

# Providing an efficient Customers Churn Prediction Model based on Improvised K-Means Clustering And Non Linear Support Vector Machine

Anuradha, Shaveta Kalsi  
Research Scholar, Assistant Professor  
Computer Science and Engineering  
DAV Institute of Engineering and Technology  
Jalandhar, India

**Abstract :** This paper proposed a new approach to enhance the performance of existing base techniques including Neural networks, Logistic Regression, Linear Support vector machines and Non-Linear support vector machine with the proposed technique Improvised K-Means with NLSVM. The Improvised K-Means algorithm resolved the random selection problem of cluster centroid of K-Means by choosing the cluster centroid by taking the mean value of the data points. The Improvised K-Means algorithm clusters are then classified with Non-Linear Support vector machine classification algorithm. This enhanced approach is used for predicting customer churn. So that proactive measures could be taken by company for churn prevention. The experimental results show that the proposed technique performs better than the existing base techniques in terms of recall and f-measure.

**IndexTerms** – Data Mining, Customer Churn Prediction, Clustering, Classification, K-Means, Non-Linear Support Vector Machine (NLSVM).

## I. INTRODUCTION

Data mining is the computing procedure of finding designs in huge dataset including strategies at the crossing point of AI and database framework. Customer churn modelling alludes to when a client (player, endorser, client, and so forth.) stops his or her association with an organization. Online organizations regularly treat a client as stirred once a specific measure of time has slipped by since the client's last connection with the website or administration. The client agitate examination highlight encourages you recognize and center around higher worth clients, figure out what activities ordinarily go before a lost client or deal, and better comprehend what elements impact client spending. When you improve client maintenance, you significantly improve the primary concern. A typical hotspot for information is an information store or data warehouse [1].

A Predictive Churn Model is an instrument that characterizes the means and phases of client stir, or a client leaving your administration or item. Having a prescient beat model gives you mindfulness and quantifiable measurements to battle against in your maintenance endeavors. Agitate forecast model limits the client danger of beat to hold clients. The model allots high agitate probabilities to churners dependent on past clients' conduct and create maintenance measures for churners and non-churners to remain [3].

### *Support Vector Machine (SVM)*

A support vector machine (SVM) directly isolates information with a hyperplane (one measurement not exactly the component space, for example a line in two measurements) with reference to its nearest information focuses, called bolster vectors. The SVM attempts to augment the help vectors to accomplish the most extreme edge conceivable between the help vectors and the hyperplane.

### *Non-Linear Support Vector Machine (NLSVM)*

In reality, our dataset is by and large scattered up somewhat. To take care of this issue detachment of information into various classes based on a straight direct hyperplane can't be viewed as a decent decision. For this Vapnik proposed making Non-Linear Classifiers by applying the portion stunt to greatest edge hyperplanes. In Non-Linear SVM Classification, information focuses plotted in a higher dimensional space. It regularly happens that our information focuses are not straightly divisible in a p-dimensional (finite) space. To comprehend this, it was proposed to delineate dimensional space into an a lot higher dimensional space. We can draw modified/non-straight hyperplanes utilizing Kernel stunt. Each part holds a non-direct bit capacity [2].

### *Clustering*

Clustering can be defined as a process which helps in forming a group of data objects into classes of similar objects. A cluster of similar objects can be called as one group. In the cluster analysis, a dataset is firstly partitioned into groups based on similarity in data and then the class labels are assigned to the groups. In this process, a group of distinct data objects is classified as similar data objects. Clustering can be helpful for marketers as it discovers different groups in their customer

database. Customer groups of organizations can be characterized on the basis of purchasing patterns. Cluster analysis serves as a tool in data mining as it gains an insight into the distribution of data in order to observe the features of every cluster. Clustering technique is useful for exploring the characteristics in data [4].

### ***K-Means Clustering***

For the clustering of data, most commonly used algorithm is K-means clustering but K-means algorithm have many limitations like K-means algorithm takes initial clusters randomly which is not necessarily true in real world applications. The K-means algorithm is sensitive to initial centres selection. The Improved K-Means Clustering algorithm is applied for dimensionality reduction to remove outliers and noisy data. To overcome the disadvantages of K-Means clustering algorithm, an improved K-Means algorithm has been proposed for implementation of clustering technique in this work. Anyhow, the advantages of K-Means clustering algorithm include the fact that whenever the value for k is kept small, the algorithm performs faster than other clustering methods even if variables are large. Moreover, this algorithm obtains tighter clusters [5].

