office in Sokoto to registered, upgrade and replace their SIM card.

Discussion on Fig. I

Node 1: since there is no input of SIM replace at first node, we jump to node 2
Node 2: if the SIM replaced is ≤ 28, the possibility of upgrading 17 SIM is 8.8%
Node 3: if the SIM replacedis > 28, the possibility of upgrading 31 SIM is 11.3%
Node 4: if the SIM replaced is ≤20, the possibility of upgrading 26 SIM is 12.5%
Node 5: if the SIM replaced is >20, the possibility of upgrading 36 SIM is16.7%
Node 6: if the SIM replaced is ≤39, the possibility of upgrading 31 SIM is 14.7%
Node 7: if the SIM replaced is> 39, the possibility of upgrading 32 SIM is 10.5%
Node 8: if the SIM replaced is ≤ 16, the possibility of upgrading 15 SIM is 12.5%
Node 9: : if the SIM replacedis > 16, the possibility of upgrading 26 SIM is 25.0%
Node 10: if the SIM replaced is≤23, the possibility of upgrading 36 SIM is 22.2%
Node 11: if the SIM replaced is >23, the possibility of upgrading 17 SIM is22.2 %
Node 12: if the SIM replaced is ≤ 34, the possibility of upgrading 31 SIM is 17.4%
Node 13: if the SIM replaced is > 34, the possibility of upgrading 42 SIM is 18.2 %

Decision tree

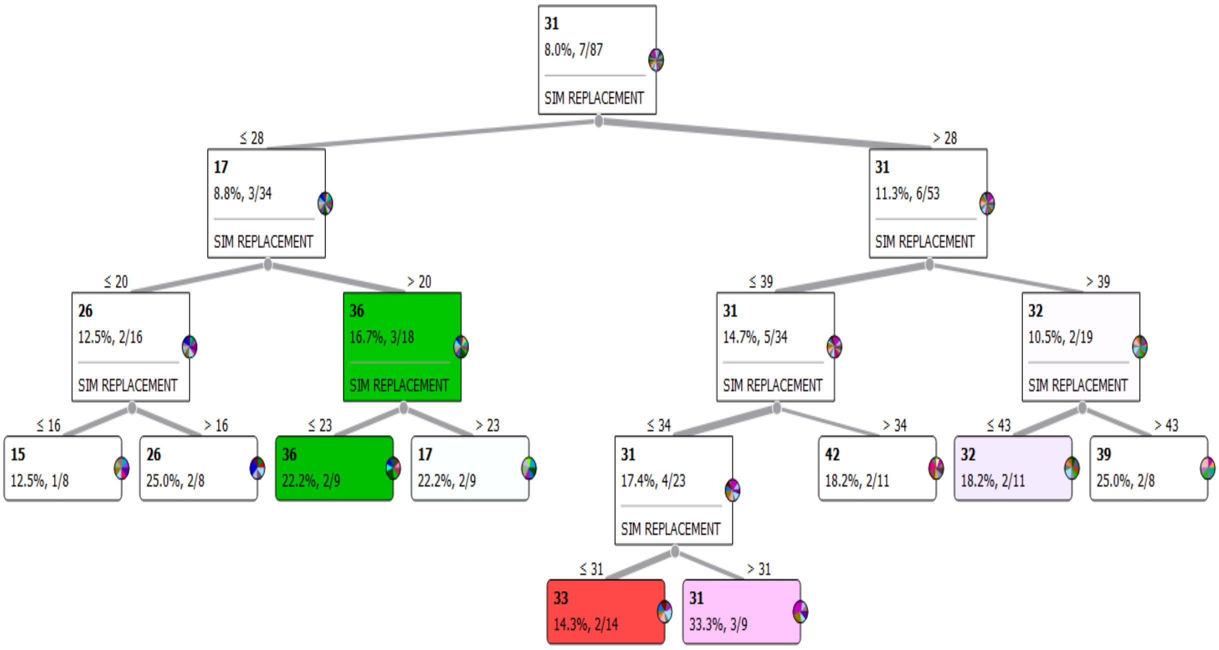


Fig. 1 Decision tree, showing how prediction is made in a telecommunication company.

