# Customer Churn Prediction in Banking Industry Using K-Means and Support Vector Machine Algorithms

Micheal Arowolo
University of Missouri
71 PUBLICATIONS 489 CITATIONS
SEE PROFILE

Bilkisu Jimada-Ojuolape
Kwara State University
12 PUBLICATIONS 167 CITATIONS
SEE PROFILE

Saheed Yakub
Al-Hikmah University,Ilorin, Nigeria
25 PUBLICATIONS 162 CITATIONS
SEE PROFILE

Contents lists available at Crown Academic Publishing

International Journal of Multidisciplinary Sciences and Advanced Technology

Journal home page: http://www.ijmsat.com

# Customer Churn Prediction in Banking Industry Using K-Means and Support Vector Machine Algorithms

Abdulsalam Sulaiman Olaniyi [1], Arowolo Micheal Olaolu [1], Bilkisu Jimada- Ojuolape [2], Saheed Yakub Kayode [3]

[1] Department of Computer Science, Kwara State University Malete, Nigeria

[2] Department of Electrical and Computer Engineering, Kwara State University, Malete, Nigeria.

[3] Department of Computer Computer Science, Al-Hikmah University, Ilorin, Nigeria.

Corresponding Author: abdulsalamny@gmail.com, arowolo.olaolu@gmail.com

Original article Received 1 January 2020, Accepted 15 January 2020, Available online 18 January 2020

**ABSTRACT**

This study proposes a customer churn mining structure based on data mining methods in a banking sector. This study predicts the behavior of customers by using clustering technique to analyze customer's competence and continuity with the sector using k-means clustering algorithm. The data is clustered into 3 labels, on the basis of the transaction in and outflow. The clustering results were classified using Support Vector Machine (SVM), an Accuracy of 97% was achieved. This study enables the banking administrators to mine the conduct of their customers and may prompt proper strategies as per engaging quality and improve proper conducts of administrator capacities in customer relationship.

.

**Keywords:** Customer Churn, Banks, K-Means and SVM.

## 1. Introduction

Investigating customer churn for huge data in customer retention is an open research in machine learning technology. Customer churn is the loss of customers who changes from one sector, for example, a bank, telecommunication network, among others, to another contender within a certain time. Customer churn misclassification utilizing clustering can prompt massive financial losses and even hurt the association's development (Rohini, and Devaki, 2017). Customer churn management is critical, particularly for businesses like the banking industry where information are been utilized, analysis of huge data having substantial number of repetitive, redundant and noisy information must be wiped out (Yi, Qixin, Chongqing, and Qing, 2016). Data mining clustering system has been an answer for producing the prediction of the results by utilizing data organization and ranking strategies which can generate solutions. Customer relationship board strategizes, oversee and examine customer connections and information all through the lifecycle of the customer in the organization, with the intension of managing the relationships with customers in terms of business and driving transactions development. Analyzed data of customers are gathered to make proper and supportive choices. Customer churn will result in the loss of organizations; customer churn predicts the individuals who are going to churn. Customer churn prediction has gotten a developing consideration amid the most recent decade. As one of the critical measures to retain customers, churn prediction has been an apprehension in the financial and research world (Wang, and Chen, 2010). Recently, various data mining technique have been received for churn prediction, including customary statistical techniques, for example, logistic regression, non-parametric statistical models like for example k-nearest neighbor, decision trees, and neural networks (Guangqing, and Xiuqin, 2012). In this investigation, mining a customer behavior is examined by utilizing k-means clustering procedure and SVM classification based on customer's features, to help the financial industries to recognize distinctive types of customers and the churn behaviors

## 2. Literature Review

Amjad, Reham, Osama, Ruba, and Hossam (2015), proposed a hybrid data mining learning approach for foreseeing customers churn. In their work, three models were carried by stages the clustering and predicting the performance. Customers' information was filtered utilizing the k-means algorithm and Multilayer Perceptron Artificial Neural Systems (MLP-ANN) for prediction. Using the clustering with MLP-ANN, the model uses self-sorting out maps (SOM) with MLP-ANN on the data, the precision and churn rate values were determined and compared with other state-of-art, their work demonstrated that the three-crossover model outperforms single normal and common models.

Wenjie, Meili, Mengqi and Guo (2018) proposed up a clustering algorithm called semantic driven subtractive clustering technique (SDSCM) using a Hadoop map reduced structure. The proposed model ended up being fast, compared with different techniques and gave some showcasing procedures as per clustering algorithm to guarantee benefit amplification.

