The International Journal of Engineering and Science (IJES)
|| Volume || 8 || Issue || 8 Series II || Pages || PP 55-61 || 2019 ||
ISSN (e): 2319 – 1813 ISSN (p): 23-19 – 1805

Application Programming Interface (API) for Real time Churn Prediction Based on Customer Classification using Decision Tree and k-nearest neighbor

Turiho Jean Claude1, 2, Kipruto W. Cheriuyot1, Mutoni Anne Kibe1, Uwitoze Alfred3

[1]Jomo Kenyatta University of Agriculture and Technology (JKUAT) Nairobi, P.O. Box 62 000, Kenya
[2]University of Lay Adventists of Kigali (UNILAK) Kigali, P.O. Box 6392, Rwanda
[3]University of Rwanda (UR) P.O. Box 4285, Kigali
Corresponding Author: Turiho Jean Claude

--ABSTRACT---

In Telecommunication Industry, customers are considered as one of the most important asset for several companies within a marketplace. Customer churn prediction is live topic for both decision makers and researchers in data mining. One of the reason is that customer's behaviors are keeping changing with time and injection of new competitors, and new products. In this perceptive, using different algorithms on the same customers and data, accuracy statistics show different levels for different datasets. In this paper, a novel Customer Churn Prediction approach is presented based on working with real and historical data by applying Decision tree and k-nearest algorithms. This novel approach let knowing real information about customer behaviors by taking data from the source- switch. API pulls CDR data converted in CSV format in a file which is exported to any DBMS application where all Data mining steps are respected using Knime analytics platform. In this paper, two Data mining algorithms: Decision tree and K-nearest. In fact, the API between switch and CSV file is in charge to trigger two most important actions: (i) converting CDR data format to CSV data, (ii) pulling data in CSV file after deleting its existing content (iii) refreshing after a defined time. Accuracy statistics measures used are True Positives, False Positives, True Negatives, False Negatives, Recall, Precision, Sensitivity, Specifity, F-measure, Accuracy, and Cohen's Cappa. They have shown good and accuarate results in real time.

***Keywords**: Churn, non-Churn, Call Data Record, Churn Prediction, decision tree, K-nearest.*

Date of Submission: 21-08-2019 Date of acceptance: 05-09-2019

I. INTRODUCTION

The customer are considered as one of the most important asset for a business in numerous dynamic and competitive companies within a marketplace. In competitive market, companies in which the customers have numerous choice of service providers, they can easily switch a service or even a provider [1]. Such kind of customers are referred to as churned. There are multiple reasons of churning. In [2] is listed some of them: dissatisfaction, higher cost, low quality, lack of features and privacy concerns. Churn phenomenon appears in a big number of organizations: financial services, airline ticketing services, social network analysis, online gaming, banking sector and telecommunication industry [1]. Prediction churn studies purpose in all above organizations is to establish and maintain the long-term relationship with their existing customers [3]. According to some authors [4], loyal customers are not only qualified as profitable for the company but also as best marketing agent in the marketpalce. Currently, telecommunication industries store a big amount of information about their customers such as local/international call records, short messages, voice mail, demographics, financial details, and other usages behavior of the customers [1]. Since long time, based on customers information stored by telecommunication industry, data mining algorithms and a lot of approaches have been develeped nd applied in order to predict if customer is about to churn or not. In [1], authors concluded that there is a greater challenge of which data mining techniques and algorithms can be chosen to build customer churn prediction model. One of the multiple reasons is that maybe churn and non-churn customers in the time behaviors are almost similar. This similarity is the source of increasing the classification error rate. If is the case, some decision taken are wrong at a given instance. The approach we propose is Real time customer churn prediction model using decision tree and k-nearest neighbor algorithms apart focusing on local/international call records, short messages, voice mail, demographics, financial details, is especially focusing on Competitor calls.

II. RELATED WORK

This section focuses on related works that help to explore the state-of-the-art tehcniques that have been developped for customer churn prediction.