Node 14: if the SIM replaced is ≤43, the possibility of upgrading 32 SIM is 18.3 %

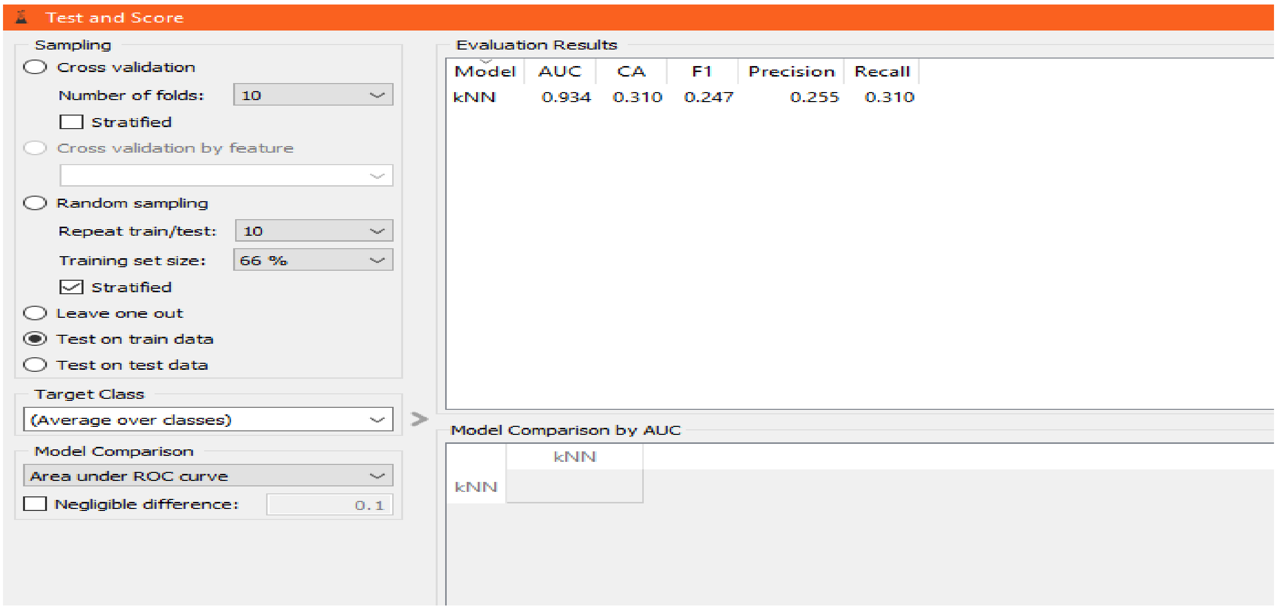
Node 15: if the SIM replaced is> 43, the possibility of upgrading 39 SIM is 25.0%

Node 16: if the SIM replaced is ≤ 31, the possibility of upgrading 33 SIM is 14.3 %

Node 17: if the SIM replaced is > 31, the possibility of upgrading 31 SIM is 33.3 %

Each node contain the number of SIM upgraded and prediction value in percentage, and SIM replacement.Excluding root node, ontop of each node is the number of SIM replaced or registered.

K-Nearest Neighbor



Test and Score

Sampling

○ Cross validation

Number of folds: 10

□ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

● Test on train data

○ Test on test data

Target Class

(Average over classes)

Model Comparison

Area under ROC curve

□ Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.934	0.310	0.247	0.255	0.310

Model Comparison by AUC

	kNN
kNN	

Fig. 2 K Nearest Neighbor classifying the data

Discussion on Fig. 2

Areal under curve AUC: the area under the receiver operating curve = 0.935

Classification Accuracy CA: is the proportion of correctly classified = 0.310

F1 : is the weight harmonic mean of precision and recall = 0.247

Precision: is the proportion of true positive among instance classified as positive = 0.255