Fathian, Hoseinpoor, and Minaei (2016) compared single standard classifiers and ensemble classifiers for churn prediction, they built up an aggregate of 14 prediction models that were grouped in four classifications: fundamental Classifier; (Decision Tree, Artificial Neural Network, K-Nearest Neighbor, SVM), Classifier with SOM + basic classifier, Classifier with SOM + reducing features with PCA + basic classifier; and Classifier with SOM +reducing features with PCA +bagging and boosting ensemble classifier.

Alwis, Kumara and Hapuarachchi (2018) assert with the capacity to recognize potential churn customers, cluster customers with comparative conduct and mine the applicable examples embedded in the gathered information. The essential information gathered from clients were utilized to make a prescient churn model that get customer churn rate of five transmission industries. Four model structure, classified the applicable factors with the use of the Pearson chi-square test, cluster analysis, and association rule mining. Using WEKA, the cluster results delivered the customers involvement, premium regions and purposes behind the churn decision to improve promotion and marketing activities. Using the rapid miner, the association rule mining with the FP-development component was communicated rules to recognize intriguing quality examples and patterns in the gathered information affect the incomes and development of the media transmission organizations. At that point, the C5.0 decision tree, Bayesian network, logistic regression, and the neural network algorithms were developed using the IBM SPSS Modeler 18. A comparative evaluation was performed to find the optimal model and test the model with precise, predictable and dependable outcomes.

Hossam (2018) proposed a swarm intelligent neural network model for customer churn prediction, using particle swarm optimization and feed-forward neural network. The PSO is used to tune the loads of the input features and improve the structure of the neural network simultaneously to build the prediction control. The proposed model handles the imbalanced class dissemination of the information using a progressed oversampling strategy. The assessment results demonstrated that the proposed model can improve the inclusion rate of churn customers comparing with other cutting-edge classifiers. In addition, the model has high interpretability, where the allocated features can indicate the significance of their corresponding features in the class.

## 3. Methodology

The dataset used in this investigation was provided by a noteworthy Nigerian Bank industry (Ecobank). The dataset contains 5 attributes properties of haphazardly chosen 200 customers within a period of three months. The attributes cover the transaction and details of customers transaction flow, with the data shows that a customer is either churned (left the bank) or still active.

The churn prediction model carries out two procedures. The first procedure uses Clustering algorithm on the data, to separate customers to groups that represents various practices using the k-means algorithm. In the second procedure, SVM is applied on the clustered result for classification. The performance of the developed model is evaluated and assessed using the assessment criteria. The proposed system for customer churn is shown in figure 3.1 below.

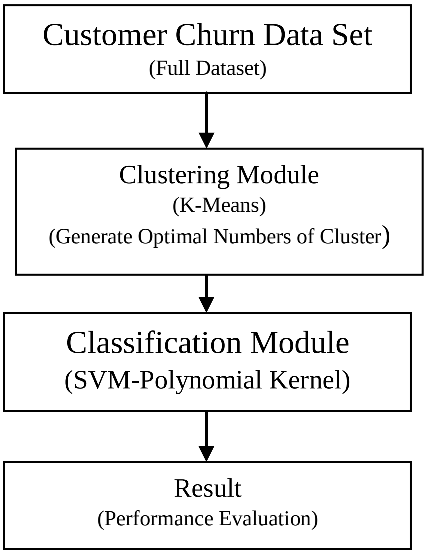

**Figure 1.** Customer Churn Framework

### 3.1 Clustering

Clustering is used to arrange an extensive document accumulation into unmistakable gatherings of comparative reports. It perceives general topics covered up inside the corpus. Uses of document clustering go past arranging document accumulations into information maps. This can encourage ensuing learning recoveries and accesses. Document clustering for instance, has been used to improve the proficiency of text arrangement and find occasion scenes in transiently requested records. Likewise, rather than showing query items as one not insignificant list, some earlier investigations and rising search engines utilize an archive clustering approach to deal with consequently compose list items into important classifications and in this manner supports cluster-based perusing (Philip, and Thomson, 2017).