Social interactions have been a discussed topic in [8]. Author discussed the dynamics of social interactions for customer churning within the telecommunication networks. He associates customer's churn behavior with incoming and outgoing calls. The study conluded that the customer is likely to churn if this one maintains a relationship with a customer who has churned from his/her actual network..

In Prediction of Customer Attrition of Commercial Banks based on SVM Model [5], authors proposed a methodology based on SVM classifier and random sampling techniques. In order to minimise the imbalance class problem which is caused due to the lack of availability of data the random sampling has been used to change data distribution.

2015 a new custpmer churn prediction model has been preposed basically focusing to identify most likely customers exhibiting churn behaviors [6]. This model is basically based on simple k-means for clustering and applied JRip for rules extractions.

Customer classification model based on rough set theory was proposed in 2017 by [7]. This model has the purpose of effeciently classifying customer churn. Authors have expressed that rough set therory based classification model can be better than linear regression, J48, voted perception of neural etwork.

From above authors, a good work was done in customer churn prediction in telecommunication industry. However, no one of all developed models to be considered as standard model to overcome customer churn prediction in more appropriate fashion. On this topic: which approach is the standard one, authors in [1] have established a long list of discussions on customer churn prediction. From all developed models and discussions on the existing models we concluded that there is no study which has focused on competitors calls as one factor influencing churn for building a model. In following section, methodology used is described.he context of churn management, predictive modeling uses historical transactions and characteristics of a customer to predict future customer behavior. Churning customers can be divided into two main groups, voluntary and non-voluntary churners. Non-voluntary churners are the easiest to identify, as these are the customers who have had their service withdrawn by the company. Voluntary churn is more difficult to determine, because this type of churn occurs when a customer makes a conscious decision to terminate his/her service with the providerIn the context of churn management, predictive modeling uses historical transactions and characteristics of a customer to predict future customer behavior. Churning customers can be divided into two main groups, voluntary and non-voluntary churners. Non-voluntary churners are the easiest to identify, as these are the customers who have had their service withdrawn by the company. Voluntary churn is more difficult to determine, because this type of churn occurs when a customer makes a conscious decision to terminate his/her service with the provider

III. METHODOLOGY

In this section, we develop the empirical study in details. (1) Explain the problem statement. (2) Develop in details the empirical steps. And (3) Evaluation setup.

3.1 Problem stament

Customer churn prediction is the answer to the problem of classifying customer loyalty into classes: Class one: churn, and class two: non-churn.According to [1] voluntary churns are difficulty to predict while involuntary churns are easier to predict by using simple queries. On other hand, litterature revealed that existing studies have been published but still there is no agreement on awarding the best approach to solve customer churn prediction. Indifferent results shown by different classification techniques might the source.

Data mining studies analyzing interpersonal relationships and social contagion, however, concentrated to a significant extent on the relevance of social interactions in the process of client acquisition and depended on the use of undirected networks. But no data mining studies focused on customer churn prediction behaviors based on the importance of social interactions within a directed network.

3.2 Empirical Setup

In this study, we designed a customer churn prediction model where we used a combination of live and historical data and especially focusing on information including competitor calls among customer information to be analysed.

We used public data from the kaggle website for this research. After receiving the dataset, a simulation was performed to fill in our database in real time.

1.2.1 Data acquisition

The following scenario shows how data are acquired figure 1. Acquiring data is the first and most important phase in our model. Data are directly captured from telecommunication company switch. Call Data Record are raw data for this model. Then, after being converted are loaded into a database. Helped by API, data are pushed into CSV file after deleting the old ones by a trigger.