Recall: is the proportion of true positive instances in the data 0.310

The classification accuracy predicted is **0.310**

K-Mean

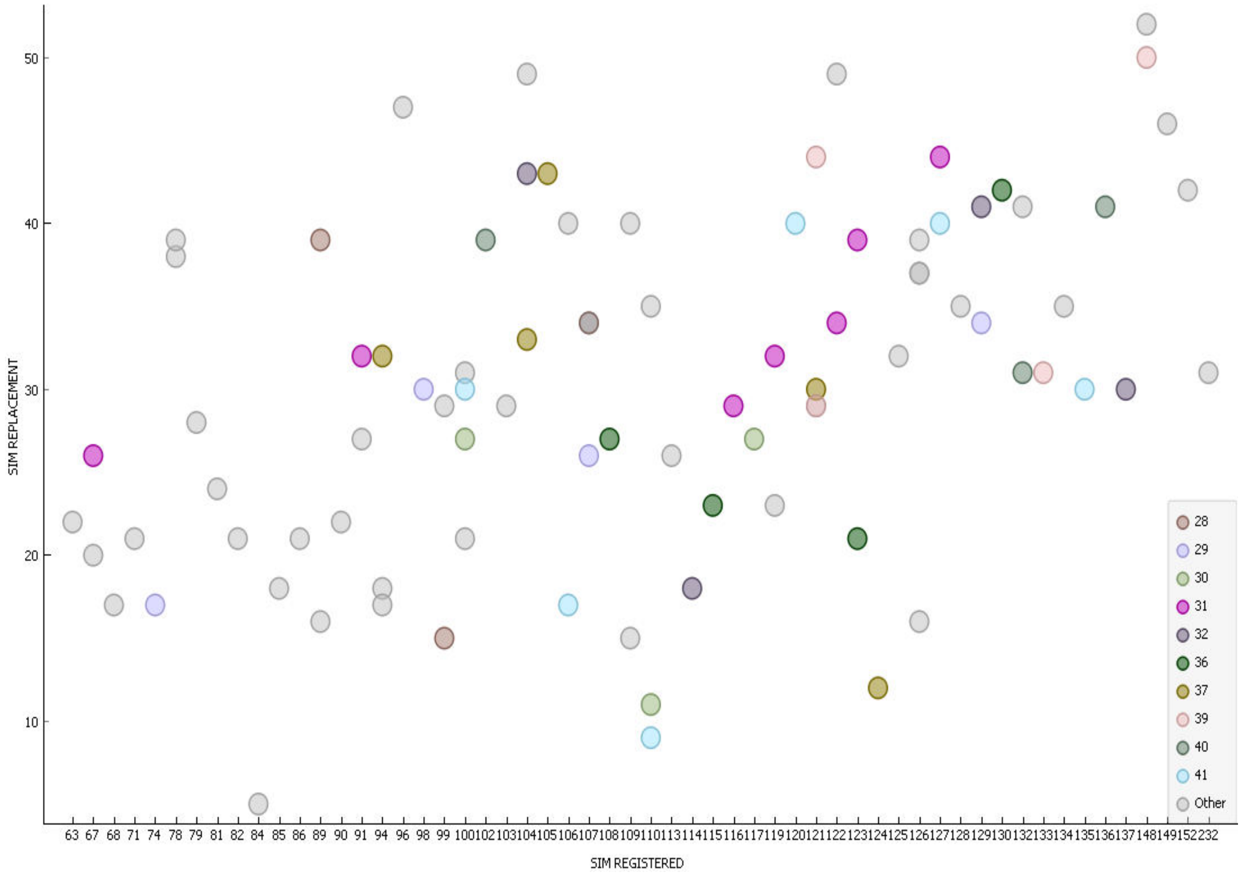


Fig. 3 k-mean grouping of the telecommunication data

Discussion on Fig. 3

Pink class

When the SIM upgraded is 28 , SIM registered 89 and the SIM replaced is39

white class

When the SIM upgraded is 29, SIM registered 74 and the SIM replaced is 17

sage green class

When the SIM upgraded is 30, SIM registered 117 and the SIM replaced is 27

Lime class

When the SIM upgraded is , SIM registered and the SIM replaced is

Pink class

When the SIM upgraded is 31, SIM registered 119 and the SIM replaced is 32

Air force blue class

When the SIM upgraded is 32, SIM registered 114 and the SIM replaced is 18

Green class

When the SIM upgraded is 36, SIM registered 108 and the SIM replaced is 27

Gold class

When the SIM upgraded is 37, SIM registered 105 and the SIM replaced is 43

Turkish blue

When the SIM upgraded is 40, SIM registered 136 and the SIM replaced is 41

Light blue

When the SIM upgraded is 41, SIM registered 135 and the SIM replaced is 30

Orange cluster all the related activities into a single colour,the data file is applies to the K-Mean widget and K-mean is being visualize by applying it to a scater plot. The above result classify the the SIM upgraded, the SIM registered and the SIM replaced with the same colour.

Density based spacial cluster of apllication (DBSCAN)

