Analysis of Customer Churn Prediction in Telecom Sector Using Logistic Regression and Decision Tree

Manoj Kumar Sahu[1], Dr. Rajeev Pandey[2], Dr. Sanjay Silakari[3]

[1] M. Tech. Research scholar, Department of computer science, UIT RGPV Bhopal, M.P., India

[2] Assistant professor, Department of computer science, UIT RGPV Bhopal, M.P., India

[3] HOD, Department of computer science, UIT RGPV Bhopal, M.P., India

***Abstract*:** Customer churn prediction in Telecom Industry is a core research topic in recent years. A huge amount of data is generated in Telecom Industry every minute. On the other hand, there is lots of development in data mining techniques. Customer churn has emerged as one of the major issues in Telecom Industry. Telecom research indicates that it is more expensive to gain a new customer than to retain an existing one. In order to retain existing customers, Telecom providers need to know the reasons of churn, which can be realized through the knowledge extracted from Telecom data. In this paper we can focuses on various data mining techniques for predicting customer churn, In this we can build the classification model using logistic regression and decision tree and model evaluation measures are computed and compare between the models.

Keywords: ***Churn prediction, data mining, telecom system, Customer retention, classification system, random forest, CART.***

I.INTRODUCTION

Studies revealed that gaining new customers is 5 to 10 times costlier than keeping existing customers happy and loyal in today s competitive conditions, and that an average company loses 10 to 30 percent of customers annually (Kotler 2009). Many companies, being aware of this fact, are engaged in satisfying and retaining the customers. Especially in the subscription oriented industries, such as telecommunications, banking, insurance, and in the fields of customer relationship management, etc., companies working with numerous customers, the revenues of the companies are provided by the payments made by these customers periodically. It is very important to be able to keep customers satisfied in order to be able to sustain this revenue with the least expenditure cost.

The objectives of this study are:

Reviewing the relevant studies about churn analysis on telecommunications industry presented in the last five years, particularly in the last two years, and introducing these up-to-date studies in the literature,

Determining the data mining methods frequently used in churn implementations,

Shedding a light on methods that can be used in further studies.

Data Mining and Customer Churn Analysis

In today's technological conditions, new data are being produced by different sources in many sectors. However, it is not possible to extract the useful information hidden in these data sets, unless they are processed properly. In order to find out these hidden information, various analyses should be performed using data mining, which consists of numerous methods. The Churn Analysis aims to predict customers who are going to stop using a product or service among the customers. And, the customer churn analysis is a data mining based work that will extract these possibilities. Today's competitive conditions led to numerous companies selling the same product at quite a similar service and product quality.

In the midst of this competition, the cost of gaining new customers is more than retaining the existing customers. For this reason, existing customers are very valuable. With the Churn Analysis, it is possible to precisely predict the customers who are going to stop using services or products by assigning a probability to each customer. This analysis can be performed according to customer segments and amount of loss (monetary equivalent). Following these analyses, communication with the customers can be improved in order to persuade the customers and increase customer loyalty. Effective marketing campaigns for target customers can be created by calculating the churn rate or customer attrition. In this way, profitability can be increased significantly or the possible damage due to customer loss can be reduced at the same rate (Argüden 2008). For example, if a service provider which has a total of 2 million subscribers, gains 750.000 new subscribers and losts 275.000 customers; churn rate is calculated as 10%. The customer churn rate has a significant effect on the financial market value of the company. So most of the companies keep an eye on the value of the customer at monthly or quarterly periods (Seker 2016).

II. LITERATURE REVIEW

According to [1] ,Since the beginning of data mining the discovery of knowledge from the Databases has been carried out to solve various problems and has helped the business come up with practical solutions. Large companies are behind improving revenue due to the increase loss in customers. The process where one customer leaves one company and joins another is called as churn. This paper will be discussing how to predict the customers that might churn, R package is being used to do the prediction. R package helps represent large dataset churn in the form of graphs which will help to depict the outcome in the form of various data visualizations. Churn is a very important area in which the telecom domain can make or lose their customers and hence the business/industry spends a lot of time doing predictions, which in turn helps to make the necessary business conclusions. Churn can be avoided by studying the past history of the customers. Logistic Regression is been used to make necessary analysis. To proceed with logistic regression we must first eliminate the outliers that are present, this has be achieved by cleaning the data (for redundancy, false data etc) and the resultant has been populated into a prediction excel using which the analysis has been performed.