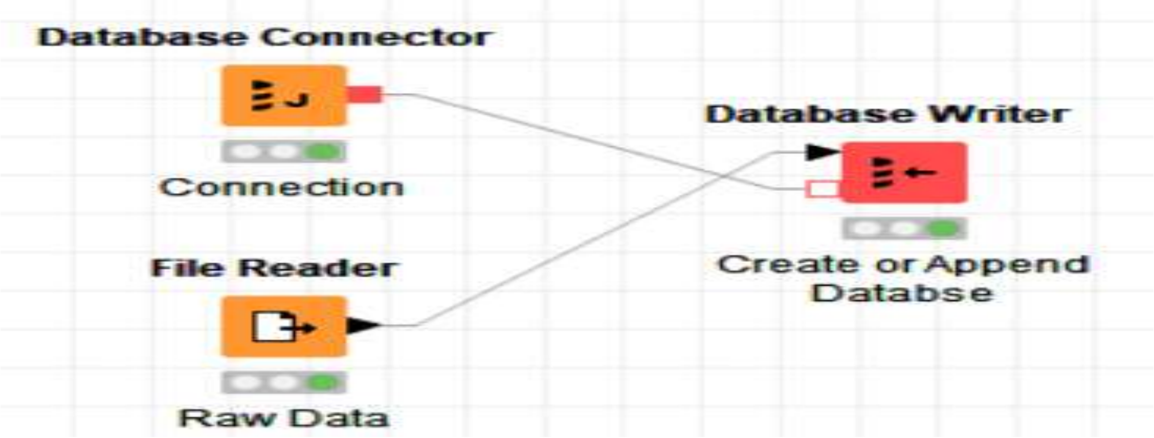


Figure 1 Data acquisition

1.2.2 Data preprocessing

First of all, we have done attributes selection on all training data. Weka's ReliefFAttributeEval evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class. Can operate on both discrete and continuous class data. We fount the following:

Table 1 Ranked Attributes

Ranked attributes	
1.0000	CompetitorsCalls
0.2516	MonthlyRevenue
0.0015	MonthsInService
0.0010	TotalRecurringCharge
0.0009	OverageMinutes
0.0005	Occupation
0.0005	OffPeakCallsInOut
0.0004	PercChangeMinutes
0.0004	PercChangeRevenues
0.0003	ReceivedCalls
0.0003	OutboundCalls
0.0002	MonthlyMinutes
0.0002	RoamingCalls
0.0002	InboundCalls
0.0002	DirectorAssistedCalls

Fifteen attrabutes have been selected for this model as listed in table 1.

Then attributes are filtered according to above results, rows are sampled relative 60% and stratified sampling have used.

Third, missing values (if any) are handled using mean in case of number values and fixed value in case of string values.

Finally, data are splitted into train data and test data. The splitting of dataset into training and test sets is a common process of data mining algorithms for building predictive model [8]. Usually train data is used to train the model while test set is used to measure how well the proposed model has been trained (model performance evaluation). We suppose that the test data is fresh data where the class label value is obtained from the proposed predictive model. We then gathered the predictions result on the inputs from the test set from the qualified classifier and then compared them to obtain the empirical results of these test data.

All above preprocessing steps are summarized in figure 2.

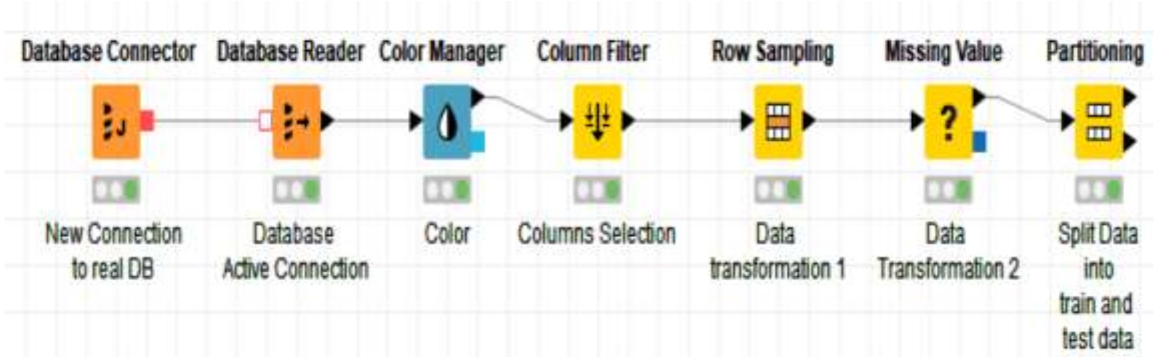


Figure 2 Data preprocessing phase

1.2.3 Customer Churn Prediction model overall

This process allows us to evaluate the performance of the proposed model on the given test data.

