2018 International Conference On Business Innovation (ICOBI), 25-26 August 2018, NSBM, Colombo, Sri Lanka

Customer Churn Analysis and Prediction in Telecommunication for Decision Making

P.K.D.N.M. Alwis
Department of Computing and Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka,
madushani.niroshi@gmail

B.T.G.S. Kumara
Department of Computing and Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka

H.A.C.S. Hapuarachchi
Department of Computing and Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka

Abstract*— With the rapid development of communication technology, the field of telecommunication faces complex challenges due to the number of vibrant competitive service providers. Customer Churn is the major issue that faces by the Telecommunication industries in the world. Churn is the activity of customers leaving the company and discarding the services offered by it, due to the dissatisfaction with the services. The main areas of this research contend with the ability to identify potential churn customers, cluster customers with similar consumption behavior and mine the relevant patterns embedded in the collected data. The primary data collected from customers were used to create a predictive churn model that obtain customer churn rate of five telecommunication companies. For model building, classified the relevant variables with the use of the Pearson chi-square test, cluster analysis, and association rule mining. Using the Weka, the cluster results produced the involvement of customers, interest areas and reasons for the churn decision to enhance marketing and promotional activities. Using the Rapid miner, the association rule mining with the FP-Growth component was expressed rules to identify interestingness patterns and trends in the collected data have a huge influence on the revenues and growth of the telecommunication companies. Then, the C5.0 Decision tree algorithm tree, the Bayesian Network algorithm, the Logistic Regression algorithm, and the Neural Network algorithms were developed using the IBM SPSS Modeler 18. Finally, comparative evaluation is performed to discover the optimal model and test the model with accurate, consistent and reliable results.

Keywords—bayesian network, c5.0 decision tree, logistic regression, neural network

I. INTRODUCTION

Decision making is a key feature of every organization. The quality of decisions made is dependent on some amount of knowledge generated from existing or researched information. The use of modern analytical tools to generate such knowledge is reasonable for any profit-driven firm. Taking decisions on customers is one of the key points in most companies, especially companies in the service sector. The ability of these companies to predict customer churn is critically inadequate. Customer churn is the action of the customer who is like to leave the company and it is one of the mounting issues of today's rapidly growing and competitive telecommunication industry. To minimize the customer churn, prediction activity to be an important part of the telecommunication industry's vital decision making and strategic planning process.

A. *Churn Prediction*

Today numerous telecom companies are prompt all over the world. Telecommunication market is facing a severe loss of revenue due to increasing competition among them and loss of potential customers [1]. Churn is the activity of the telecommunication industry is the customers leaving the current company and moving to another telecom company. Many companies are finding the reasons of losing customers by measuring customer loyalty to regain the lost customers. To keep up with the competition and to acquire as many customers, most operators invest a huge amount of revenue to expand their business in the beginning [2]. In the telecommunication industry each company provides the customers with huge incentives to attract them to switch to their services, it is one of the reasons that customer churn is a big problem in the industry nowadays. To prevent this, the company should know the reasons for which the customer decides to move on to another telecom company. The Telecom Churns can be classified into two main categories: Involuntary and Voluntary. Involuntary are easier to identify. Involuntary churn is those customers whom the Telecom industry decides to remove as a subscriber. They are churned for fraud, non-payment and those who don‘t use the service. Voluntary churn is difficult to determine because it is the decision of the customer to unsubscribe from the service provider. Voluntary churn can further be classified as incidental and deliberate churn [3]. The former occurs without any prior planning by the churn but due to change in the financial condition, location, etc. Most operators are trying to deal with these types of churns mainly.

B. *Churn Management*

Churn management is very important for reducing churns as acquiring a new customer is more expensive than retaining the existing ones [4]. Churn rate is the measurement for the number of customers moving out and in during a specific period of time. If the reason for churning is known, the providers can then improve their services to fulfill the needs of the customers. Churns can be reduced by analyzing the past history of the potential customers systematically [5]. A large amount of information is maintained by telecom companies for each of their customers that keep on changing rapidly due to a competitive environment. The information includes the details about billing, calls and network data. The huge availability of information arises the scope of using Data mining techniques in the telecom database. The information available can be analyzed in different perspectives to provide various ways to the operators to predict and reduce churning. Only the relevant details are used in the analysis which contributes to the study from the information given. Data mining techniques are used for discovering the interesting patterns within data and it helps to learn to predict whether a customer will churn or not based on customer‘s data stored in the database.

