Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

Churn Prediction in Telecommunication Industry using Decision Tree

Nisha Saini[1], Monika[2], Dr. Kanwal Garg[3]
[1]Research Scholar, [2,3]Assistant Professor,
[1,2,3] Department of Computer Science and Applications,
Kurukshetra University, Kurukshetra

***Abstract-* Telecommunication industry provides customers an opportunity to choose from various service providers. However, certain factors such as low switching costs and deregulation by the government have contributed to the risk of customers switching to competitors. Customer churn can therefore be defined as the switching of a customer from services of one provider to another. Since it can be a costly risk, it needs to be managed properly. The presented paper aims at predicting customer churn using Decision Trees, one of the most widely used classification technique based on data mining.**

Keywords- Churn Prediction, Classification, Decision tree, CHAID, Data Mining.

I. INTRODUCTION

One of the key aspirations of any telecommunication company is to maintain a loyal customer base but since the customers have been provided with facility of switching from one service provider to another, telecommunication companies' are facing more problems. According to a study, acquiring a new customer is about 5-6 times costlier as compared to retaining an old one [1][2]. Customer churn is one of the major issues that the telecom industry is facing today. Hence customer churn may prove to be a costly risk if not managed carefully. Various costs are associated with customer churn and include loss of revenue, costs of customer retention and reacquisition, advertisement costs, organizational as well as planning and budgeting chaos [2]. Therefore it becomes quite necessary to identify the possible churning customers so that the losses can be prevented.

Data mining methods can be defined as the process of finding unknown patterns in huge data sets [8]. These methods find their applications in the area of CRM such as fraud detection, customer churn prediction etc. Classification is one of the data mining tasks that focuses on classifying unknown cases based on a set on known examples [9]. Hence, classification techniques based on data mining can be used for predicting churn in telecommunication industry[8][10].Wai-Ho Au et al. [3], Erfaneh and Tarokh [4], Wei Yu et al. [5], Chao et al. [6], Adnan and Asifullah [7] have also focused on using data mining techniques for churn prediction in telecommunication in their work. Various techniques have been used by various researchers but use of data mining techniques for predicting customers churn has turned out to be an efficient approach with high accuracy in results.

The rest of the paper is organized as follows: In the Second Section, a rich literature survey has been provided to put a spotlight on the related work that has been done in the field of churn prediction. The third section focuses on the data used for research purpose. The fourth section comprises of the algorithms implemented on the dataset. In the next section, the results obtained after implementation have been mentioned and analyzed. Finally the conclusion has been presented with some future scope of work.

II. RELATED WORK

Clement et al. [2] have presented new features categorized as contract-related, call pattern description, and call pattern changes description features derived from traffic figures and customer profile data. The given features were evaluated using Naïve Bayes and Bayesian network and obtained results were compared to results obtained using decision tree. Results have shown that probabilistic classifiers have shown higher true positive rate than decision tree but decision tree performs better in overall accuracy.

Essam et al. in [11] have introduced a simple model based on data mining to track customers and their behavior against churn. A dataset of 500 instances with 23 attributes has been used to test and train the model using 3 different techniques i.e., Decision trees, SVM and Neural networks for classification and k-means for clustering. Results indicate that SVM has been stated as the best suited method for predicting churn in telecom.

Umman Tuğba Şimşek Gürsoy [12] have compared regression techniques with decision tree based techniques. Results have shown that in logistic regression analysis churn prediction accuracy is 66% while in case of decision trees the accuracy measured is 71.76%. Hence decision tree based techniques are better to predict customer churn in telecom.

V. Umayaparvathi and K. Lyakutti [13] have used Neural Networks and Decision trees to build the churn prediction model. According to the results, Decision trees have 98.88% of predictive accuracy and an error rate of 1.11167%. Similarly neural networks have shown the predictive accuracy of 98.43% with 1.5616% of error rate. As is indicated by the results, decision trees have outperformed neural networks for churn prediction. According to the authors, selection of right combination of attributes and fixing the proper threshold values may produce more accurate results.

Saad et al. [14] have applied different machine learning algorithms such as linear and logistic regression, ANN (Artificial Neural Networks), K-means clustering, Decision Trees to identify churners and active customers. The best results were obtained using exhaustive CHAID, a variant of standard decision trees.

Ning Lu [15] has proposed a model with an “Implementation zone” where customers with highest churn probability can be addressed for retentive actions. The author has also proposed a further improvement in performance by analyzing other classification techniques as well or using a hybrid approach for more accurate results.