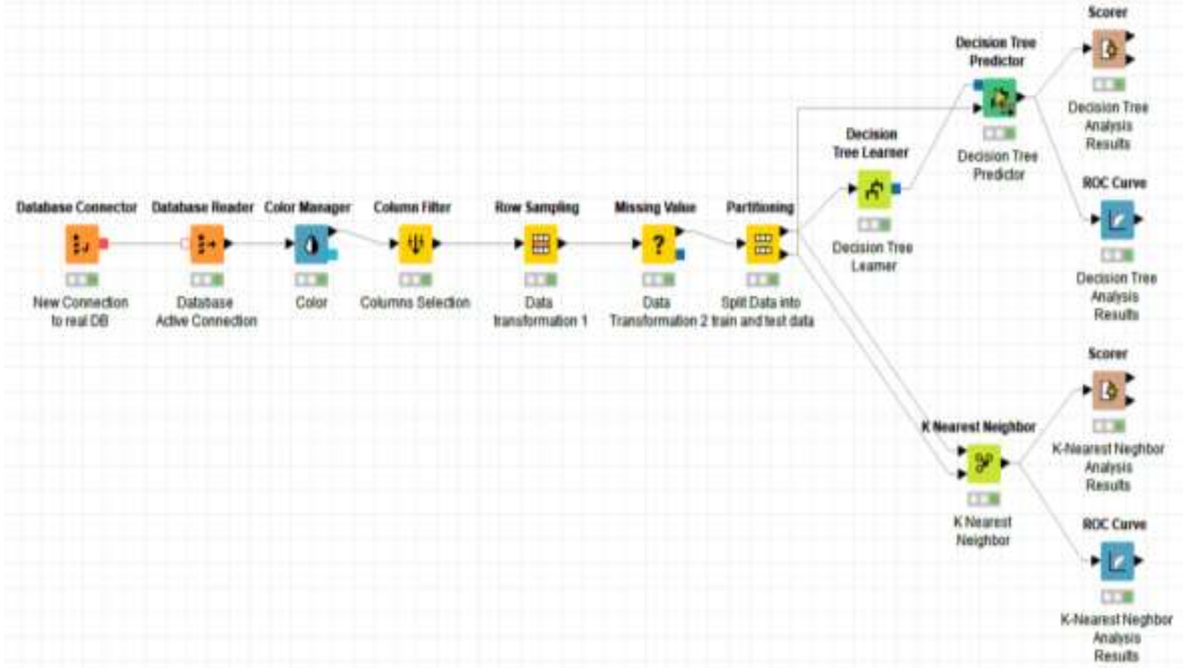


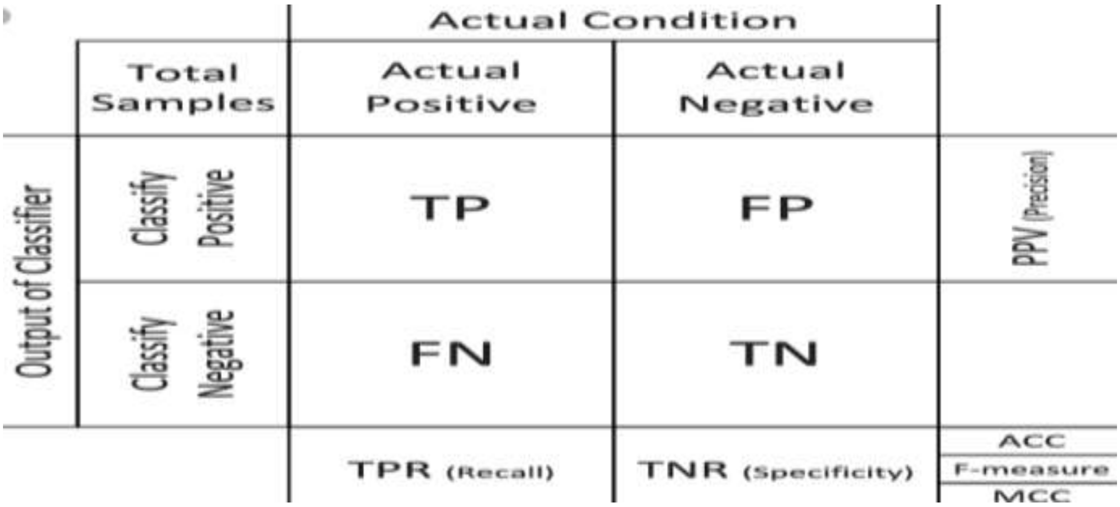
Figure 3 Customer Churn Prediction Evaluation Step