C. *Research Objectives*

The main objective of this research is to produce a predictive model with better results that assess customer churn rate of telecommunication companies using the predictive analytics algorithm for data mining.

The supporting objectives examined are to:

i. Cluster customers into various categories to enhance marketing and promotional activities.

ii. Mine the relevant patterns embedded in the collected data have a huge influence on the revenues and growth of the Telecommunication companies.

II. METHODOLOGY AND EXPERIMENTAL DESIGN

Data mining and statistical algorithms were used in the data analysis, model building and model deployment in this research. Weka 3.8, Minitab 17, RapidMiner Studio 8.1 and IBM SPSS Modeler 18 were the analytical tools used in the respective analysis and mining process.

A. *Data Collection*

The questionnaire was used as the tool to collect the data primarily from customers. The Google drive plug-in was used to design the questionnaire. Training data was collected from the 200 respondents and 50 responses were received from respondents on the questionnaire for testing data. The data was collected during the period (October – November) of 2017.

TABLE 1: THE VARIABLES USED IN DATASET FOR THIS RESEARCH

No	Variable Name	Description
1	Age, Gender, Occupation	Demographic variables considered
2	The number of networks	Identifies the number of mobile networks a customer is connected to and actively using
3	Frequently used network	Identifies the most frequently used mobile network by the consumer
4	Tariffs	The type of customer, whether a prepaid or post-paid customer
5	Tenure	Length of time a customer has been with a particular subscriber
6	Credit purchase amount (CpM)	Approximates the amount used to purchase call credits a month in rupees
7	Data purchase amount (DpM)	Approximates the amount used to purchase data bundles a month in rupees
8	Internet usage	Identifies whether customer have used internet facility or not
9	Product innovation	Determines whether product innovation is necessary for sustaining customers
10	Churn	Identifies whether customer have changed networks or not

B. *Data Pre-processing*

The training and testing dataset used in this research may be included missing data, repeated data or inconsistent data. To handling missing data and removing duplicated data values data pre-processing is done. The RapidMiner tool is used at this stage to pre-process the data for analysis and mining. In doing cluster analysis, the Pearson chi-square and predictive model building, the data types to be converted into numerical values.

TABLE 2: CODES FOR ALTERNATIVES

Variable	Alternative	Code
Gender	Female	0
	Male	1
Occupation	Student	1
	Government Employee	2
	Private Employee	3
	Own Business	4
	Others	5
Network often used	Dialog	1
	Mobitel	2
	Airtel	3
	Hutch	4
	Etisalat	5
Tenure	Less than 1 year	1
	1-3	2
	3-5	3
	Above 5	4
Churn	No	0
	Yes	1
Tariffs	Pre-paid	1
	Post-paid	2
	Both	3
Usage of Internet	No	0
	Yes	1
Product innovation	No	0
	Yes	1
	Not sure	2

C. *Research Framework*

In figure 1, a research framework developed to address problems of this research. This research framework details the sectorial areas of concentration and the data mining algorithms adapted in creating the predictive model. It includes a model deployment and evaluation strategies that will assess its effectiveness and efficiency.

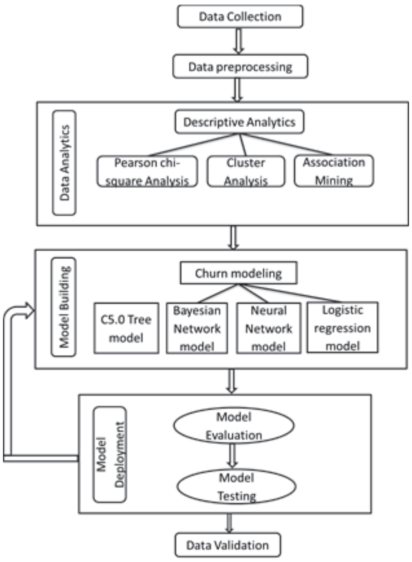


Figure 1: Research framework of Customer Churn Analysis and Prediction in Telecommunication

III. RESULTS

A. Pearson Chi-square Test

Pearson Chi-square test is used to evaluate the variables which are associated with the decision of churn that can be used in the predictive model building. Pearson and likelihood ratio chi-square tests are conducted using Minitab. The test produced significant results (p–value is less than α level of 0.05) to indicate that some of the variables have an association with the decision to churn.

