Enhanced Subscriber Churn Prediction Model for the Mobile Telecommunication Industry

By

E. N. Onwuka, B. A. Salihu and *I. A. Abdulrahman
Department of Telecommunication Engineering
Federal University of Technology,
Minna, Niger state, Nigeria.
Email: onwukaliz@futminna.edu.ng, salbala@futminna.edu.ng, Ibrahim4u001@gmail.com*

ABSTRACT
Subscriber churn is a major cause of worry of many industries which require low or zero switching cost. Telecommunication industry can be considered as the most affected and top the list with approximate annual churn rate of 30%. Recently Mobile Network Operators (MNOs) have implemented customer relation management with intention to reduce the number of Subscriber churn, but it is still faced with high churn rate in the industry. It is important to recognize the potential churners before they churn. At this era of Big Data, the telecos have the advantage of using user generated data to predict customer churn. Service usage metrics such as account ID, service ID, Activation date, Deactivation date and others like network performance indicators and traditional demographic information such as Zip code, Age, Sex, population density, cell site coverage are employed by MNOs for churn prediction. The challenge lies in developing effective prediction techniques, this work is aimed at using the Genetic Algorithm for optimal selection of churn attributes from call detail records (CDR) and Artificial Neural Network for churn prediction based on the selected attributes. The WEKA (Waikaito Environment for Knowledge Analysis) tool was used for this work.

Key words: Artificial Neural Network, Churn, Genetic Algorithm, Mobile Network Operators, Social Network Analysis.

INTRODUCTION

In the present day highly competitive mobile telecommunication industry, subscribers can select from many mobile network providers and without difficulty move from one service provider to another. The main emphasis of Mobile Network Operators (MNOs) has therefore shifted from building a large customer base to keeping customers in-house. For this reason, it is important to know which customers are likely to switch to a competitor in the near future. Those customers are called churned subscribers. In light of the above, subscriber churn has become a very important issue to the MNOs. Because subscribers switch operators for a variety of reasons. Faced with this threat, MNOs should be equipped with the most efficient and effective methods of predicting a subscribers' behaviour to know the future decision\n to be taken by the subscriber with respect to churn. Big data and analytics is coming as a handy tool to help manage churn. Fortunately, telecommunication is one industry that has a deluge of human behavioural data. Effective churn management requires developing a data driven predictive technique that enables operators to keep their existing customers [1]. Customer retention is clearly more important than customer acquisition as it cost more to acquire a new customer than to retain the old ones. In the US wireless market, for example, the retention cost of a customer was estimated at $10 while the cost to acquire a new one is at $40 [2]. This is partly because service providers acquire considerable information about existing customers and can analyze this valuable information to understand their preferences and behaviour [3].In the Nigeria scene, as at March 2016, the total number of Global system for mobile communication (GSM) subscribers was 147,398,854, an increase of 5,756,018, or 4.06% relative to March 2015 [4]. This statistics points to the fact that the Nigerian mobile market is nearing saturation, which means that subscriber churn will be on the increase.

RELATED WORK

Churn prediction is an important area of focus for data analytics for telecom providers. This is evident in the fact that in 2009 the ACM Conference on Knowledge Discovery and Datamining (KDD) hosted a competition on predicting mobile network churn using a large dataset posted by Orange Labs [5]. A study by [1] compared various machine learning and classification algorithms for churn prediction, not focusing on any particular machine learning tool. However, the work the work did not identify a really effective prediction technique with appreciable accuracy. [6] Considered the use of customer demographics, contractual data, customer service logs, complaint data, billing and payment information for churn prediction in mobile telecommunication industry. These are a part of the attributes contained in a CDR, the authors did not explain what influenced the choice of these attributes or whether a different set of attributes could yield a better result. In [7] investigation about retention factors in telecommunication industry was conducted by examining other features and variables such as complaint data and service usage metrics. The focus of the study is understanding the factors related to customer behaviour. The methodology used was a binary logistic regression model and a two level linear model. Four attributes was considered in the study: service plan complexity, handset sophistication, length of call, and recipient's network. Again it was not clear whether the chosen attributes were optimal. [13] In their work, focused on each subscriber in isolation to analyze their usage, billing, and service data. Three attributes were considered from the call detail records (CDR).An accuracy of 89.4% was achieved which was appreciable. Nevertheless, it was not clear how the three attributes were chosen or whether they were the best attributes from the CDR for churn prediction. It is believed that if the optimal attributes for churn prediction were used, accuracy will be further improved. Therefore to further improve accuracy in churn prediction, this research work used an optimization technique to optimize the 21 attributes of CDR Dataset with a view to selecting optimal attributes that could better predict customer churn. Five attributes were selected with genetic algorithm (GA), which are as follows, international plan, voicemail plan, total day minutes, total international minutes and customer service. These attributes were considered good based on the hypothesis that states that good attribute sets contain features that are highly correlated with the class, yet uncorrelated with each other [7]. Artificial neural network (ANN) was applied on the selected attributes to develop the prediction model.