This model has used two data mining algorithms: decision tree and k-nearest neighbor. For organizing classifiers and visualizing their performance two techniques are used: receiver operating characteristics (ROC) graphs and scores for decision tree. On the other hand scatter matrix and scorer.

IV. RESULTS AND DISCUSSION

In part, we examined and assessed the findings of the proposed customer churn prediction model through state-of - the-art evaluation measures (True Positives, False Positives, True Negatives, False Negatives, Recall, Precision, Sensitivity, Specifity, F-measure, Accuracy, and Cohen's Cappa) as indicated in the following table.

Table 2 Confusion matrix and relevant evaluation index



		Actual Condition		
	Total Samples	Actual Positive	Actual Negative	
Output of Classifier	Classify Positive	TP	FP	PPV (Precision)
	Classify Negative	FN	TN	
		TPR (Recall)	TNR (Specificity)	ACC F-measure MCC

The above table contains all accuracy statistics used for this study. The mathematical equations of these evaluation measure given

$$Precision = \frac{TP}{TP+FF} \qquad (1)$$

$$Recall = \frac{TP}{TP+FN} \qquad (2)$$

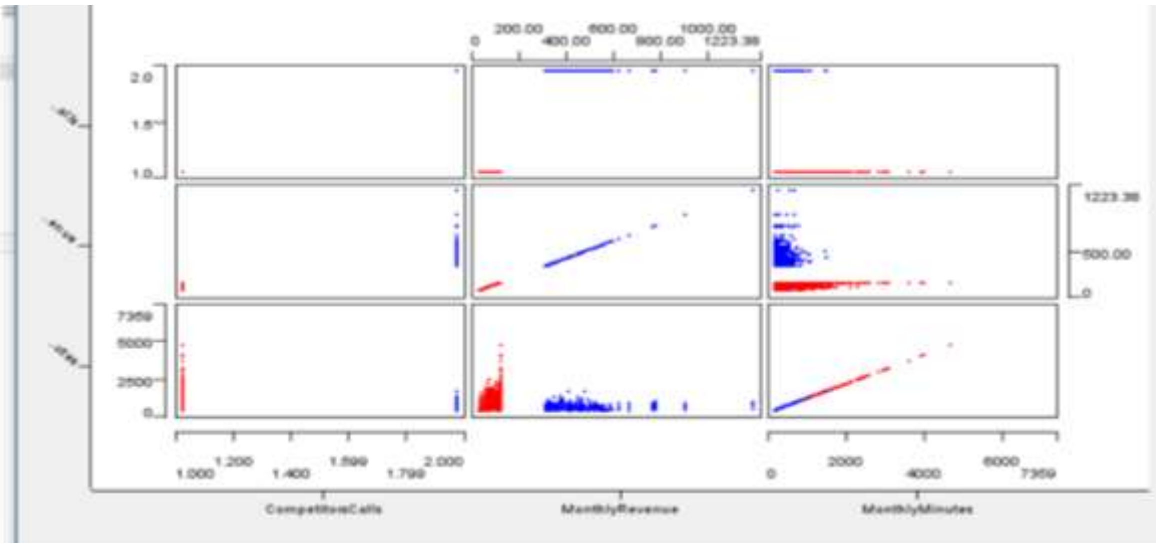
$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (3)$$

$$F-measure = 2 \times \frac{Precision \times Recall}{Precision+Recall} \qquad (4)$$