## **II. RELATED WORK**

Ali et al. [2] examined the forecast of beat conduct of client by utilizing different information mining systems. the lead when customer never again needs to remain with his relationship with the organization. Client stir the board the administrators is accepting critical occupation in customer the board. Nowadays media transmission associations are focusing on perceiving high worth and potential unsettling customers to construct advantage and bit of the general business. It is grasped that creation new customer is all the more exorbitant instead of holding existing customer. There is a present issue that customers leave the association in light of darken reasons. It will over the long haul help with analyzing customer's lead and portray whether it is a beat customer or not. In this investigation, we used online educational record open at Kaggle for desire for Customer blend direct using different classifiers i.e SVM (Support Vector Machine), Bagging, Stacking, C50/J48, PART, Naïve Bayes, Baysen Net, Adaboost and see that our model gave 99.8% precision level using Bagging Algorithms.

Xiao et al. [3] has clarified agitate forecast displaying, illustrating, the path toward selecting imprints to customers is extravagant and dull, and along these lines there are more unlabelled than named tests. To deal with this issue, this examination joins semi regulated learning, cost-sensitive learning procedure Meta cost with troupe methodology self-assertive subspace to propose a semi-oversaw gathering model reliant on Meta cost. This model joins the going with three stages: (1) use the Metacost method to adjust the class name of the hidden stamped getting ready set to secure another readiness set; (2) name the unlabelled models explicitly; (3) train a couple of fundamental classifiers with the last planning set and use them to gather the models in the test set. The precise results for two customer mix figure enlightening files display that the proposed model raises customer shake desire execution differentiated and by and large used coordinated assembling and semi regulated models.

Bart Baesens et al. [4] in their assessment considers In this paper, we extensively look at the exhibition of cutting edge methods to manage class lopsidedness with regards to beat expectation. An as of late created expected most extreme benefit rule is utilized as one of the primary execution measures to offer more bits of knowledge from the point of view of money saving advantage. The exploratory outcomes demonstrate that the connected assessment metric has an incredible impact on the exhibition of systems. It is additionally discovered that outfit techniques are the most conspicuous strategies and one fascinating finding is that the first Bagging and arbitrary backwoods learning calculation without considering class irregularity get the best outcomes as for the benefit based measure.

Koen W. De Bock et al. [5] in their assessment considers Using genuine information of poker players at bwin, single calculations, CART choice trees and summed up added substance models are benchmarked to their outfit partners, arbitrary timberlands and GAMens. The outcomes demonstrate that stir expectation is an important methodology to distinguish and profile those clients in danger. Moreover, the presentation of the groups is more hearty and superior to the single models. This investigation features the benefits of the groups RF and GAMens, that is, vigorous and better forecast execution, than the single calculations, CART and GAM.

Samineh Jafari et al. [6] in their assessment considers client beat expectation model utilizing a standard CRISP-DM (Cross Industry Standard Process for Data Mining) procedure dependent on RFM (Recency, Frequency, Monetary) and arbitrary woodland strategy. The database of Solico nourishment enterprises gathering is investigated. Utilizing this model, the clients tending to turn over are distinguished and powerful advertising systems will be made arrangements for this gathering. Client conduct examination shows that length of relationship, the relative recurrence and the normal bury buy time are among the best indicators. Just the records full of feeling on beat expectation were considered; this issue is the beginning stage of stir the executives.

Mestre M.R. et al. [7] depicted the examination of client conduct to discover stir clients. They tell that from the profiles of assortment of clients and their changing conduct after some time, associations can make promoting methodologies to know the gathering of clients and to choose whether those clients are gainful or not. They proposed a calculation which is cross breed from various leveled grouping and hidden markov model (HMM). They contrast increased technique and the non-expanded strategy with genuine information and engineered information to speak to that their proposed model performs better in foreseeing client

conduct. They utilize diverse bunching calculations for division of the clients. At that point they use choice hypothesis to check whether their proposed model is monetarily advantageous for an association.

Nabavi S. et al. [8] depicted the information mining capacities, structure and usage of client beat forecast model with CRISP-DM dependent on RFM and Random Forest Technique. Client conduct examination tells that the length of relationship, normal buy time, and relative recurrence are the best indicators. For the division of stir clients they utilize irregular woodland system and supported tree as a crossover method.

