Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression

Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, Prof. V. A. Kanade
Department of Computer Science,
PVG's COET,
Pune, India

***Abstract*— Customer churn prediction in Telecom industry is one of the most prominent research topics in recent years. It consists of detecting customers who are likely to cancel a subscription to a service. Recently, the mobile telecommunication market has changed from a rapidly growing market into a state of saturation and fierce competition. The focus of telecommunication companies has therefore shifted from building a large customer base into keeping customers in house. For that reason, it is valuable to know which customers are likely to switch to a competitor in the near future. The data extracted from telecom industry can help analyze the reasons of customer churn and use that information to retain the customers. We have proposed to build a model for churn prediction for telecommunication companies using data mining and machine learning techniques namely logistic regression and decision trees. A comparison is made based on efficiency of these algorithms on the available dataset.**

Keywords—Churn Prediction; Logitic Regression; Decision Trees; CRM(Customer Relationship Management)

I. INTRODUCTION

Customer churn refers to when a customer switches from one service provider to another. Churn is a problem for any provider of a subscription service or recurring purchasable. The focus of this paper is mainly on telecom industry because of its tremendous growth in the recent years. With easy communication and a number of service providers almost everyone today has a telecom subscription. Churn is especially important to mobile phone service providers because it is easy for a subscriber to switch services. Phone number portability has removed the last important obstacle. Churn Prediction model can help analyze the historical data available with the business to find the list of customers which are at high risk to churn. This will help the telecom industry to focus on a specific group rather than using retention strategies on every customer. Individualized customer retention is difficult because businesses usually have a big customer base and cannot afford to spend much time and money for it. However, if we could predict in advance which customers are at risk of leaving, we can reduce customer retention efforts by directing them solely toward such customers.

This is where the churn prediction model 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.

II. RELATED WORK

A lot of research has been done in the field of CRM(Customer Relationship Management) in various industries for retention of customers and develop strategies to build an efficient model so that specific group of customers can be targeted for retention. Various data mining and statistical techniques have been used for churn prediction of which some famous techniques include Decision trees, Regression models, Neural Networks, Clustering, Bayesian Models, SVM, etc. In [1] the authors have proposed a hybrid learning model to predict churn in mobile Telecommunication networks. Their model is built using WEKA, a well-known tool of Machine Learning. They have proposed that DM(Data Mining) can detect the customers with high propensity to churn, but not necessarily providing the reason of churn. The goal of their study was to show that hybrid models built on DM techniques can explain the churn behaviour with more accuracy than single methods; and that in some extend the reason of churn can be revealed. They used Logistic

978-1-5090-0669-4/16/$31.00 © 2016 IEEE

Regression in parallel with Voted Perceptron for classification, and then combined with clustering for churn prediction.

Qureshi et al. [2] in their examination take into account active churners in the Telecom industry by applying various methods of data mining such as K-Means Clustering, Logistic Regression , Neural Network, Linear &,Exhaustive CHAID, CART, QUEST, &CHAID. They found that Exhaustive CHAID performed well related to all other methods. In [5] the authors have used decision tree and neural network for churn prediction. They have delineated the process of churn prediction right from data acquisition to churn analysis. They have also focused on pre-processing where data cleaning, feature abstraction is used before giving it as input to the algorithms. In their study, they observed that decision tree model surpasses the neural network model in the prediction of churn and it is also easy to construct. They have proposed that selecting the right combination of attributes and fixing the proper threshold values may produce more accurate results. Their study limits itself with prediction of churn and no steps were analyzed to include retention policies.

A recent review based on the techniques used for churn prediction [3] Decision tree based techniques, Neural Network based techniques and regression techniques are generally applied in customer churn. In the review they have stated that decision tree based techniques specially C5.0 and CART have outperformed some of the existing data mining techniques such as regression in terms of accuracy.