TABLE 3: SUMMARY OF ASSOCIATION OF EACH ATTRIBUTES THE CHURN DECISION

Variable	P-Value	Association
Marital Status	0.645	No
Gender	0.038	Yes
Age	0.005	Yes
Occupation	0.011	Yes
Monthly Income	0.001	Yes
Purpose of mobile phone usage	0.107	No
No of mobile network connected	0.003	Yes
Mobile network often used	0.006	Yes
Tenure	0.000	Yes
CpM	0.004	Yes
Tariffs	0.029	Yes
Internet usage	0.020	Yes
DpM	0.105	No
Product Innovation	0.021	Yes

A. Cluster Analysis

Cluster Analysis is used to discover groups with identical features in collected data. These groups explained the interest areas and churn decision with the reasons for targeted marketing and product development. Using Weka 3.8 the k-means clustering produced four clusters out of the 200 collected data.

TABLE 4: CLUSTER INSTANCES WITH PERCENTAGE

Cluster Number	Clustered Instances	Percentage (%)
Cluster 0	69	17
Cluster 1	34	32.5
Cluster 2	65	16
Cluster 3	32	34.5

TABLE 5: FINAL CLUSTER CENTROIDS

Variable	Full Data (200.0)	Cluster 0 (69.0)	Cluster 1 (34.0)	Cluster 2 (65.0)	Cluster 3 (32.0)
Gender	0.515	1	1	0	0
Age	33.24	35.49	54.79	22.93	26.40
Occupation	2.305	2.6232	4.0882	1.18	2
Monthly Income	31827.5	38833.33	61617.64	10984.61	27406.25
NoOfMobileNetworkConnected	2.125	2.5217	3	1.30	2
MobileNetworkOftenUsed	2.195	2.2609	4.52	1	2
Tenure	3.085	3.7391	4	1.95	3
Tariffs	1.325	1.1884	1.38	1.35	1.5
CpM	1003.25	1105.79	1948.52	504.61	790.62
Internet Usage	0.8	1	1	0.3846	1
Churn	0.675	1	1	0	1

Telecom providers can leverage this cluster model to allocate customers' for conducted promotional activities. It is observed in the clusters that the churners are mostly businessman and private employees who are generally males and government employees who are female. These churners spend a lot of call credit per month and used prepaid service package. The customers do not intend to churn are mostly students who are generally females. Telecom providers, especially those who have endured a churn of customers need to pay attention in this cluster to the reason for churn as presented by these customers.

B. Association Rule Mining

Association rule mining used to determine interestingness patterns and trends between variables in the dataset. It is contracted to identify strong rules explored in the dataset using some measures of interestingness. The RapidMiner Studio 8.1 tool was used in creating the Association rules model for collected data. In creating the model, the Frequent

Pattern Growth (FP-Growth) algorithm was used to mine associations between variables that result in a churn decision with particular interest and focus on confidence. The generated Association rules model presented in Figure 2.

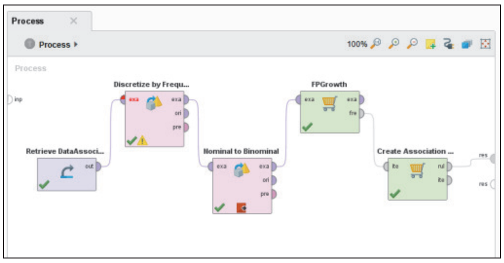


Figure 2: Association rules model

In the Table 6 showed ten (10) generated association rules were selected based on filtering the conclusion as the decision of churn is yes and sorted in descending order in line with confidence. The sorted rules have a maximum confidence of 95.5 percent and a minimum of 84.6 percent.