Wang C. et al. [9] proposed another system to anticipate clients' obtaining conduct utilizing buy arrangements of clients. At that point consecutive buy examples are removed utilizing affiliation rule. Utilizing buy value-based records of clients, profile of the clients is manufacture that depicts the preferences of the clients. At that point a gathering of clients recognized who have comparative acquiring conduct utilizing the figurings of relationships among clients. Exchange grouping is utilized to bunch every one of the exchanges of the clients. At that point SOM procedure is utilized to recognize client buy successions. Successive buy examples are removed utilizing affiliation principle mining. They foresee the client conduct utilizing client's buy succession base on exchange information.

Basiri J. et al. [10] talked about new methodology, the arranged weighted averaging (OWA) strategy to improve the forecast exactness of existing agitate the executives frameworks. In this paper, they utilized the qualities of both sacking and boosting and LOLIMOT calculations and proposed OWA way to deal with consolidate these calculations.

### III. PROPOSED WORK

A Hybrid clustering with classification model is proposed technique comprising of enhanced K means (Clustering) technique with support vector machine (classification) technique which is used to mine the data to extract the useful patterns and to improve the accuracy of the classifier than the existing hybrid technique. It will be used to identify churn customers and effective marketing strategies could be planned for this group of customers. Partitioning algorithms comes under clustering methods that subdivide the datasets into k groups, where k represents the number of groups. The implementation of partitioning algorithm on large data sets can be done through K-means clustering. In K-means clustering algorithm, every cluster is formed by taking the center or calculating the mean of the data points belonging to that cluster. The K-means clustering algorithm is mainly used to find out the data groups which have not been labeled in the dataset. This is helpful in confirming various business assumptions such as identifying unknown groups in multivariate data sets. After dividing the data from customer base into clusters, an informed decision can be made about the buyer of this product.

According to the existing techniques surveyed, when K-means clustering algorithm is combined with Non Linear SVM, it will give better results but the issue is the problem of selecting the initial centroids in basic K-means clustering algorithm. This problem is mainly focused in the proposed technique and steps taken for resolving this issue are mentioned below:

- i. Data is partitioned into k equal parts. Then the arithmetic mean of each part is taken as the centroid point.
- ii. K-means is applied on the input dataset by finding the Euclidean distance of each data point from the centroid and clusters are defined. If the distance of centroid of the present nearest cluster is less than or equal to the previous distance, then the data point remains in that cluster and there is no need to find its distance from other cluster centroids.
- iii. Apply clustering on the dataset for dimensionality reduction and then classify that reduced dataset using Support Vector Machine classifier.
- iv. The initial centroids are randomly selected in case of simple K-means algorithm but it is not so in proposed algorithm. The proposed work is to select the initial centroids by partitioning the data into k equal parts and then the arithmetic mean of each part is taken as the centroid point. The efficiency and accuracy of enhanced K-means algorithm is more than simple K-means.

Firstly, collect dataset from UCI repository and represent it by a list  $D (d_1, d_2, d_3, \dots, d_n)$  which specifies a data set consisting n objects. After this, apply pre-processing and filtering process on the data by using replace missing value filter and then read the authenticated dataset. Later on, select a value of k for the data. In next step, divide the dataset into k equal parts. To calculate the centroid point, find the arithmetic mean of each part as:

$$\text{Mean} = (\sum d_i) / n \quad (3.1)$$

Where  $\sum d_i$  = sum of all data instances or records and n is the total number of data instances in each partition of data set.

Compute the Euclidean distance of each data point  $d_i$  to all the centroids as  $\text{edist} (d_i, c_j)$ . Euclidean distance of data record 1 to data record 2 is calculated as follows:

$$D (d_1, d_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (3.2)$$

Where  $d_1 = (x_1, x_2)$  and  $d_2 = (y_1, y_2)$

For each  $d_i$ , examine the closest centroid and assign  $d_i$  to that centroid and Set  $\text{Near\_edist}[i] = \text{edist}(d_i, c_j)$  //  $c_j$ : nearest centroid. Then, for each cluster  $j$ , recalculate the centroids. After this repeat the following steps for each data-point  $d_i$

- a. Compute its distance from the new centroid of the present nearest cluster.
- b. If this distance is less than or equal to the previous distance, the data-point stays in that cluster,

Else

- i. Compute  $\text{edist}(d_i, c_j)$  from all cluster centroids; End for.
- ii. Assign the data-point  $d_i$  to the cluster with the nearest Centroid.
- iii. Set  $\text{Near\_edist}[i] = \text{edist}(d_i, c_j)$ ; End for loop.