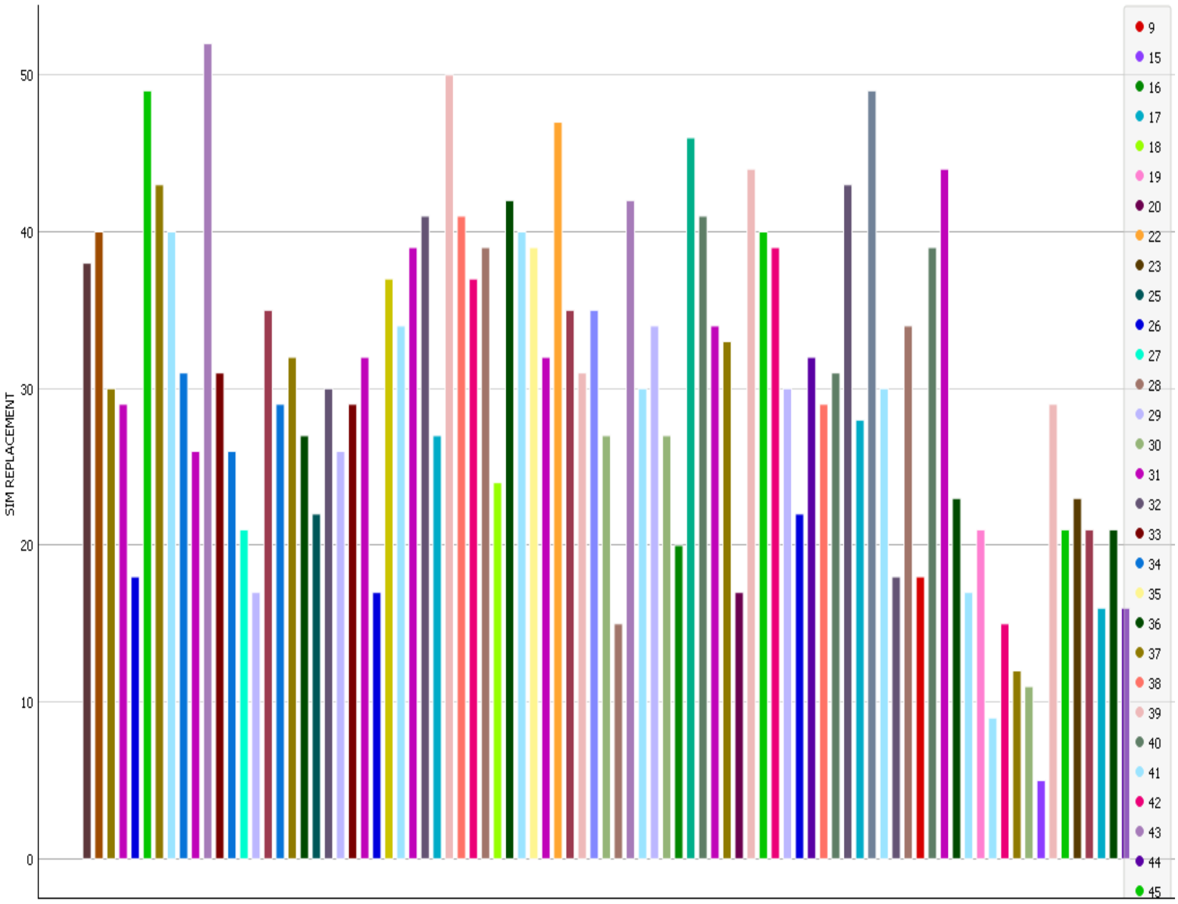


Fig. 4 DBSCAN grouping telecommunication data

Discussion on Fig.4

RedClass: if the SIM upgraded is 9 and 18 SIMs is replaced

Orchid Violet Class: if the SIM upgraded 15 and 5 SIMs is replaced

Green Class: if the SIM upgraded16and 20 SIMs is replaced

Light Blue Class: if the SIM upgraded17and 16 SIMs is replaced

Lime Class: if the SIM upgraded18 and is 24 SIMs replaced

Creamy pink: if the SIM upgraded is 19 and 21SIMs replace

Purple class: if the SIM upgraded is 20 and 17 SIMs is replaced

Barn red class: if the SIM upgraded is 23 and 23 SIMs is replaced

Green class: if the SIM upgraded is 25 and 22 SIMs is replaced

The widget applies the DBSCAN clustering algorithm to the data and its visual by bar plot. The widget also shows the sorted graph by differentiating it by colour, SIM register is plot again SIM upgrade.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we analyze the impacts of customer engagements on MTN dataset. Forecasting through researches showed that in year 2030, more than 90% of world population will be connected via internet and most of it will be using wireless mode to connect especially GSM or future telecom architecture. It will generate huge amount of data for telecommunication industry for which the industry is not ready to deal with. Most of this data will be redundant, therefore, proper data mining tools and techniques are required to dig out for the required data and dump the redundant data. In the future, we aimed at analyzing a larger data covering a period of not less than three years to forecast the future engagements of customers in the telecommunication sector

REFERENCES

[1] Aregbeyen, A. (2011). The Determinants of Bank Selection Choices by Customers: Recent and Extensive Evidence from Nigeria. *International Journal of Business and Social Science.Vol.* 2, No. 22, pp.276-288.