#### 3.1.1 K-implies bunching

K-means is a well-known and significant clustering technique that was exhibited by Mc. Queen in 1967. The fundamental stride in K-mean clustering includes; the initial step that selects k objects that have their middle (mean). In this strategy the rest of the s are not chosen yet are doled out to group regarding the similitude of the item with bunch. These relationships are estimated for the sake of the distance between cluster means and objects after computation, the new focus point is determined for the sake of facts. These steps are repeated until the required function is accomplished. In k-means clustering, the most critical point is to discover the numbers of cluster that is optimum as the separation between cluster means and objects. The algorithm works until no new cluster component leaves a cluster and goes into other group and no new focus point is set for any cluster. At the point when this objective is accomplished the algorithm is ceased (Nadeem, Umar, and Shahzad, 2018).

### 3.2 Classification

An enormous measure of information is accessible and there is need to classify this information, however these are accessible in a large portion of times. There is need to classify this information according to its types, such as; sound/video, configuration and so forth. Classification depends on the sort of information found or data mining functionalities, for example; characterization, discrimination, association, classification, clustering, and so forth. A few frameworks tend to be in general far reaching frameworks offering a several data mining functionalities together (Neha, and Vikram, 2015).

#### 3.2.1 Support Vector Machine (SVM)

The SVM classifier manages linear permutation of subset of the training set by finding a maximum edge over stimulated plane. The SVM plots the information into high dimensional feature space closing to unbounded with the assistance of most vital part, if vectors are nonlinearly distinguishable input features (Nadeem, Umar, and Shahzad, 2018) and order the information by the most elevated scope hyper-plane.

$f(\bar{x}) = sgn(\sum_{i}^{M} y_i \alpha_i \emptyset(\overline{x_i}, \bar{x}) + \delta)$

Where;

M = the numbers of samples in training information collection.

Xi= demonstrates vector bolster when ai> 0

∅= demonstrates a center capacity

X = unidentified vector sample

δ = is a doorstep.

(ai) is a parameter that is the results of curved quadratic programming issue regarding direct requirement. In this strategy it demonstrates that Polynomial kernel and Gaussian radial basis functions (RBF) are habitually incorporated for kernel functions. The (δ) is another parameter that is the consequence of picking any I where ai> 0 and the condition is Karush–Kuhn– Tucker condition (Burges, 1998).

### 3.3 Evaluation Rate

This study was developed using MATLAB (2016A) application, with Windows 10 containing 4G Smash and 250G HD.

To evaluate the developed model, this study performs the model prediction execution based on the evaluation measures that are determined based on the confusion matrix (True Positive TP, False Positive FP, True Negative TN and False Negative FN). The evaluation measures can be assessed using the metrics below (Farshid, and Shaghayegh, 2018).

i. Accuracy is the ratio of the correct classifications to total number of classifications. Accuracy = $\frac{TN+TP}{TN+TP+FN+FP}$

ii. Sensitivity= TP/ (TP+FN) %

iii. Specificity = TN/ (TN+FN) %

iv. Precision: TP/ (TP+FP)

### 4. Results

The dataset acquired from Ecobank Nigeria, as stated in section3 was filtered; k-means clustering algorithm is applied on the data so as to filter it from outliers and unrepresentative information. The attributes in the cluster using k-means is determined, the results of the cluster are shown below in table1 below. SVM classification model is developed using the filtered data and evaluated using the data.

The first cluster with a class label of 3 consists of 99 observations with 58.5859 % male customers and 41.41% of female customers, it was also noted that the customers that fell into this cluster have an account balance between NGN20,000 to NGN50,000.

The second cluster consists of 79 with class label of 1 has observations with 49.3671 % male customers and 50.6329% of female customers, the customers that fell into this group have an account balance between NGN50,000 to NGN76,000.

The third cluster comprises of 22 with class label of 2, and has observations with 54.5455 % male customers and 45.4545% of female customers; the clients that fell into this group have an account balance above NGN75, 000.

Table 1. Results of Cluster Analysis Using K-Means Algorithm

| Main attributes definition | | | | |
|---|---|---|---|---|
| Clusters | Instances | Balances (NGN) | Male | Female |
| 1 (Class 3) | 99 (49.5)% | <=50,000 | 58.5859 % | 49.3671 % |
| 2 (Class 1) | 79 (39.5)% | <=75,000 | 49.3671 % | 50.6329% |
| 3 (Class 2) | 22 (11)% | >75,000 | 54.5455% | 45.4545% |

K-means algorithm has the capacity to cluster the data into class labels of 1, 2, and 3. The dataset was passed into the SVM with the 10 folds cross validation. SVM used the polynomial kernel to perform the classification. The result of the 10-fold cross validation is shown in figure 4.2 below with a training time of 17.8721secs and a classification accuracy of 97%.