After all this, take best average sum of all Euclidean distances and obtain the final output and train the classifier using reduced dataset. Finally, evaluate the performance on the basis of the following parameters:

$$\text{a. Precision} = \text{TP} / \text{TP} + \text{FP} \quad (3.3)$$

$$\text{b. Recall} = \text{TP} / \text{TP} + \text{FN} \quad (3.4)$$

$$\text{c. F-Measure} = 2 * \text{Recall} * \text{Precision} / \text{Precision} + \text{Recall} \quad (3.5)$$

$$\text{d. Accuracy} = \text{TP} + \text{TN} / \text{TP} + \text{FP} + \text{TN} + \text{FN} \quad (3.6)$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN False Negative

The flowchart of the proposed technique is represented graphically as follows:

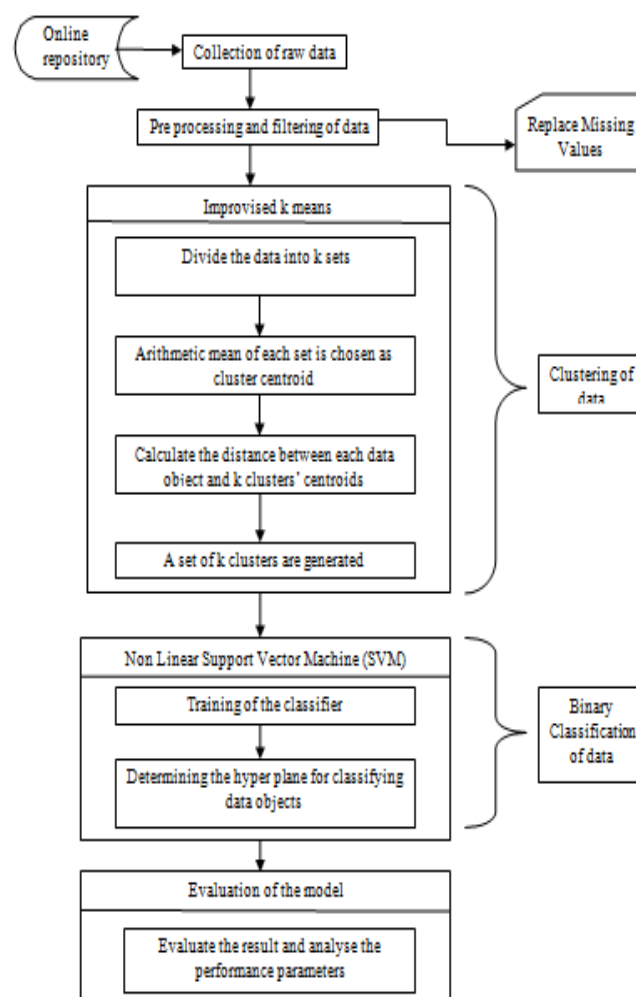


Figure 1: Flowchart of Proposed Technique

#### IV. EXPERIMENTAL RESULTS

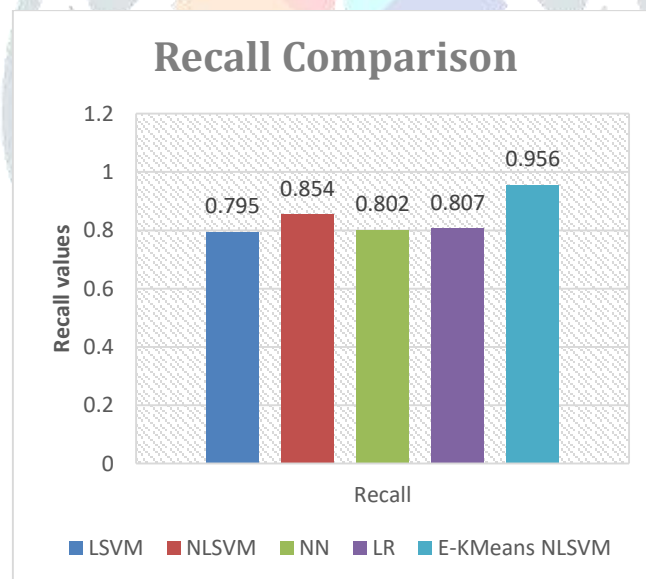
The diversion has been done in Java Net Beans. NetBeans is an open-source wander committed to giving rock solid programming change things (the NetBeans IDE and the NetBeans Platform) that address the necessities of specialists, customers and the associations who rely upon NetBeans as an explanation behind their things; particularly, to enable them to develop these things quickly, successfully and easily by using the characteristics of the Java organize and other significant industry principles.