TABLE 6: TOP (10) GENERATED RULES

No	Premises	Conclusion	Support	Confidence	Laplace
1	InternetUsage= Yes, Gender =Male and Tenure = 3-5 years	Churn_Yes	0.105	0.955	0.995
2	Tariffs=Prepaid, Gender=Male and Tenure= 3-5 years	Churn_Yes	0.100	0.952	0.995
3	Gender=Male and Tenure=3-5 years	Churn_Yes	0.130	0.929	0.991
4	MobileNetwork OftenUsed=Mobitel and Tenure=3-5 years	Churn_Yes	0.115	0.920	0.991
5	InternetUsage= Yes and MobileNetwork OftenUsed=Mobitel	Churn_Yes	0.105	0.913	0.991
6	Tariffs=Prepaid and Tenure= 3-5 years	Churn_Yes	0.230	0.902	0.980
7	InternetUsage= Yes, Tariffs=Prepaid and Tenure= 3-5 years	Churn_Yes	0.190	0.884	0.979
8	Tenure= 3-5 years	Churn_Yes	0.270	0.871	0.969
9	Tariffs=Prepaid, Gender=Female and Tenure= 3-5 years	Churn_Yes	0.130	0.867	0.983
10	InternetUsage= Yes, Tariffs=Prepaid and Gender=Female	Churn_Yes	0.110	0.846	0.982

B. Predictive Model Building

Using the valid variables identified in the Pearson Chi-square test, the four predictive models are created with IBM SPSS Modeler 18.0 data mining software. The four classification modeling techniques; C5.0 tree, the Bayesian network, Neural Network and Logistic regression are used to create predictive models. The optimal model is recommended based on individual models and performance metrics.

An auto classifier was applied in the created C5.0 tree model in Figure 3, to test whether the selected C5.0 algorithm will be determined as one of the best algorithms to create the predictive model.

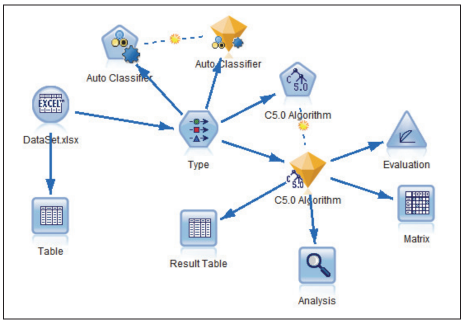


Figure 3: C5.0 algorithm tree model

The C5.0 algorithm was listed in the suggested churn algorithms which were applied to the data. In Figure 3, the matrix was applied to create a table showing the relationship between fields of Churn by $C-Churn. In the created above model, analysis and evaluation are used to create a report and a chart for comparing the accuracy of predictive models.

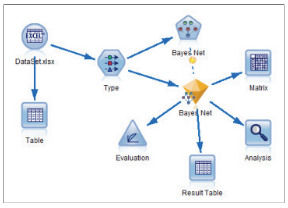


Figure 4: Bayesian network model

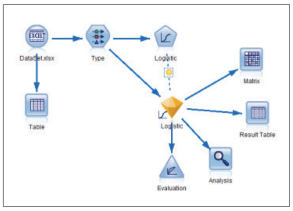


Figure 5: Neural network model

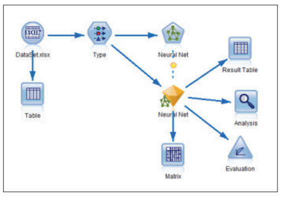


Figure 6: Logistic regression model

As the result of generating the logistic regression model, it built up a statistical model which consists of two mathematical equations to calculate the ability of a person being churner or non-churner.

Equation 1: Calculating Y’

$$Y' = 0.0682*Gender+(-0.00182)*Age +0.04558*Occupation+0.00001458* MonthlyIncome+(-0.7214)*Tenure +(-0.2053)*Tariffs+0.00001024*CpM+2.659$$

Equation 2: Calculating P(1)

$$P(1) = \exp(Y')/(1 + \exp(Y'))$$

Equation 1 consists of most relevant variables which are most affected by the churn decision. The variables values should be replaced by this equation and then the value of Y' can be calculated. Then the calculated Y' value should be replaced with the equation 2 and calculate the value of P(1). Prediction of being a churn or non-churn customer is depending on this P(1) value.

If the P(1) value is equal or greater than 0.5, then the prediction result is positive and the person will be a churner. If the P(1) value is less than 0.5, the result is close to 0 (zero). It means the prediction result is negative and the person will be a non-churner.

C. Model Evaluation

The four models are evaluated by testing the significance of the predictive model generated. The performance metrics of all the models were correlated for optimal performance using Area Under Receiver Operating Characteristic Curve (AUROC).