$$Specificity = \frac{TN}{TN+FP} \qquad (5)$$

$$Sensitivity = \frac{TP}{TP+FN} \qquad (6)$$

The equation (1) mathematically expresses the precision measure, used to evaluate the correct degree of prediction power of the proposed model. Then, equation (2) represents the recall measure, which is very important because prediction models intend to predict true churn customers as much as possible. However, there exists trade-off between precision and recall. Therefore, a comprehensive measure is required for precision and recall. Here, we use equation (4) (F-measure) that calculates the harmonic mean of these two measures (e.g., precision and recall) resulting in achieving the balance between the said trade-off. Accuaracy represented in equation (3), its numerical value represents the proportion of true positive results (both true positive and true negative) in the selected population. Specificity equation (5) measures the proportion of actual negatives that are correctly identified as such. And finally, sensitivity measure in equation (6) which is very important because prediction models intend to predict true churn customers as much as possible.



9

Figure 4 Scatter Matrix (K-nearest neighbor)

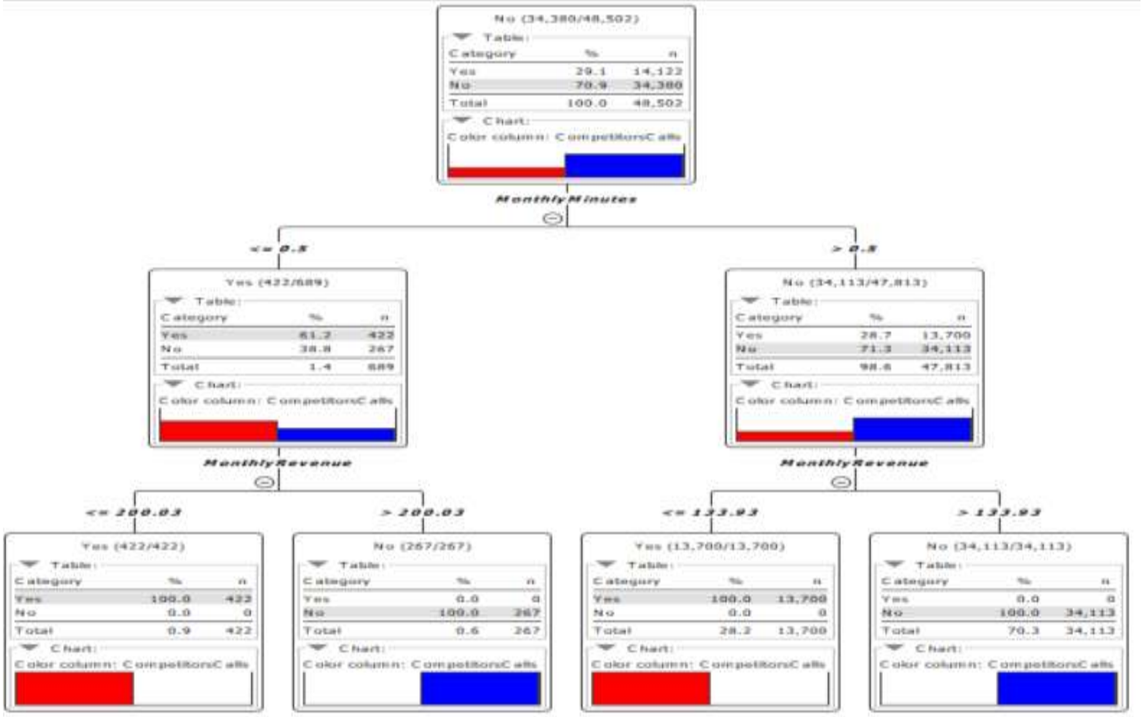


Figure 5 Decision tree learner

Table 3 Accuracy Statistics

row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy	Cohen's kappa
No	1606	142	158	94	0.94	0.92	0.94	0.53	0.93		
Yes	158	94	1606	142	0.53	0.63	0.53	0.94	0.57		
Overall										0.882	0.505

The results presented in figures 4-5, and in table 3 captured at given time demonstrate that customer churn prediction in real time using decision tree and k-nearest neighbor is useful because all accuracy measures applicable.