Recall and F-measure are used as parameters to evaluate the performance of the proposed technique in contrast to the existing technique.

The results presented in Table 1 shows the comparison of Recall value for five classifiers and it can be concluded that proposed Hybrid technique which is Improved K-Means NLSVM has attained a highest peak among all other cases. It can be noted that the value for Neural Network is 0.802, Logistic Regression is 0.807, Linear Support Vector Machine is 0.795 and Non Linear Support Vector Machine has 0.854 recall value. Whereas in the case of proposed Hybrid technique which consists of Improved K-Means clustering with Non Linear Support Vector Machine, the recall value comes out to be 0.956.

**Table 1: Recall Comparison**

Algorithms	Recall
LSVM	0.795
NLSVM	0.854
NN	0.802
LR	0.807
Improved K-Means NLSVM	0.956



**Figure 2: Recall Comparison of proposed technique with the existing techniques**

Figure 2 shows the recall comparison of proposed technique with the existing base techniques. The recall values is calculated on a scale between [0-1]. The value closer to 1 is better. The proposed technique Improved K-Means with NLSVM is having value 0.956 which is highest among all other techniques. The parameters shown in form of graphs and tables depicts that the proposed technique is better than the existing techniques.

The results presented in Table 2 shows the comparison of F-Measure parameter for all the classifiers and it can be concluded that proposed Improved K-Means NLSVM has attained a maximum value amongst other classifiers. It can be noted that the value for Neural Network is 0.801, Logistic Regression is 0.807, Linear Support Vector Machine is 0.795 and Non Linear Support Vector Machine has 0.852 recall value. Whereas in the case of proposed Hybrid technique which consists of Improved K-Means clustering with Non Linear Support Vector Machine, the recall value comes out to be 0.956.

**Table 2: F-Measure Comparison**

Algorithms	F-Measure
LSVM	0.795
NLSVM	0.852
NN	0.801
LR	0.807
Improvised K-Means NLSVM	0.956

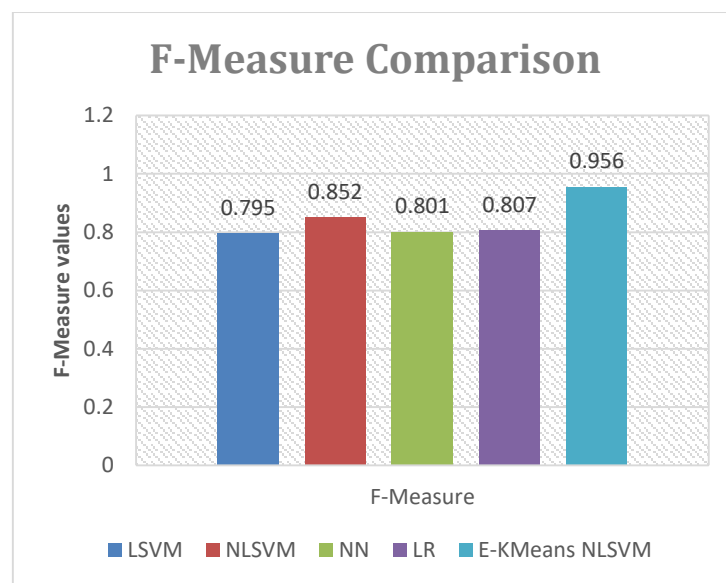
**Figure 3: F-Measure Comparison of proposed technique with the existing techniques**

Figure 3 shows the F-Measure comparison of proposed technique with the existing base techniques. The F-measure value is calculated on a scale between [0-1]. The value closer to 1 is better. The proposed technique Improvised K-Means with NLSVM is having value 0.956 which is highest among all other techniques. The parameters shown in form of graphs and tables depicts that the proposed technique is better than the existing techniques.