The confusion matrix shows that the class 1 was classified correctly and 3 classified incorrectly, for the class 2, 20 were classified correctly and 2 were classified incorrectly, the class 3 has 98 and were correctly classified and 1 was incorrectly classified.

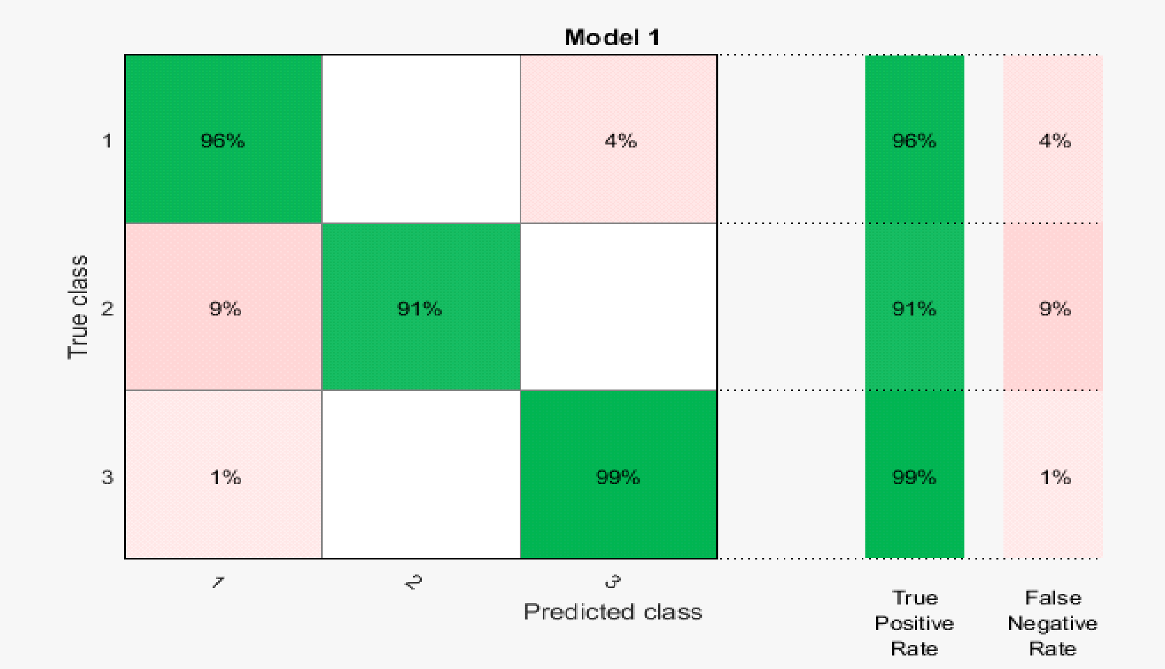

Figure 3. Confusion matrix with true positive and negative rate.

The receiving operating characteristic (ROC) curve shows the analytic capacity of SVM classifier. The curve demonstrates an exceptional output result of the SVM classifier.