METHODOLOGY

In this work secondary data collection technique also referred to as "data mining" was used. CDR data were obtained from Harvard university database https://dataverse.harvard.edu/dataset.xhtml. This work was carried out in 5 stages, starting with the data collection, data pre-processing, genetic algorithm based optimization of attributes, churn predictor using Artificial Neural Network, and finally performance evaluation of the enhanced technique using accuracy benchmark.

The dataset used in this study is the Call Detail Record (CDR) generated by Telecom Italia in the city of Milano. The call detail record (CDR) is a telecommunications system generated metadata of caller behaviour. It, however does not record caller conversation but only contains information necessary for billing and system management such as caller and callee location, time and duration of calls etc. This has been identified as an important data source for quantifying and analysing user experience and behaviour. A CDR is usually anonymised before being released for research, this is done to protect the privacy of callers. This dataset used in this work is sourced from https://dataverse.harvard.edu/dataset.xhtml. Each user records contains 21 attributes that makes it suitable for the research. It was collected during a three-month duration and consists of 5035 call records. The dataset was already anonymized before it was made available by Harvard University for research purposes.

The optimization model used in the work was borrowed from the work of [8]. It uses CFS (correlation-based feature selector) evaluator in the WEKA tool [14], which is a filter algorithm that ranks attributes according to a correlation based heuristic evaluation function.

Equation (1) is known as Pearson's correlation coefficient, it is used to measure the correlation between two variables [10]. It shows that the correlation between attributes and class is a function of the number of attribute variables in the CDR Dataset and the size of the inter-correlations among them. The equation is used in this study as a measure of the merit of attributes selection in classification tasks. For prediction, it is clear that redundant attributes should be eliminated, i.e., if a given attributes predictive ability is covered by another then it can be safely removed.

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k+k(k-1)\overline{r_{ii}}}} \qquad (1)$$

Where:

r_{zc} is the correlation between the summed components and the outside variable

k is the number of components,

$\overline{r_{zi}}$ is the average of the correlations between the components and the outside variables while r_{ii} is the average intercorrelation between component. [10]

PREDICTION MODEL

The ANN tool in WEKA was used in addition to a data set of 5,035 instances, with three (3) hidden layers and an output layer indicating the churn class. The hidden layer of the neural network was defined as "a" where

$$\left(a = \frac{(number\ of\ attributes + number\ of\ classes)}{2}\right).$$

The learning rate of the neural network defines the amount by which the weights are updated and it is set to 0.3 so that the network can converge smoothly as the weight is updated. Using a high value would result to skiping over the global minimums. The momentum of the neural network applied to the weight during updating is set at 0.2 and the error per epoch is set at 0.081. The snapshot of the Network is shown in Fig. 1.