Vladislav and Marius in [17] have presented quality measures of six churn prediction models including regression analysis, naïve Bayes, decision trees, neural networks etc. They have also pointed out the links between churn prediction and customer lifetime value. According to the authors, new prediction models need to be developed and combination of proposed techniques can also be used.

Khalida et al. in [20] have used a specific training sample set was used to conduct an experiment on customer churn factor using decision tree. According to the authors, rule information can be easily understood by decision tree. An attempt has been made to identify various factors responsible for customer churn such as area.

Amal et al.[21] have reviewed that generally Decision tree based techniques, neural network trees and Regression techniques are applied in churn prediction. Decision tree based techniques outperform all other in terms of accuracy. On the other hand, neural networks outdo other techniques due to size of data sets.

From the presented literature work, it can be concluded that in most of the cases, Decision trees have outperformed other techniques for predicting churn in telecommunication industry.

III. DATA ACQUISITION

Churn analysis is done on the basis of historical data. This data may be accessible from the warehouse of respective company. The data set used in this study was acquired from an online source[1]. This data set is a longstanding customer data set of about 33,000 customers (active and disconnected) and include demographic as well as service details such as their location, call minutes used during different times of day, charges incurred for services, lifetime account duration etc. Phone number is the unique attribute to identify a particular customer. A customer is considered as active if he is still using the network and in case the services are terminated either voluntarily or involuntarily [11], the customer is disconnected.

IV. RESEARCH METHODOLOGY

In the current study, variants of decision data have been implemented in SPSS data mining tool.

A. Decision Tree

A decision tree is a classification scheme which generates a flow chart like structure where an internal node represents a test on an attribute, each branch represents outcome of the test and leaf node represents classes. Decision tree partitions the input space into cells where each cell belongs to one class [16]. The decision tree is developed into two phases: building and pruning. In the building phase, data set partitioning is done till the records in a single partition contain identical values. On the other hand, in the second phase branches containing noisy data are removed [17].

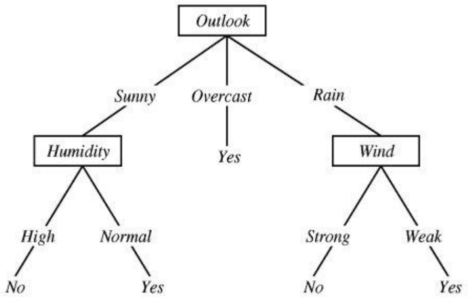

Figure 1. A Decision Tree

There are three tree growing method in decision tree namely CHAID, CART and QUEST. Exhaustive CHAID is another variant of CHAID.

A.1 CHAID (Chi-squared Automatic Interaction Detection)

This algorithm selects a set of predictors and their interactions and predicts the optimal value of the dependent variable. In the end a classification tree is obtained [15]. At each step, CHAID chooses the independent (predictor) variable that has the strongest interaction with the dependent variable. Categories of each predictor are merged if they are not significantly different with respect to the dependent variable[18].The CHAID analysis is advantageous when we are looking for patterns in complicated datasets. The variables can be categorical or interval in nature. Moreover not all the predictor variables need to be measured in the same level. CHAID is a useful method of summarizing the data and is analogous to stepwise regression.

A.2 CART (Classification and Regression Trees)

CART technique is more suited for continuous dependent variable and categorical predictor variable. CART splits the feature space recursively into non-overlapping regions. In order to predict the value of dependent categorical variable, a classification tree is generated [14].CART incorporates both testing with a test data set and cross-validation to assess the goodness of fit more accurately. CART can use the same variables more than once in different parts of the tree. This capability can uncover complex interdependencies between sets of variables. CART can be used in conjunction with other prediction methods to select the input set of variables.

A.3 QUEST (Quick, Unbiased, Efficient Statistical Tree)

QUEST technique provides unbiased feature selection and is also able to handle categorical variables with several categories [14]. QUEST stands for Quick, Unbiased and Efficient Statistical Tree. The QUEST node provides a binary classification method for building decision trees, designed to reduce the processing time required for large C&RT analyses while also reducing the tendency found in classification tree methods to favor predictors that allow more splits. Predictor fields can be numeric ranges, but the target field must be categorical. All splits are binary[19].

A.4 Exhaustive CHAID

Exhaustive CHAID is a variant of CHAID, where the algorithm performs a more thorough merging and testing of predictors for similar pairs until only one pair remains. Therefore, it takes much more computing time [14]. Splitting and stopping steps in Exhaustive CHAID algorithm are the same as those in CHAID. Merging step uses an exhaustive search procedure to merge any similar pair until only single pair remains. Also like CHAID, only nominal or ordinal categorical predictors are allowed, continuous predictors are first transformed into ordinal predictors before using the algorithm[18].