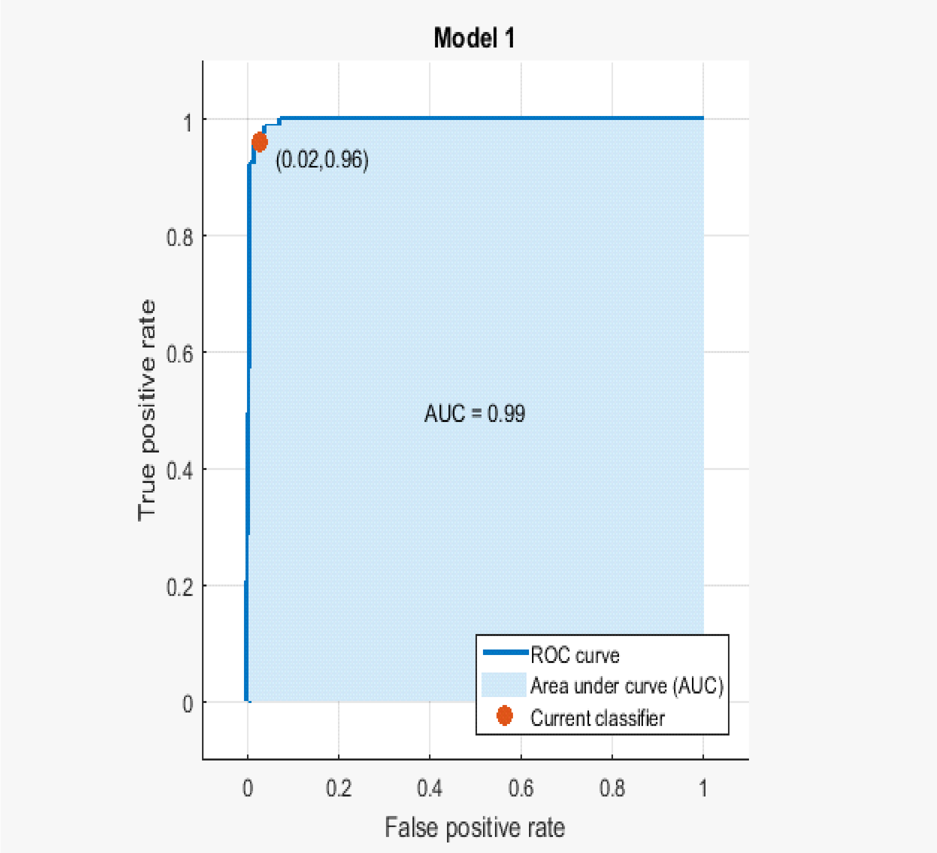

Figure 4. ROC Curve.

Table 2 below gives the performance evaluation of the experiment performed.

TP = 20; TN = 76; FP = 3; FN = 2.

Table 2. Evaluation Measure of the Experiment Performed

| **Evaluation Measures** | **Experiment Results** |
|---|---|
| Accuracy (%) | 97 |
| Sensitivity (%) | 91 |
| Specificity (%) | 97 |
| Precision (%) | 87 |
| Computational Time (Sec) | 17.8721 |

This study carried out an investigation on a customer churn analysis on a bank data, using K-means clustering and SVM approach for classification. The K-Means clustering divided the information into 3 groups as shown table 2, SVM was used for the classification, and the results were analyzed using an evaluation measure shown in table2. The chart for the evaluation measure is shown in figure 5 below.

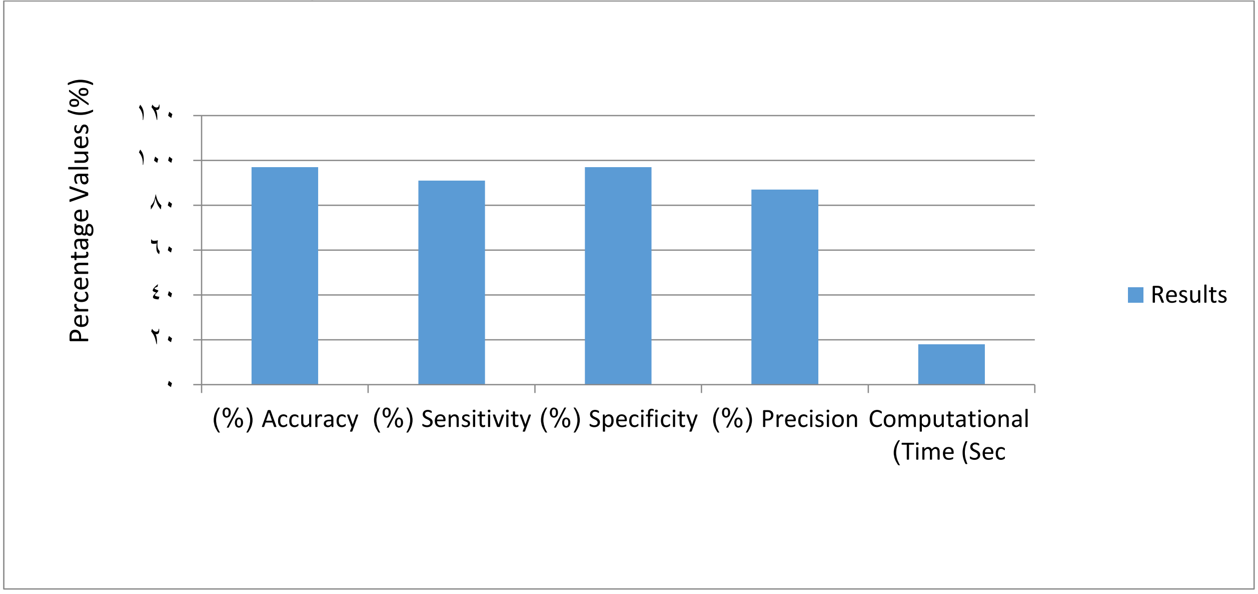

Figure 5. Evaluation Measure Chart

.