TABLE 7: CONFUSION MATRIX WITH TRAINING DATA

Model	Accuracy (%)	AUC Value
C5.0 algorithm model	85	0.888
BA model	79	0.886
LR model	72	0.762
NN model	70	0.759

The variables were equally tested for validity and reliability. The validity of the model indicates that it measures what it is intended for while reliability test produces consistent results. The tests assessed the efficiency and effectiveness of the model in predicting customer churn in Telecommunication.

		no	yes	% correct
C5.0	no	45	20	69.2
	yes	10	125	92.5
Overall Percentage				85%
BN		no	yes	% correct
	no	47	18	72.3
	yes	24	111	82.2
Overall Percentage				79%
LR		no	yes	% correct
	no	27	38	41.5
	yes	18	117	86.6
Overall Percentage				72%
NN		no	yes	% correct
	no	21	44	32.3
	yes	16	119	88.1
Overall Percentage				70%

TABLE 8: ACCURACY AND AUC VALUE OF EACH MODEL

Contrasting the four models, the C5.0 algorithm of decision tree proved optimal model with 85% accuracy and AUC value as 0.888 for the customer churn analysis and prediction in Telecommunication based on the chosen variables and attributes.

D. Model Testing

The optimal model based on the results of the evaluation is tested on the dataset designed to test the model. The C5.0 algorithm model was used to test the data as it was identified as the most optimal among the models. The chosen optimal model was tested using the test data collected from customers. The test data has 50 observations, 7 variables and coded the same as the coding in Table 2. The distribution of the dataset is along with all the gender, age, monthly income, occupation and the other demographic and operational variables used to develop the model. Predictions are then made to indicate which customers are likely to churn and those that are not. The predictor variable and target variables used in building the predictive churn model were tested for significance.

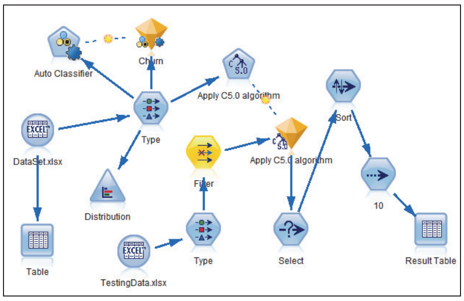
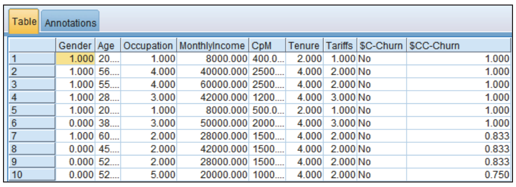


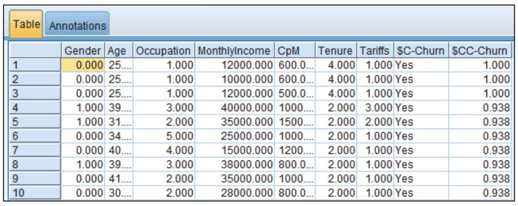
Figure 7: Test model for C5.0 algorithm

The test data is applied by mapping the dataset to the model designed by the C5.0 algorithm as indicated in Figure 7. Further model screening and applications are initiated to define the output in determining the likelihood of churn. The test results presented as the model predicted that 36 customers will churn with confidence from 100% to 55.6%. It was further explained by the results that over 62% of the churn customers have a confidence of above 80%. According to the Figure 8, the results also indicate the churn customers staying their network above 5 years. It is expensive to acquire new customers than to retain existing ones, the prediction of churners and the reasons proffered earlier need close attention. The top 10 churners and non-churners predicted by the model are presented in Figures 8 and 9 respectively. The source of the test data set can be connected to the database or server of the company to produce a real-time output of churn results for decision making.



	Gender	Age	Occupation	MonthlyIncome	CpM	Tenure	Tariffs	$C-Churn	$CC-Churn
1	1.000	20....	1.000	8000.000	400.0...	2.000	1.000	No	1.000
2	1.000	56....	4.000	40000.000	2500....	4.000	2.000	No	1.000
3	1.000	55....	4.000	60000.000	2500....	4.000	2.000	No	1.000
4	1.000	28....	3.000	42000.000	1200....	4.000	3.000	No	1.000
5	1.000	20....	1.000	8000.000	500.0...	2.000	1.000	No	1.000
6	0.000	38....	3.000	50000.000	2000....	4.000	3.000	No	1.000
7	1.000	60....	2.000	28000.000	1500....	4.000	2.000	No	0.833
8	0.000	45....	2.000	42000.000	1500....	4.000	2.000	No	0.833
9	0.000	52....	2.000	28000.000	1500....	4.000	2.000	No	0.833
10	0.000	52....	5.000	20000.000	1000....	4.000	2.000	No	0.750