V. RESULTS AND ANALYSIS

The present research work aims at comparing the performance of variants of decision tree using SPSS statistics software (SPSS statistical package is one of the most popular statistical packages which can perform highly complex datamanipulation and analysis with simple instructions). In decision tree, there are classification and regression models in the form of tree structure. The variable for the root node is selected based on its predictive significance represented by its p-value. Each node represents one of the attributes of the customers. SPSS statistics software provides an inbuilt implementation of all the techniques namely, CHAID, exhaustive CHAID, CART and QUEST. Input data was fed and the above mentioned data mining techniques were applied. The results hence obtained were analyzed. Each technique gave a classification table, an associated risk factor table, a decision tree as the output. The decision tree so obtained had the root node as the factor which influenced the churn to the largest extent. The root node also denotes the attribute on the basis of which the splitting procedure was performed. In order to maintain simplicity, only classification table and the associated risk factor table for each technique has been shown below. Risk table represents the risk estimate and the standard error. It is a measure of tree’s predictive accuracy. For categorical dependent variable, risk estimate is the proportion of cases incorrectly classified after the adjustment of prior probabilities and misclassification cost. For scale dependent variable, risk estimate is within node variance. After analyzing the results so obtained, it has been observed that out of all the possible techniques, Exhaustive CHAID has outperformed and has given the most accurate results. This technique also has the minimum risk error associated with it.

TABLE I
CART

Classification

Sample	Observed	Predicted		
		False.	True.	Percent Correct
Training	False.	1401	49	96.6%
	True.	113	107	48.6%
	Overall Percentage	90.7%	9.3%	90.3%
Test	False.	1336	63	95.5%
	True.	139	123	46.9%
	Overall Percentage	88.8%	11.2%	87.8%

Growing Method: CRT
Dependent Variable: Churn

TABLE II
Risk Factor for CRT

Risk

Sample	Estimate	Std. Error
Training	.097	.007
Test	.122	.008

Growing Method: CRT
Dependent Variable: Churn

TABLE III
QUEST

Classification

Sample	Observed	Predicted		
		False.	True.	Percent Correct
Training	False.	1370	14	99.0%
	True.	192	51	21.0%
	Overall Percentage	96.0%	4.0%	87.3%
Test	False.	1458	8	99.5%
	True.	198	42	17.5%
	Overall Percentage	97.1%	2.9%	87.9%

Growing Method: QUEST
Dependent Variable: Churn

TABLE IV
Risk Factor for QUEST

Risk

Sample	Estimate	Std. Error
Training	.127	.008
Test	.121	.008

Growing Method: QUEST
Dependent Variable: Churn

TABLE V
CHAID RESULTS

Classification				
		Predicted		
Sample	Observed	False.	True.	Percent Correct
Training	False.	1407	43	97.0%
	True.	152	82	35.0%
	Overall Percentage	92.6%	7.4%	88.4%
Test	False.	1344	55	96.1%
	True.	154	94	37.9%
	Overall Percentage	91.0%	9.0%	87.3%

Growing Method: CHAID
Dependent Variable: Churn

TABLE VI
Risk Factor for CHAID

Risk		
Sample	Estimate	Std. Error
Training	.116	.008
Test	.127	.008

Growing Method: CHAID
Dependent Variable: Churn

TABLE VII
Exhaustive CHAID

Classification			
	Predicted		
Observed	False.	True.	Percent Correct
False.	2786	64	97.8%
True.	252	231	47.8%
Overall Percentage	91.1%	8.9%	90.5%

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: Churn

TABLE VIII
Risk Factor for Exhaustive CHAID

Risk	
Estimate	Std. Error
.095	.005

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: Churn

The accuracy of the techniques is calculated using the following formula,

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

VI. CONCLUSION

In order to maintain a loyal customer base, all the service providers in telecommunication industry aim to retain the customers with themselves. Since the cost associated with acquiring a new customer is much higher than retaining an old one, churn prediction becomes even more necessary. The presented paper has tried to analyze a large customer dataset for churn by using Decision Tree Classification Technique. After implementing all the possible variants of decision tree in SPSS it was observed that Exhaustive CHAID technique proved to be more efficient and accurate than others to predict the customers who are likely to churn in nearby future.

VII. REFERENCES

[1] Amal M. Almana, Mehmet Sabih Aksoy, Rasheed Alzaharni, “A Survey on Data Mining Techniques in Customer Churn Analysis for Telecom Industry”, International Journal of Engineering Research and Applications, Vol. 4, Issue 5, ISSN: 2248-9622, May 2014, Pp. 165-171.