## 5. Conclusion

The development and arrangement of different administrations by the financial industry expands the likelihood of losing profitable clients. Quick development of data innovation in various organizations such as the financial industries that generates huge databases, reasonable investigation can be made to predict the conduct of customers and build up the connections of customers, in order to satisfy, attract and retain them.

Data mining procedures can be successfully used to extract hidden information and learning in customers' information. Directors can use this learning during the process of decision making. Customer clustering and analyzing conducts of each cluster can be an essential step in actualizing customer relationship executives' frameworks.

In this study, a data mining method was proposed to build up a framework containing collectives and classification of churned customers. K-means algorithm was used to conduct the customer's segmentation. Customers were separated into three groups based on their features. The groups of customers were investigated and classified using SVM to predict the level of the future of customer quality. In this study, customer churn prediction was carried out using the customer attractiveness and customer churn behavior. The proposed system enables the banking organization to predict the likelihood of investigating the conduct of past, current and future clients..

## References

Rohini , M., and Devaki, P. (2017). Analysis of Customer Churn by Big Data Clustering. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, No. 3, March 2017pp.6157-6162.

Yi, W., Qixin, C., Chongqing, K., and Qing, X. (2016). Clustering of Electricity Consumption Behavior Dynamics Toward Big Data Applications. IEEE Transactions on Smart Grid, Vol. 7, No. 5, pp. 1-13.

Guangqing, L., and Xiuqin, D. (2012).Customer Churn Prediction of China Telecom Based on Cluster Analysis and Decision Tree Algorithm. Communications in Computer and Information Science. Vol. 315, pp. 219-327.

Wang, Y., Chen, Z.(2010). The Application of Classification Algorithm Combined with K-means in Customer Churning of Telecom. Journal of Jiamusi University (Natural Science Edition). Vol. 28, No. 2, pp. 175–179.

Amjad, H., Reham, D., Osama, H., Ruba, O., and Hossam, F. (2015). Hybrid Data Mining Models for Predicting Customer Churn. International Journal of Communications, Network and System Sciences, Vol. 8, pp. 91-96.

Nadeem ,A.N., Umar, S., and Shahzad, M.S. (2018). A Review on Customer Churn Prediction Data Mining Modeling Techniques. Indian Journal of Science and Technology, Vol. 11, No. 27, pp. 1-7. DOI: 10.17485/ijst/2018/v11i27/121478.

Hossam, F. (2018). A Hybrid Swarm Intelligent Neural Network Model for Customer Churn Prediction and Identifying the Influencing Factors. Information Journal. Vol. 9, No. 288, pp. 2-18. doi:10.3390/info9110288.

Burges, C.J.C (1998). A tutorial on support vector machines for pattern recognition. Vol. 2, pp. 121–67. Data Mining and Knowledge Discovery (1997). Springer. Vol. 2, No. 2, pp.121–67.

Zhao, Y., Li, B., Li, X., Liu, W., and Ren, S. Customer churn prediction using improved one-class support vector machine. In Proceedings of the International Conference on Advanced Data Mining and Applications, ADMA 2005, Wuhan, China, 22–24 July 2005; pp. 300–306.

Fathian M, Hoseinpoor Y, Minaei-Bidgoli B (2016) Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods. Kybernetes. Vol. 45, No. 5, pp. 732–743.

Alwis, P.K.D., Kumara, B.T.G., and Hapuarachchi, H.A.C. (2018). Customer Churn Analysis and Prediction in Telecommunication for Decision Making. International Conference on Business Innovation. Vol. 1. pp. 40-45.

Wenjie, B., Meili, C., Mengqi, L., and Guo. L. (2016). A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn.IEEE Transactions on Industrial Informatics. Vol. 12, No. 3. Pp. 1270-181. DOI: 10.1109/TII.2016.2547584.

Farshid, A., and Shaghayegh, A. (2018). Customer Behavior Mining Framework Using Clustering and Classification Technique. Journal of Industrial Engineering International. Pp. 1-18. Doi.org/10.1007/s40092-018-0285-3.

Philip, S., and Thomson, N. (2017). Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors. Pp. 1-22. arXiv:1703.03869v1.

Neha, M., and Vikram, S. (2015). A Survey of Classification Techniques in Data Mining. IJCSMS (International Journal of Computer Science & Management Studies) Vol. 16, No. 1. pp. 9
-12