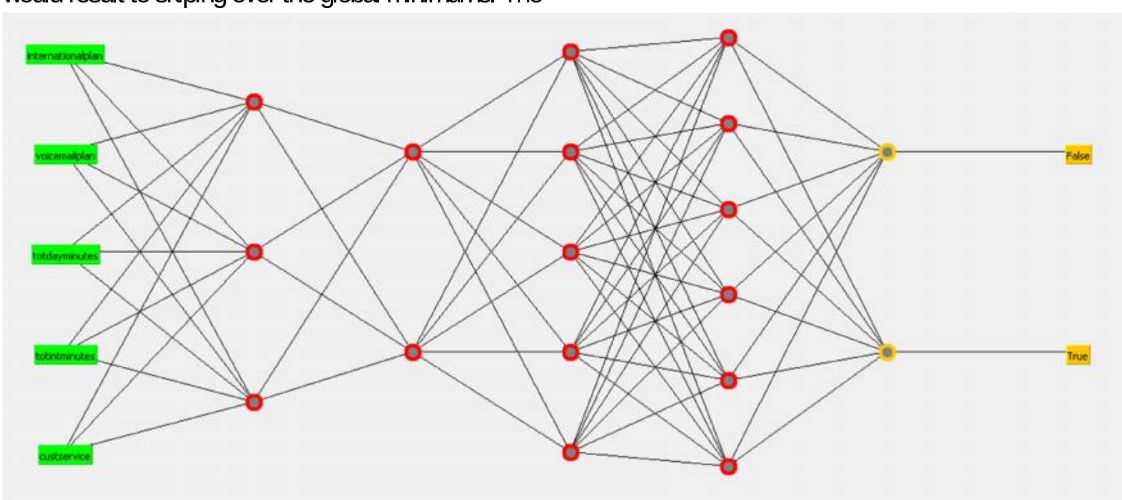


Fig. 1. Snapshot of the ANN showing the prediction of the selected attributes and the class (True or False)

Genetic search method was selected in the WEKA tool to search and optimize the population of the attributes. The following parameters was considered: Crossover probability, Mutation rate, Report frequency.

Cross over probability: Crossover rate is the most important evolution operation. It has been revealed that standard genetic algorithms settings of large population size should have a crossover probability of 0.6 and mutation rate of 0.033 [11]. This produces faster convergence.

Mutation rate: To avoid disruption and steering the algorithm away from the most promising region of the search landscape the mutation rate was set at 0.033.

DATA PRE-PROCESSING:

Attributes selection was done using a supervised attribute filter in WEKA tool, we choose CFS subset evaluator, which evaluates the worth of a subset of attributes by considering the individual predictive ability of redundancy between them. It is capable of evaluating nominal class data, binary class data, date class data, numeric class data, missing class values. Also, Genetic Search was used to optimize the attributes and result was obtained after the search based on the fitness function of the individual attributes for churn prediction.

COMPUTER SIMULATION AND RESULTS

The simulation was done using WEKA software tool installed on a Compaq computer system with Intel(R) Dual Core(TM) i5 CPU T3400 @ 2.16GHz processor and 3GB RAM.

The data was divided into training and testing sets. The training set was 2,533 (58%) instances while the testing set 1,822 (42%) instances. (The training set contains 1692 (93%) instances are labelled Non-Churners and 130 (7%) churners).

Table 1: information on the CDR used in the study**.**

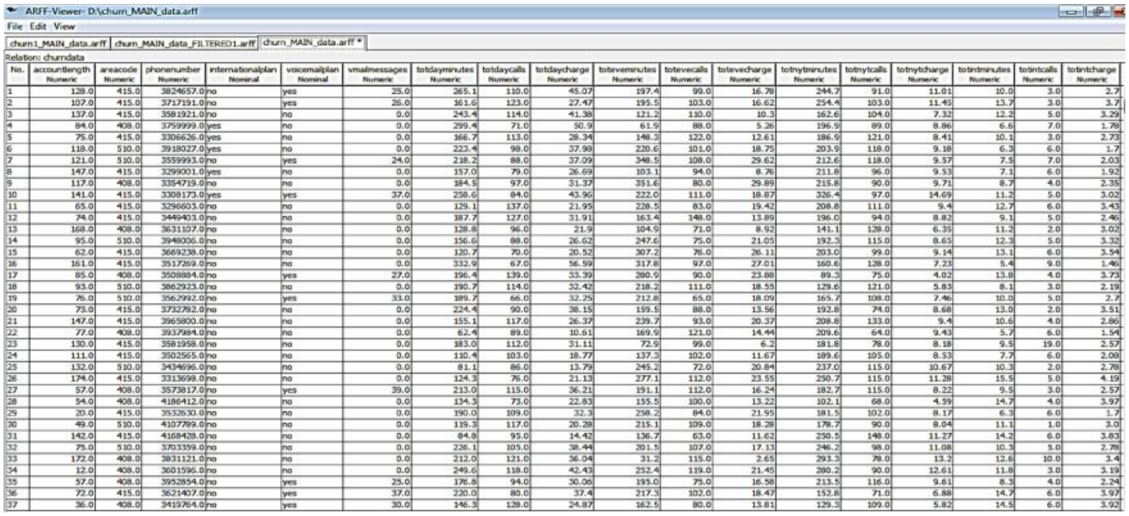
No of attributes per subscriber	Period of data collection	No of subscribers
21	3 months	5035