## V. CONCLUSION AND FUTURE SCOPE

Organizations need to be more concern about customers and always work for Customer Relationship Management. Losing present customers causes additional costs due to the loss in sale and it generated the need of attracting more new customers. Retention is the most valuable asset for organization. So, this paper proposed a new approach to enhance the performance of existing base techniques including Neural networks, Logistic Regression, Linear Support vector machines and Non-Linear support vector machine with the proposed technique Improvised K-Means with NLSVM. The Improvised K-Means algorithm resolved the random selection problem of cluster centroid of Kmeans by choosing the cluster centroid by taking the mean value of the data points. The Improvised K-Means algorithm clusters are then classified with Non-Linear Support vector machine classification algorithm. This enhanced approach is used for predicting customer churn. So that proactive measures could be taken by company for churn prevention. The experimental results show that the proposed technique performs better than the existing base techniques in terms of recall and F-measure. In future, some feature selection techniques can also be used with the proposed technique to optimize the attributes of the data and improve the performance to some more extent. Also, the same approach can be used for churn prediction in case of some other applications like telecommunication, banking sector and other retail industries.

**REFERENCES**

- [1] Adnan Amina, Feras Al-Obeidat, Babar Shah, Awais Adnan, Jonathan Loo, Sajid Anwar, "Customer churn prediction in telecommunication industry using data Certainty", Elsevier, Journal of Business Research, 2019, pp. 290–301.
- [2] Muhammad Ali , Aziz Ur Rehman, Shamaz Hafeez and Dr. Muhammad Usman Ashraf, "Prediction of Churning Behavior of Customers in Telecom Sector Using Supervised Learning Techniques", International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2018
- [3] Jin Xiao, Lan Huang and Ling Xie, " Cost-Sensitive Semi-Supervised Ensemble Model for Customer Churn Prediction", 15th International Conference on Service Systems and Service Management (ICSSSM), 2018
- [4] Bart Baesens, Zhu, Bing, and Seppe KLM vanden Broucke. "An empirical comparison of techniques for the class imbalance problem in churn prediction." Information Sciences 408, 84-99, (2017)
- [5] Siddhi Khandge, Dalvi, Preeti Kanade, Ashish Deomore and Aditya Venkatesh Bankar."Analysis of customer churn prediction in telecom industry using decision trees and logistic regression." IEEE, In Colossal Data Analysis and Networking (CDAN), Symposium on, 2016, pp. 1-4.
- [6] Michael Ewing, Ali Tamaddoni, Stanislav Stakhovych,, "Comparing churn prediction techniques and assessing their performance: A contingent perspective", Journal of service research 19, no. 2 ,2016, pp. 123-141.
- [7] Ali Tamaddoni, Jahromi, Stanislav Stakhovych, and Michael Ewing, "Managing B2B customer churn, retention and profitability", Industrial Marketing Management 43, no. 7, 2014, pp. 1258-1268.
- [8] Coussement, Kristof, "Improving customer retention management through cost sensitive learning", European Journal of Marketing 48, no. 3/4 , 2014, pp. 477-495.
- [9] Koen W. De Bock, Coussement and Kristof,. "Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning." Journal of Business Research 66, no. 9 ,1629-1636,(2013)
- [10] Nabavi and Samineh Jafari. "Providing a Customer Churn Prediction Model using Random Forest Technique" In Proc 5th IEEE-Conference on Information and Knowledge Technology (IKT), pp. 202-207,(2013)
- [11] Mestre Maria Rosario and Victoria Pedro., "Tracking of consumer behaviour in e-commerce". 16th International Conference on Information Fusion, Istanbul, Turkey, 2013, pp. 1214-1221
- [12] Nabavi Sadaf and Jafari Shahram., "Providing a Customer Churn Prediction Model using Random Forest Technique". 5th IEEE-Conference on Information and Knowledge Technology (IKT), 2013, pp. 202-207
- [13]Wang Chong and Wang Yanqing, "Discovering Consumer's Behaviour Changes Based on Purchase Sequences". 9th IEEE-International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012),pp. 642-645
- [14] Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr,"A Proposed Churn Prediction Model", International Journal of Engineering Research and Applications, Vol. 2, Issue 4, June-July 2012, pp.693-697.
- [15] Javad Basiri, Fattaneh Taghiyareh and Behzad Moshiri, "A Hybrid Approach to Predict Churn", IEEE Asia-Pacific Services Computing Conference, 2010, pp. 485-491.