Figure 8: Results of test predictions_Yes



	Gender	Age	Occupation	MonthlyIncome	CpM	Tenure	Tariffs	$C-Churn	$CC-Churn
1	0.000	25....	1.000	12000.000	600.0...	4.000	1.000	Yes	1.000
2	0.000	25....	1.000	10000.000	600.0...	4.000	1.000	Yes	1.000
3	0.000	25....	1.000	12000.000	500.0...	4.000	1.000	Yes	1.000
4	1.000	39....	3.000	40000.000	1000....	2.000	3.000	Yes	0.938
5	1.000	31....	2.000	35000.000	1500....	2.000	2.000	Yes	0.938
6	0.000	34....	5.000	25000.000	1000....	2.000	1.000	Yes	0.938
7	0.000	40....	4.000	15000.000	1200....	2.000	1.000	Yes	0.938
8	1.000	39....	3.000	38000.000	800.0...	2.000	1.000	Yes	0.938
9	0.000	41....	2.000	35000.000	1000....	2.000	1.000	Yes	0.938
10	0.000	30....	2.000	28000.000	800.0...	2.000	1.000	Yes	0.938

Figure 9: Results of test predictions_No

V. DISCUSSION AND CONCLUSION

Data mining is a symbolic tool in the Telecommunication industry that can exploit the large volume of data generated for pattern analysis. The recent increasing embrace of the predictive algorithm of data mining has given room for companies to assess their future success, challenges, and targets. The research brings to fore the relevant untapped customer data and knowledge for churn prediction and customer classification for better decision making. Clustering customers were developed in this research to determine the involvement of customers, interest areas and reasons for the churn decision. The results of the cluster analysis can be used in promotional and direct marketing purpose to access marketing strategies in the industry. In addition, the association rule mining was provided the significant results that present relevant knowledge of factors that have a huge influence on the revenues and growth of the Telecommunication companies. Telecommunication companies must grasp on this finding and work to maintain their clients. C5.0 Decision tree model, the Bayesian Network model, Logistic Regression model, and the Neural Network model were used and compared for the most optimal model that predicts accurately. The C5.0 algorithm of decision trees model proved optimal among the models with 85 percent accuracy and AUC value as 0.888. The C5.0 algorithm model of the decision tree can be recommended for churn management. The models can be used by industry with the IBM SPSS Modeler or any other appropriate tool with the same algorithm. The Telecommunication companies can connect the models directly to their servers or database to produce real-time results.

ACKNOWLEDGMENT

This study was done by the Department of Computing and Information Systems in the faculty of Applied Sciences in the Sabaragamuwa University of Sri Lanka. The author would like to thanks to the Dr.B.T.G.S.Kumara (Head, Department of Computing & Information Systems, Sabaragamuwa University of Sri Lanka) and Mr.H.A.C.S.Hapuarachchi (Lecturer, Department of Sports Sciences and Physical Education, Sabaragamuwa University of Sri Lanka) for their contribution in stimulating suggestions and encouragement, assistance, guidance, and cooperation.

REFERENCES

[1] V. Umayaparvathi and K. Iyakutti, "A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics," International Research Journal of Engineering and Technology (IRJET), vol. 03, no. 04, April 2016.

[2] Shin-Yuan Hung and Hsiu-Yu Wang, "Applying Data Mining to Telecom Churn Management," Department of Information Management, National Chung-Cheng University, Taiwan, ROC,.

[3] M.Balasubramanian and M.Selvarani, "Churn Prediction in Mobile Telecom Systems Using Data Mining Techniques," Department Of Computer Science, Annamalai University, Chidambaram, April 2014.

[4] Rahul J. Jadhav and Usharani T. Pawar, "Churn Prediction in Telecommunication Using Data Mining Technology," International Journal of Advanced Computer Science and Applications, vol. 2, no. 2, February 2011.

[5] K.Dahiya and S.Bhatai, "Customer churn analysis in telecom industry," 4th International Conference on Realibility, Infocom Tehnilogies and Optimization(ICRITO), 2015.

[6] Amjad Khan and Zahid Ansari, "Comparative Study Of Data Mining Techniques In Telecommunications-A Survey," Dept of Electronics and Communication, P.A. College of Engineering, Mangalore, India.