Table 2: Details of the attributes contained in Call Detail Records

S/n	Attribute	Data type	Attribute Description
1	account length	number	Customer's length
2	area code	number	Customer's area code
3	phone number	number	Customer's phone number
4	International. Plan	number	Customer's international calls
5	voicemail plan	number	Customer's voicemail service
6	vmail messages	number	Customer's voicemail messages
7	total day minutes	number	Customer's total day minutes call
8	total day calls	number	Mean of out minutes of use from other service provider
9	total day charge	Number	mean of all minutes of use from other service provider
10	total eve minutes	number	mean of in minutes of use either from the same or other service provider
12	total eve calls	Number	mean of out minutes of use either from the same or other service provider
13	total night minutes	number	Customer's total night calls
14	total night calls	Nominal	Total amount of night calls by the customer
15	total night charge	Number	The mean number of messages during the period
16	total inter. Calls	Number	Mean monthly revenues
17	total inter. Charge	Number	Associated product (yes or no)
18	customer service	Number	Customer's calls to MNO's customer service
19	State	number	Customer's state of residence
20	customer gender	Nominal	Customer's gender (1=male,0=female)
21	customer's marital status	Nominal	Customer's marital status 1=yes, 0=no)

Source [11]

Fig 2 shows the snapshot of the complete attributes (21) before optimisation, it shows the values of each attributes as used by each subscribers over a period of three. As seen, it contains both numeric and nominal type of data.