According to [2] , Telecom Customer churn prediction is a cost sensitive classification problem. Most of studies regard it as a general classification problem use traditional methods, that the two types of misclassification cost are equal. And, in aspect of cost sensitive classification, there are some researches focused on static cost sensitive situation. In fact, customer value of each customer is different, so misclassification cost of each sample is different. For this problem, we propose the partition cost-sensitive CART model in this paper. According to the experiment based on the real data, it is showed that the method not only obtains a good classification performance, but also reduces the total misclassification costs effectively.

Customer churn research [5] is an important aspect of customer relationship management, based on the comparison of each algorithm and ensemble learning, theory and Practice, based on ensemble learning and selective ensemble learning is an effective means to predict customer churn, there are still a lot of problems, for example, how to choose the method of integration, how to choose the strategy, which makes the final ensemble classifier has the best generalization ability, and how to select the parameters of each algorithm as well as the kernel function calls to achieve the best results will be the focus of future research. On the whole, there is no absolute good classifier, in the face of different data using the appropriate classifier or classification method, can be classified forecast to do the best. When choosing a classifier, it is time to consider all aspects, horizontal and vertical contrast to find the most suitable, several classifiers are required to predict, based on the analysis of each focus, in order to get the most satisfactory results.

III PROBLEM DEFINITION

In a business setting, the term, client attrition merely refers to the purchasers exploit one business service to a different. client churn or subscriber churn is additionally kind of like attrition, that is that the method of shoppers shift from one service supplier to a different anonymously. From a machine learning perspective, churn prediction could be a supervised (i.e. labeled) downside outlined as follows: Given a predefined forecast horizon, the goal is to predict the longer term churners over that horizon, given the info related to every subscriber within the network. The churn prediction downside diagrammatical here involves three phases, namely, i) the training part, ii) testing part, iii) prediction section. The input for this downside includes the info on past necessitate every mobile subscriber, along with all personal and business data that's maintained by the service supplier. Additionally, for the training section, labels are provided within the type of an inventory of churners. When the model is trained with highest accuracy, the model should be able to predict the list of churners from the important dataset that doesn't embody any churn label. Within the perspective of information discovery method, this downside is categorized as prognostic mining or prognostic modeling.

IV PROPOSED WORK

In the proposed system R [8] programming will be used to build the model for churn prediction. It is widely used among statisticians and data miners for developing statistical software and data analysis. R is freely available and a powerful statistical analysis tool which has not yet been explored for building model for churn prediction [7].

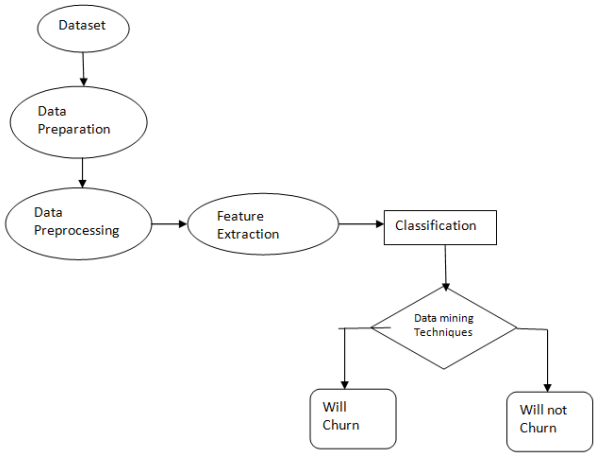

Figure1. Churn Prediction Framework

This is where the churn prediction model [4] can help the business to identify such high risk customers and thereby helps in maintaining the existing customer base and increase in revenues. Churn prediction is also important because of the fact that acquiring new customers is much costly than retaining the existing one. As the telecom users are billions in number even a small fraction of churn leads to high loss of revenue. Retention has become crucial especially in the present situation because of the increasing number of service providers and the competition between them, where everyone is trying to attract new customers and lure them to switch to their service. With a large customer base and the information available about them data mining techniques proves to be a viable option for making predictions about the customers that have high probability to churn based on the historical records available. The data mining techniques can help find the pattern among the already churned customers and provide useful insights which can then be used strategically to retain customers.