Hence as seen in the literature survey the most popular techniques used for churn prediction include decision trees, neural networks and logistic regression out of which we have proposed to study the performance of decision trees C5.0 and logistic regression on a telecom data publicly available.

Although neural networks also perform very well in case of churn prediction it has limitations such as it performs well for large datasets only and takes a lot of time for training even with small datasets. Also with neural networks it is difficult to analyze the features that lead to churn. Neural networks is a interesting tool, but has operative problem for marketers: they are like black boxes; it is very difficult to understand the underlying factors that explain the churn. If we cannot know the factors, we have no insight to build churning prevention strategy. Hence Decision tress and logistic regression is used which help analyze the factors causing churn in an effective and understandable way.

Logistic Regression helps understand the degree to which each feature affects the decision of churn and decision tree provides a graphical overview of the available data from which rules can be generated and strategies can be build for customer retention. Previously decision trees and logistic regression has been used for churn predication for credit card customers. The study and comparison of these 2 algorithms has not yet been explored for telecom industry. One being a popular data mining technique and other is a statistical model can provide a fair idea to telecom industry about which algorithm to adopt and suits their data.

III. PROPOSED SYSTEM

In the proposed system R 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.

The system will have three main options namely View performance analysis – which displays the results obtained by applying logistic regression and decision tree on the available dataset, Testing – to construct a list of customers which have a high probability to churn from the input, given that the attributes of the input data are same as the available dataset used for training, Training and testing – which builds a model along with generating a churn list if any other type of dataset is provided. In performance analysis the results after using logistic regression and decision trees on the available dataset is illustrated using confusion matrix analysis. In the next operation the user can provide data for testing the system provided the features of the data are same as that used for training using the publicly available dataset.

If the user wishes to use the system for their data of different format and features they can do so by using the training and testing option where the system will use the new training data to build the model first and then use it for testing. Hence the system is not limited to use only for some specific type of data. To make the system robust data cleaning will be included in the training and testing phase explicitly. The system is made usable and lucid through a web interface that will graphically depict the results of the system.

The system building has three main phases i.e. developing the web interface module, feature extraction module and prediction module. The web interface will provide a graphical overview of the results obtained, which can be created using R package called Shiny. The feature extraction module will consist of estimation of parameters in logistic regression and generation of rules for decision trees. In decision tree algorithm the information gain is calculated for each feature and the maximum value feature is used to split the dataset, this is continued until all the features are used and then the features which do not provide enough information are pruned from the tree to get an optimized tree for the best possible estimate.

Logistic regression uses maximum likelihood estimation for transforming the dependent variable into a logistic variable. Logistic regression uses the linear regression function to estimate the value of dependent variable by estimating the parameters for the linear equation. As shown in (1) α, b_1, $b2...b_n$ are the parameters to be calculated using the training data and the equation will then be used to predict P(X) which represents the dependent variable value if values of features $x_1, x_2, \ldots x_n$ are given.

$$P(X) = \alpha + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n, \quad (1)$$

Logistic regression is used when the output is binary i.e. of the form yes or no, 0 or 1, etc. As the output obtained by the above equation is a real valued number it needs to be converted in a form appropriate for making the prediction. Hence for this purpose logic or the sigmoid function is used to convert the output of linear regression into a probability value and its equation is as follows:

$$Q(X) = \frac{1}{1 + e^{-(\alpha + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)}}. \quad (2)$$

$$0 \leq Q(X) \leq 1$$
$$-\infty < P(X) < +\infty$$

As the value of Q(X) is a number between 0 and 1 this value can be considered as the probability of the particular outcome. For example if the output is 0.8 it means there are 80% chances of getting the output as 1 and it can therefore be safely predicted that for the given set of input attributes the output would be 1. The prediction for the negative case or 0 can be calculated as 1 minus the probability of positive case. For the above example it will be 0.2 which is far less and hence the prediction is given to be as 1. Normally 0.5 is used as threshold to decide what prediction is to be given. Any value above 0.5 is predicted as positive case and anything below as negative case value.