ARFF-Viewer- D:\churn_MAIN_data.arff

File Edit View

churn1_MAIN_data.arff | churn_MAIN_data_FILTERED1.arff | churn_MAIN_data.arff *

Relation: churndata

No.	accountlength Numeric	areacode Numeric	phonenumber Numeric	internationalplan Nominal	voicemailplan Nominal	vmailmessages Numeric	totdayminutes Numeric	totdaycalls Numeric	totdaycharge Numeric	toteveminutes Numeric	totevecalls Numeric	totevecharge Numeric	totnytminutes Numeric	totnytcalls Numeric	totnytcharge Numeric	totintminutes Numeric	totintcalls Numeric	totintcharge Numeric
1	128.0	415.0	3824657.0	no	yes	25.0	265.1	110.0	45.07	197.4	99.0	16.78	244.7	91.0	11.01	10.0	3.0	2.7
2	107.0	415.0	3717191.0	no	yes	26.0	161.6	123.0	27.47	195.5	103.0	16.62	254.4	103.0	11.45	13.7	3.0	3.7
3	137.0	415.0	3581921.0	no	no	0.0	243.4	114.0	41.38	121.2	110.0	10.3	162.6	104.0	7.32	12.2	5.0	3.29
4	84.0	408.0	3759999.0	yes	no	0.0	299.4	71.0	50.9	61.9	88.0	5.26	196.9	89.0	8.86	6.6	7.0	1.78
5	75.0	415.0	3306626.0	yes	no	0.0	166.7	113.0	28.34	148.3	122.0	12.61	186.9	121.0	8.41	10.1	3.0	2.73
6	118.0	510.0	3918027.0	yes	no	0.0	223.4	98.0	37.98	220.6	101.0	18.75	203.9	118.0	9.18	6.3	6.0	1.7
7	121.0	510.0	3559993.0	no	yes	24.0	218.2	88.0	37.09	348.5	108.0	29.62	212.6	118.0	9.57	7.5	7.0	2.03
8	147.0	415.0	3299001.0	yes	no	0.0	157.0	79.0	26.69	103.1	94.0	8.76	211.8	96.0	9.53	7.1	6.0	1.92
9	117.0	408.0	3354719.0	no	no	0.0	184.5	97.0	31.37	351.6	80.0	29.89	215.8	90.0	9.71	8.7	4.0	2.35
10	141.0	415.0	3308173.0	yes	yes	37.0	258.6	84.0	43.96	222.0	111.0	18.87	326.4	97.0	14.69	11.2	5.0	3.02
11	65.0	415.0	3296603.0	no	no	0.0	129.1	137.0	21.95	228.5	83.0	19.42	208.8	111.0	9.4	12.7	6.0	3.43
12	74.0	415.0	3449403.0	no	no	0.0	187.7	127.0	31.91	163.4	148.0	13.89	196.0	94.0	8.82	9.1	5.0	2.46
13	168.0	408.0	3631107.0	no	no	0.0	128.8	96.0	21.9	104.9	71.0	8.92	141.1	128.0	6.35	11.2	2.0	3.02
14	95.0	510.0	3948006.0	no	no	0.0	156.6	88.0	26.62	247.6	75.0	21.05	192.3	115.0	8.65	12.3	5.0	3.32
15	62.0	415.0	3669238.0	no	no	0.0	120.7	70.0	20.52	307.2	76.0	26.11	203.0	99.0	9.14	13.1	6.0	3.54
16	161.0	415.0	3517269.0	no	no	0.0	332.9	67.0	56.59	317.8	97.0	27.01	160.6	128.0	7.23	5.4	9.0	1.46
17	85.0	408.0	3508884.0	no	yes	27.0	196.4	139.0	33.39	280.9	90.0	23.88	89.3	75.0	4.02	13.8	4.0	3.73
18	93.0	510.0	3862923.0	no	no	0.0	190.7	114.0	32.42	218.2	111.0	18.55	129.6	121.0	5.83	8.1	3.0	2.19
19	76.0	510.0	3562992.0	no	yes	33.0	189.7	66.0	32.25	212.8	65.0	18.09	165.7	108.0	7.46	10.0	5.0	2.7
20	73.0	415.0	3732782.0	no	no	0.0	224.4	90.0	38.15	159.5	88.0	13.56	192.8	74.0	8.68	13.0	2.0	3.51
21	147.0	415.0	3965800.0	no	no	0.0	155.1	117.0	26.37	239.7	93.0	20.37	208.8	133.0	9.4	10.6	4.0	2.86
22	77.0	408.0	3937984.0	no	no	0.0	62.4	89.0	10.61	169.9	121.0	14.44	209.6	64.0	9.43	5.7	6.0	1.54
23	130.0	415.0	3581958.0	no	no	0.0	183.0	112.0	31.11	72.9	99.0	6.2	181.8	78.0	8.18	9.5	19.0	2.57
24	111.0	415.0	3502565.0	no	no	0.0	110.4	103.0	18.77	137.3	102.0	11.67	189.6	105.0	8.53	7.7	6.0	2.08
25	132.0	510.0	3434696.0	no	no	0.0	81.1	86.0	13.79	245.2	72.0	20.84	237.0	115.0	10.67	10.3	2.0	2.78
26	174.0	415.0	3313698.0	no	no	0.0	124.3	76.0	21.13	277.1	112.0	23.55	250.7	115.0	11.28	15.5	5.0	4.19
27	57.0	408.0	3573817.0	no	yes	39.0	213.0	115.0	36.21	191.1	112.0	16.24	182.7	115.0	8.22	9.5	3.0	2.57
28	54.0	408.0	4186412.0	no	no	0.0	134.3	73.0	22.83	155.5	100.0	13.22	102.1	68.0	4.59	14.7	4.0	3.97
29	20.0	415.0	3532630.0	no	no	0.0	190.0	109.0	32.3	258.2	84.0	21.95	181.5	102.0	8.17	6.3	6.0	1.7
30	49.0	510.0	4107789.0	no	no	0.0	119.3	117.0	20.28	215.1	109.0	18.28	178.7	90.0	8.04	11.1	1.0	3.0
31	142.0	415.0	4168428.0	no	no	0.0	84.8	95.0	14.42	136.7	63.0	11.62	250.5	148.0	11.27	14.2	6.0	3.83
32	75.0	510.0	3703359.0	no	no	0.0	226.1	105.0	38.44	201.5	107.0	17.13	246.2	98.0	11.08	10.3	5.0	2.78
33	172.0	408.0	3831121.0	no	no	0.0	212.0	121.0	36.04	31.2	115.0	2.65	293.3	78.0	13.2	12.6	10.0	3.4
34	12.0	408.0	3601596.0	no	no	0.0	249.6	118.0	42.43	252.4	119.0	21.45	280.2	90.0	12.61	11.8	3.0	3.19
35	57.0	408.0	3952854.0	no	yes	25.0	176.8	94.0	30.06	195.0	75.0	16.58	213.5	116.0	9.61	8.3	4.0	2.24
36	72.0	415.0	3621407.0	no	yes	37.0	220.0	80.0	37.4	217.3	102.0	18.47	152.8	71.0	6.88	14.7	6.0	3.97
37	36.0	408.0	3419764.0	no	yes	30.0	146.3	128.0	24.87	162.5	80.0	13.81	129.3	109.0	5.82	14.5	6.0	3.92