V EXPERIMENTAL & RESULT ANALYSIS

All the experiments were performed using an i5-2410M CPU @ 2.30 GHz processor and 4 GB of RAM running Windows. After that we can install r base core on windows and Rstudio and then to identify trends in customer churn at a telecom company. The data given to us contains 3,333 observations and 21 variables extracted from a data warehouse. First we can clean the data by deleting columns which cannot be used for prediction, such as Phone and State. Then convert char variables to factors.

Logistic Regression Model

The LASSO (Least Absolute Shrinkage and Selection Operator) is a regression method that involves penalizing the absolute size of the regression coefficients. By penalizing, you end up in a situation where some of the parameter estimates may be exactly zero. The larger the penalty applied, the further estimates are shrunk towards zero. This is convenient for variable selection.Now we are using glmnet package to fit lasso for logistic regression model.

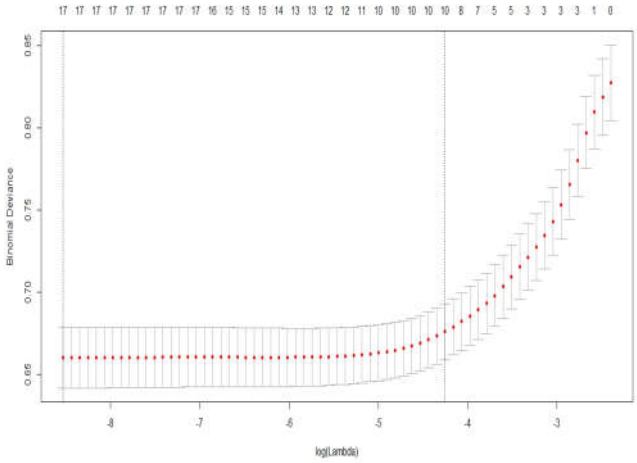

Figure 2. Binomial deviance of logistic regression model

The left line corresponds to the model with best lamda value. The right line shows the best model within 1 standard error. And then we can make prediction on the whole dataset and we can also evaluate the model with ROC curve.

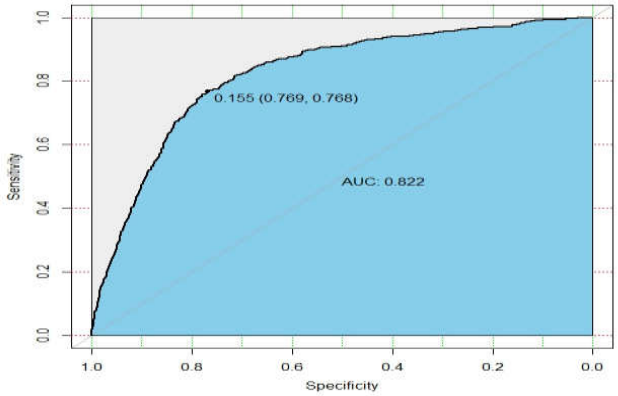

Figure 3. ROC of logistic regression model

ROC curve shows the tradeoff between sensitivity and specificity: The measure of **sensitivity** is the proportion of positive examples that were correctly classified. The measure of **specificity** is the proportion of negative examples that were correctly classified.

AUC: The area under the curve. It is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

Decision Tree Model

After logistic regression model we can also take decision tree model to analyse and predict churns in telecom sector dataset. For which we can make a tree model by using rpart package and then visualize the tree model.