[2] BenlanHea, Y. & QianWan, X. (2014). Prediction of customer attrition of commercial banks based on SVM model. 2nd *International Conference on Information*

Technology and Quantitative Management, ITQM 2014, Procedia Computer Science Vol. 31, pp.423 – 430.

[3] Ezawa, K. & Norton, S. (1995). Knowledge discovery in telecommunication services data using Bayesian Network models. Pp 100.

[4] Han, J., Altman, R. B., Kumar, V., Mannila, H., & Pregibon, D (2002). Emerging scientific applications in data mining. Communications of the ACM; 45(8): 54-58.

[5] Ngai, E.W.T. Li Xiu, D.C.K. % Chau, (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications. Vol. 36: pp. 2592–2602.

[6] Oshini Goonetilleke, T.L & Caldera, H.A. (2013). Mining Life Insurance Data for Customer Attrition Analysis. Journal of Industrial and Intelligent Information. Vol. 1: 52-58.

[7] Rehman,H. U. & Ahmed, S. (2008). An Empirical Analysis of the determinants of bank selection in Pakistan; A customer view. Pakistan Economic and Social Review. Vol. 46, no.2, pp.147-160.

[8] Siddiqi, K. O. (2011). Interrelations between Service Quality Attributes, Customer Satisfaction and Customer Loyalty in the Retail Banking Sector in Bangladesh. International Journal of Business and Management. Vol. 6, No. 3, pp.12-36.

[9] Soeini, R. A. & Rodpysh, K.V. (2012). Evaluations of Data Mining Methods in Order to Provide the Optimum Method for Customer Churn reduction: Case Study Insurance Industry", International Conference on Information and Computer Applications. Vol. 24: 290-297.

[10] ZHOA Shan, M. LIU Ai-Jun, L. (2007), "A predictive Model of Churn in Telecommunications Base on Data Mining"., IEEE International Conference on Control and Automation", Guangzhou, China.

[11] MTN official website retrieved from https://www.mtnonline.com/ on 11/01/2022

[12] ORANGE official website retrieved from https://orangedatamining.com/ on 15/01/2022