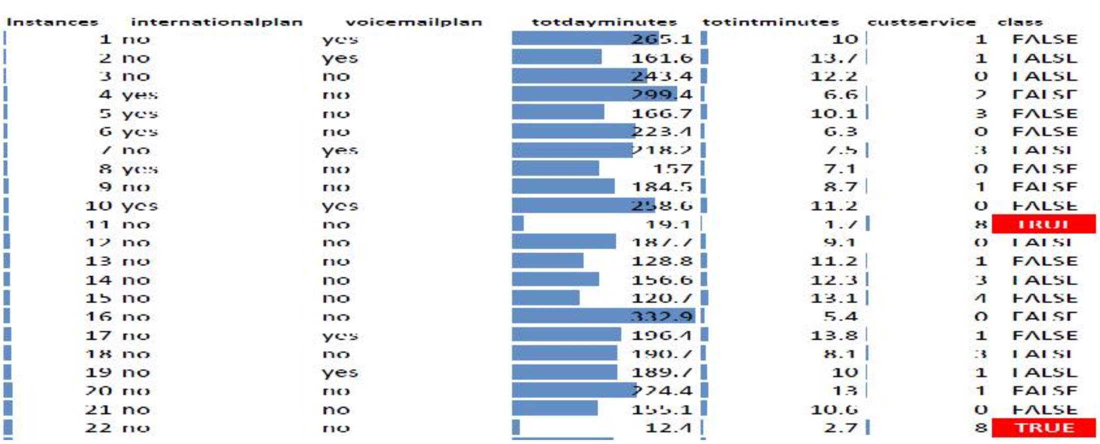
Fig 2, Snapshot of attributes of the data.

Below displays the result obtained after the optimization of the attributes using Genetic algorithm as the search method and CFS (correlation Based Feature Selector) as the evaluator.

Table 2: Optimized attributes.

No of attributes before optimization	**No of attributes after optimization**
21	5(Five) 1. international Plan 2. Voicemail Plan 3. Total day minutes 4. Total International minutes 5. Customer service

Fig 3 shows the detailed description of the optimized attributes. It shows the five selected attributes and their churn class (true or false). Evidence from the attribute selection shows that, along with irrelevant features, redundant information as related to the class was eliminated. An attribute is said to be redundant if one or more of the other features are highly correlated with it.