[2] Clement Kirui, Li Hong, Wilson Cheruiyot, Hillary Kirui, “Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining”, International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, ISSN: 1694-0814, March 2013, Pp. 165-172.

[3] Wai-Ho Au, Keith C. C. Chan, Xin Yao, “A Novel Evolutionary Data Mining Algorithm With Applications to Churn Prediction”, IEEE Transactions on Evolutionary Computation, IEEE, Vol. 7, No. 6, December 2013, Pp. 532-545.

[4] Erfaneh Gharavi, Mohammad. Jafar Tarokh, “Predicting Customers’ Future Demand using Data Mining Analysis: A Case Study of Wireless Communication Customer”, 5th Conference on Information and Knowledge Technology, IEEE, 2013, Pp. 338-343.

[5] Wei Yu, Dawn N. Jutla, Shyamala C. Sivakumar, “A Churn-Strategy Alignment Model for Managers in Mobile Telecom”, Proceeding of the 3rd Annual Communication Networks and Services Research Conference, IEEE, 2005.

[6] Chao Zhu, Jiayin Qi, Chen Wang, “An Experimental Study on four Models of Customer Churn Prediction”, Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics, USA, October 2009, Pp. 3199-3204.

[7] Adnan Idris, Asifullah Khan, “Customer Churn Prediction for Telecommunication:Employing various features selectiontechniques and tree based ensemble classifiers”, IEEE, 2012.

[8] Navid Forhad, Md. Shahriar Hussain, Rashedur M Rahman, “Churn Analysis: Predicting Churners”, IEEE, 2014, Pp. 237-241.

[9] N.Kamalraj, Dr. A.Malathi, “Applying Data Mining Techniques in Telecom Churn Prediction”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 10, ISSN: 2277 128X, October 2013, Pp. 363-370.

[10] Javed Basiri, Fattaneh Taghiyareh, Behzad Moshiri, “A Hybrid Approach to Predict Churn”, IEEE Asia-Pacific Services Computing Conference, 2010, Pp. 485-491.

[11] Essam Shaaaban, Yehia Helmy, Ayman Khedr, Mona Nasr, “A Proposed Churn Prediction Model”, International Journal of Engineering Research and Applications, Vol. 2, Issue 4, ISSN: 2248-9622, June-July 2012, Pp. 693-697.

[12] Umman Tuğba Şimşek Gürsoy, “Customer Churn Analysis in Telecommunication Sector”, Istanbul University Journal of the School of Business Administration, Vol. 39, No. 1, ISSN: 1303-1732, 2010, Pp. 35-49.

[13] V. Umayaparvathi, K. Lyakutti, “Applications of Data Mining in Telecom churn Prediction”, International Journal of Computer Applications, Vol. 42, No. 20, ISSN: 0975-8887, March 2012, Pp. 5-9.

[14] Saad Ahmed Qureshi, Ammar Saleem Rehman, Ali Mustafa Qamar, Aatif Kamal, “Telecommunication Subscribers’ Churn Prediction Model Using Machine Learning”, IEEE, 2013, Pp. 131-136.

[15] Ning Lu, Hua Lin, Jie Lu, “A Customer Churn Prediction Model in Telecom Industry Using Boosting”, IEEE Transactions on Industrial Informatics, Vol. 10, No. 2, May 2014, Pp. 1659-1665.

[16] Nikita Jain, Vishal Srivastav, “Data Mining Techniques: A Survey Paper”, International Journal of Research in Engineering and Technology, Vol. 2, Issue 11, ISSN: 2321-7308, Nov 2013, Pp. 116-119.

[17] Vladislav Lazarov, Marius Capota, “Churn Prediction”, Technische Universität München.

[18] http://10.35.2.6:50624/help/index.jsp?topic=%2Fcom.ibm.spss.statistics.algorithms%2Falg_tree-chaid.htm

[19] Rahman Mansouri, Mohamad Saraee, RasoulAmirfattahi, “Applications of Data Mining in Predicting Cell Phones Subscribers Behavior Employing the Contact Pattern”, International Conference on Data Storage and Data Engineering, IEEE 2010.

[20] Khalida, Sunarti, Norazrina, Faizin, “Data Mining in Churn Analysis Model for Telecommunication Industry” Journal of Statistical Modeling and Analytics, Vol. 1, No.19-27,ISSN: 2180-3102, 2010, Pp. 19-27.

[21] Amal M. Almana, Mehmet Sabih Aksoy, Rasheed Alzaharni, “A Survey on Data Mining Techniques in Customer Churn Analysis for Telecom Industry”, International Journal of Engineering Research and Applications, Vol. 4, Issue 5, ISSN: 2248-9622, May 2014, Pp. 165-171.