The proposed study demonstrate the benefit of incorporating analysis into real data combined with the historical data. Often, the researchers focused on a model which achieved higher accuracy rather than focusing on working with combination of real and historical data in predicting the Churn behavior. The PMML model is applied on real data integrated into existing data stored in data base.

V. CONCLUSION

With the rapid growth of digital data and associated technologies, there is an emerging trend, where industries become rapidly digitized. These technologies are providing great opportunities to identify and resolve diffuse problem of customer churn, particularly in telecommunication industry. Among these new technology we have chosen to use KNIME analytics platform. This platform facilitates the combination of real and historical data. It has integrated python and weka to its features for data analytics on the other hand.

The study proved that a customer who is intended to call competitors networks ia likely to churn at 99.25%.

REFERENCES

[1]. F. A.-O. e. a. Adnan Amin, "Customer Churn Prediction in telecommunication Industry using data certainty," Journal of business Research, pp. 290-301, 2018.

[2]. R. S. R. S. Riddhima, "Evaluating Prediction of Customer Churn Behavior Based On Artificial Bee Colony Algorithm," International Journal of Engineering and Computer Science, vol. 6, no. 1, 2017.

[3]. F. A.-O. B. S. M. A. T. C. K. H. U. R. D. S. A. Adnan Amin, "Just-in-time customer churn prediction in the telecommunication sector," The Journal of Supercomputing, 2017.

[4]. J. A. M. a. R. K. Ganesh, "Understanding the Customer Base of Service Providers: An Examination of the Differences between Switchers and Stayers.," urnal of Marketing, pp. 65-87, 2000.

[5]. H. S. Q. W. X. Benlan He, "Prediction of Customer Attrition of Commercial Banks based on SVM Model," Procedia Computer Science, vol. 31, pp. 423-430, 2014.

[6]. A. O. O. &. A. Adeyemo, "Customer Churn Analysis In Banking Sector Using Data Mining Techniques," African Journal of Computing & ICT , vol. 8, no. 3, pp. 165-174, 2015.

[7]. S. N. M. A. M. M. K. A. M. N. A. R. a. M. D. M. Makhtar, "CHURN CLASSIFICATION MODEL FOR LOCAL TELECOMMUNICATION COMPANY BASED ON ROUGH SET THEORY," Journal of Fundamental and Applied Sciences, pp. 854-868, 2017.

[8]. MichaelHaenlein, "Social interactions in customer churn decisions: The impact of relationship directionality," International Journal of Research in Marketing, vol. 30, no. 3, pp. 236-248, 2013.

[9]. E. F. M. A. H. Ian H. Witten, "Data mining : practical machine learning tools and techniques," pp. 587-605, 2011.

Turiho Jean Claude A computer science engineer, a data analyst and PhD student in the Department of Computing at the Jomo Kenyatta University of Agriculture and Technology, MSc. in Computer Science and Information Technology (Hunan University-China), BSc. In Information Management (AUCA-Rwanda). His principle research domains are machine learning, business intelligence. He currently teachs at UNILAK and has then years' experience in Data analysis. Published papers: An optimal class association rule algorithm.

Dr. Kipruto Wilson Cheruiyot is the current director of Jomo Kenyatta University of Agriculture and Technology Kigali Campus-Rwanda. He has the following academic qualifications Bsc (Hons); PGD-E (Egerton University, Kenya); Msc in Computer Application and Technology (central south university of technology Hunan, china); PhD in Computer Application and Technology (Central South University, China).

Dr. Ann Muthoni Kibe, Lecturer in School of Computing and Information Technology at Jomo Kenyatta University of Agriculture and Technology.

Dr. UWITONZE Alfred, Lecturer in College of Science and Technology at University of Rwanda.

Turiho Jean Claude " Application Programming Interface (API) for Real time Churn Prediction Based on Customer Classification using Decision Tree and k-nearest neighbor" The International Journal of Engineering and Science (IJES), 8.8 (2019): 55-61