Instances	internationalplan	voicemailplan	totdayminutes	totintminutes	custservice	class
1	no	yes	265.1	10	1	FALSE
2	no	yes	161.6	13.7	1	FALSE
3	no	no	243.4	12.2	0	FALSE
4	yes	no	299.4	6.6	2	FALSE
5	yes	no	166.7	10.1	3	FALSE
6	yes	no	223.4	6.3	0	FALSE
7	no	yes	218.2	7.5	3	FALSE
8	yes	no	157	7.1	0	FALSE
9	no	no	184.5	8.7	1	FALSE
10	yes	yes	258.6	11.2	0	FALSE
11	no	no	19.1	1.7	8	TRUE
12	no	no	187.7	9.1	0	FALSE
13	no	no	128.8	11.2	1	FALSE
14	no	no	156.6	12.3	3	FALSE
15	no	no	120.7	13.1	4	FALSE
16	no	no	332.9	5.4	0	FALSE
17	no	yes	196.4	13.8	1	FALSE
18	no	no	190.7	8.1	3	FALSE
19	no	yes	189.7	10	1	FALSE
20	no	no	224.4	13	1	FALSE
21	no	no	155.1	10.6	0	FALSE
22	no	no	12.4	2.7	8	TRUE

Fig. 3: Result showing the selected attributes and class (Churn)

Fig 5 shows the plot of the concentration points of the prediction of customer service attribute, where blue color indicates the Non churners and red color indicates Churners. Considering only the customer service attribute, it shows that a total of 194 subscribers churned representing 10% while 1628 are Non churners representing 90% of the subscribers.

Fig 5.Plot showing the concentration points of the prediction of customer service attribute.

Fig 6 shows the plot of the concentration points of the prediction of international minutes attribute, where blue color indicates the Non churners and red color indicates Churners. From the plot, it shows that a total of 175 subscribers representing 10% are likely to churn based on the number of their total international minutes while 1647 are representing 90% of the subscribers are not likely to churn.

Fig 6. Plot showing the concentration points of the prediction of total international minute attribute.

Fig 7 shows the plot of the concentration points of the prediction of Total day minutes attribute, where blue color indicates the Non churners and red color indicates Churners. It shows that a total of 142 subscribers representing 7% are likely to churn based on the number of their total day minutes while 1680 representing 93% of the subscribers are not likely churners.

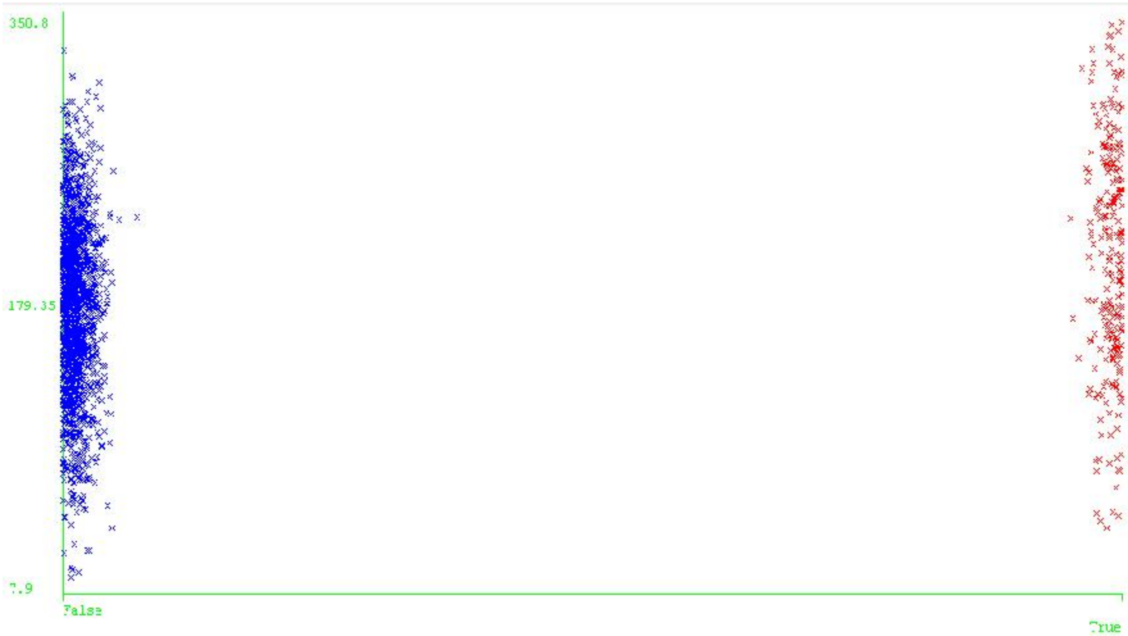


Fig 7.Concentration points of the prediction of total day minutes attribute.