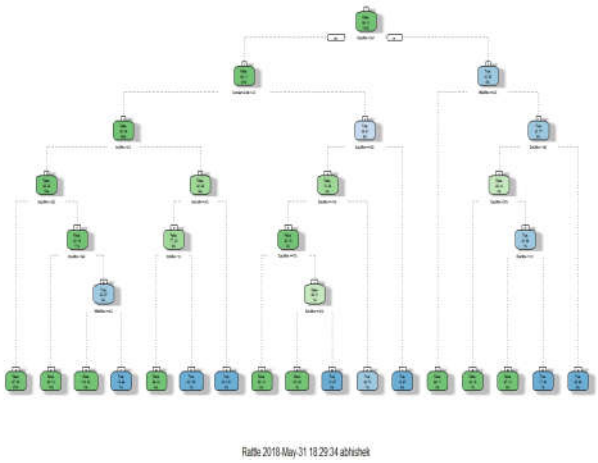

Figure 4. Tree model

To validate the model we use the printcp and plotcp functions. CP stands for Complexity Parameter of the tree. printcp() provides the optimal pruning's based on the cp value. We prune the tree to avoid any overfitting of the data. The convention is to have a small tree and the one with least cross validated error given by printcp() function i.e. xerror .and we can select the optimal cp value associated with the minimum error. According to the prints and plotcp result, the original tree with all 17 factors is the optimal decision without over fitting. There is no need to prune the tree. And then we can make prediction on the whole dataset and we can also evaluate the model with ROC curve.

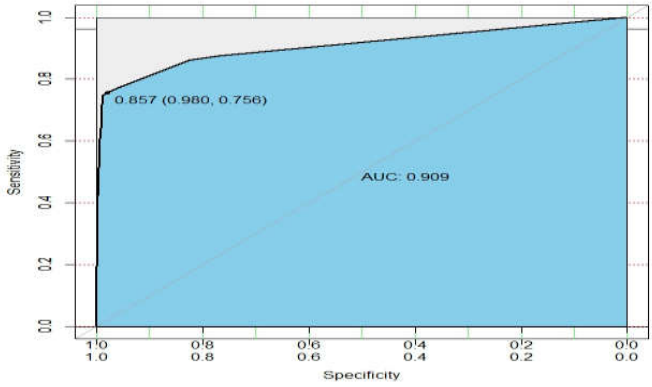

Figure 5. ROC of decision tree model

By comparing the AUC value (0.820 vs. 0.909), and the criteria in ConfusionMatrix, the Decision Tree Model performs better than Logistic Regression.

VI CONCLUSION

In this paper, it is observed that decision tree model has better in the prediction of churn because it has more accuracy than logistic regression classification model and it is also easy to construct. By comparing the AUC value (0.820 vs. 0.909), and the criteria in Confusion Matrix, the Decision Tree Model performs better than Logistic Regression.

REFERENCES

[01] Helen Treasa Sebastian and Rupali Wagh, " Churn Analysis in Telecommunication using Logistic Regression " in ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY, An International Open Free Access, Peer Reviewed Research Journal, ISSN: 0974-6471 March 2017, Vol. 10, No. (1): Pgs. 207-212.*

[02] Chuanqi Wang, Ruiqi Li, Peng Wang, Zonghai Chen, Partition cost-sensitive CART based on customer value for Telecom customer churn prediction in Proceedings of the 36th Chinese Control Conference 2017 IEEE.

[03] Guo-en Xia, Hui Wang, Yilin Jiang, Application of Customer Churn Prediction Based on Weighted Selective Ensembles in The 2016 3rd International Conference on Systems and Informatics (ICSAI 2016), IEEE 2016.

[04] Rahul J. Jadhav, Usharani T. Pawar, Churn Prediction in Telecommunication Using Data Mining Technology , in (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.2, February 2011

[05] Kiran Dahiya, Surbhi Bhatia, Customer Churn Analysis in Telecom Industry in IEEE 2015, 978-1-4673-7231-2/15

[06] N.Kamalraj, A.Malathi A Survey on Churn Prediction Techniques in Communication Sector in International Journal of Computer Applications (0975 8887) Volume 64 No.5, February 2013

[07] Kiran Dahiya,KanikaTalwar, Customer Churn Prediction in Telecommunication Industries using Data Mining Techniques- A Review in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.

[08] R Data: http://cran.r-project.org/

[09] Data Mining in the Telecommunications Industry‖, Gary M. Weiss, Fordham University, USA.