In the next phase predictions are made based on the parameter estimates obtained from logistic regression and the rules obtained from decision trees.

Figure 1 gives the overview of the proposed system:

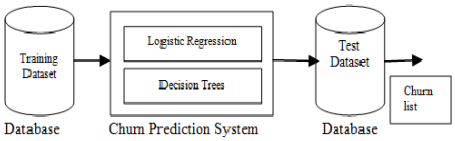

Fig. 1. System Architecture

The input to the system includes the options, training data and test data. Where Option comprise of View Performance, Train and Test, Test. Training Dataset Attributes contain State, Area Code, Phone, Day Mins, Eve Mins, Night Mins, Intl Mins, CustServ Calls, Int'l Plan, VMail Plan, Day Calls, Day Charge, Eve Calls, Eve Charge, Night Calls, Night Charge, Intl Calls, Intl Charge, Churn.

The churn prediction system will build the prediction model based on decision tress and logistic regression algorithm which will then be used on the testing dataset to measure the accuracy of the system. Finally the system will output the list of customers which have high probability and are possible churners in the future along with the accuracy of the system.

CONCLUSION

The proposed system provides a statistical survival analysis tool to predict customer churn based on comparison between decision trees and logistic regression. Selecting the right combination of attributes and fixing the proper threshold values may produce more accurate results of predicting churn customers. The proposed model suggests that data mining techniques can be a promising solution for the customer churn management. Using this model the telecom companies can predict in advance which customers are at risk of leaving, and can target those customers consequently saving a lot of revenues namely the ones which is used for replacing the lost customers and also the ones that are wasted for retaining already loyal customers.

The best churn model is not the one with best statistical precision but the one that provide best insights to further prevent churn behavior. By means of the results obtained using decision trees and logistic regression it will be easy to design retention policies and tactics to help maintain the customers as these methods provide easily deducible explanations about the reasons behind decision of churn along with the list of customers with high probability to churn.

Also the user can find which algorithm suits best for their data using our comparative analysis and thus save time which is used for finding an accurate model for their data.

ACKNOWLEDGMENT

We would like to take this opportunity to express our sincere gratitude to the Head of Department Prof.G.V.Garje and our internal project guide Prof V.A.Kanade for her guidance, immense help and encouragement during the designing of the proposed system.

REFERENCES

[1] A Hybrid Churn Prediction Model in Mobile Telecommunication Industry ,Georges D. Olle Olle and Shuqin Cai ,International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 4, No. 1, February 2014.

[2] Telecommunication Subscribers' Churn Prediction Model Using Machine Learning, Saad Ahmed Qureshi, Ammar Saleem Rehman, Ali Mustafa Qamar, Aatif Kamal, Ahsan Rehman, IEEE International Conference on Digital Information Management (ICDIM), 2013 Eighth International Conference on, 2013, pp. 131–136.

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

[4] Churn Prediction in Telecommunication Using Classification Techniques Based on Data Mining: A Survey, Nisha Saini and Monika, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 3, March 2015.

[5] Churn Prediction In Mobile Telecom System Using Data Mining Techniques, Dr. M. Balasubramaniam, M.Selvarani, International Journal of Scientific and Research Publications, Volume 4, Issue 4, April 2014.

[6] Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining, Clement Kirui1, Li Hong, Wilson Cheruiyot and Hillary Kirui, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013.

[7] Applying Data Mining Techniques in Telecom Churn Prediction, N.Kamalraj and Dr.A.Malathi, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.

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

[9] Analysis of Customer Churn in Mobile Industry using Data Mining, Aishwarya Churi, Mayuri Divekar, Sonal Dashpute, Prajakta Kamble, International Journal of Emerging Technology and Advanced Engineering, Volume 5, Issue 3, March 2015.

[10] A Proposed Churn Prediction Model, Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr / International Journal of Engineering Research and Applications (IJERA) ,Vol. 2, Issue 4, June-July 2012.