The improved churn prediction accuracy was obtained using 1,822 instances and 1692 (93%) instances are labelled Non-Churners and 130 (7%). Table 3. Shows the comparison with existing techniques and the improved technique gives a higher accuracy as compared to other techniques.

Table 3: Comparison of the improved churn prediction accuracy with existing churn prediction

	Existing Churn Prediction Accuracy	**Improved Churn Prediction Accuracy**
Decision Tree	89.6%	93.2 %
Logistics	86.8 %	93.2 %
Support vector machine	88.9%	93.2 %

The accuracy measurement was used to evaluate the performance of the technique. Table 4 shows the result for the accuracy measure. The improved technique shows a higher accuracy which is, this is attributable to the fact that the improved technique combined genetic algorithm and artificial neural network for the optimization and prediction of the attributes.

Table 4: Analysis of the improved technique.

True positive rate	False positive rate	Precision	Class
0.962	0.035	0.96	False
0.035	0.043	0.727	True

CONCLUSION

Decision tree, Logistics regression, and support vector machine are previous prediction techniques used to date, but with low accuracy. However the utilization of genetic algorithm to optimize the attributes and artificial neural network for the prediction and classification in this work shows a higher accuracy. A data set of 4,355 instances with 21 attributes is used to train and test the model. Using 4 different techniques which are decision tree, logistics regression, support vector machine and the improved technique of the utilization of Genetic algorithm and Artificial Neural network to optimize and predict, indicate that the improved technique gives the best accuracy. The next stage of the research is to use other data mining tool to carry out the research.

REFERENCES

[1] R. W. D. B. G. J. a. H. K. M. C. Mozer, ""Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry," *IEEE Trans,* vol. 2, no. 6, pp. 690-696, 2007.

[2] j. H. (. Ahn, "Customer churn analysis : Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry," *telecommunications Policy.,* vol. 7, no. 3, pp. 552-568, 2006.

[3] C. R. Y. B. DongBack Seo, "Two-level model of customer retention in the US mobile telecommunications service market," *Elsevier,* vol. iv, pp. 182-196, 2014.

[4] N. B. o. statistics., "Telecommunication report in Nigeria. Abuja:," National Bureua of Statistics NBS, Abuja, 2016.

[5] ". C. 2. Orange, "Orange, "KDD Cup 2009," Orange Telecom, may 2009. [Online]. Available: <http://www.kddcup-orange.com/>. [Accessed january 2016].

[6] J. T. Hadden, "Churn prediction: Does technology matter," *International Journal of Intelligent Technology,* vol. 1, pp. 1414-1425, 2006.

[7] D. R. Seo, "Two - level model of customer retention in the Us mobile telecommunications service market telecommunication policy," *Telecommunication statistics,* vol. 32, no. 11, pp. 182-196., 2008.

[8] R. G. a. S. Meher, "Customer churn prediction in telecommunications,School of Computer Science and Informatics," *IEEE,* vol. 4, no. 09, pp. 114-127., 2012.

[9] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," The University of Wakaito, Newzealand, 1999.

[10] E. E. Ghiselli, Theory of Psychological Measurement., New york: McGrawHill, 1994.

[11] Y. W. a. I. H. Witten., "Inducing model trees for continuous classes.," in *Ninth European Conference on Machine learning*, Amsterdem, 2008.

[12] Y. H. A. K. M. N. Essam Shaaban, "A Proposed Churn Prediction Model," *International Journal of Engineering Research and Applications (IJERA),* vol. 2, no. 4, pp. 693-697, 2012.

[13] R. K. G. a. S. K. Meher, "Customer Churn Time Prediction in Mobile Telecommunication Industry Using Ordinal Regression,," *Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining (PAKDD '08),* pp. 884-889, 2008.

[14] http://www.cs.waikato.ac.nz/ml/weka

