Customer Churn Analysis in Telecom Industry

Kiran Dahiya
Computer Science Department
Manav Rachna College of Engineering
Faridabad, India
kirandahiya23@gmail.com

Surbhi Bhatia
Computer Science Department
Echelon Institute of Technology
Faridabad India
surbhibhatia1988@yahoo.com

Abstract - With the rapid development of telecommunication industry, the service providers are inclined more towards expansion of the subscriber base. To meet the need of surviving in the competitive environment, the retention of existing customers has become a huge challenge. In the survey done in the Telecom industry, it is stated that the cost of acquiring a new customer is far more that retaining the existing one. Therefore, by collecting knowledge from the telecom industries can help in predicting the association of the customers as whether or not they will leave the company. The required action needs to be undertaken by the telecom industries in order to initiate the acquisition of their associated customers for making their market value stagnant. Our paper proposes a new framework for the churn prediction model and implements it using the WEKA Data Mining software. The efficiency and the performance of Decision tree and Logistic regression techniques have been compared.

Keywords- Churn prediction, Data mining, Decision tree, Neural Network, Customer Relationship Management

1. INTRODUCTION

Due to the rapid growth in the data communication network and advancement in the Information Technology, a massive amount of data is available. With the increase in the competition in the market, companies have devoted their time more in making their previous clients associated with them rather than convincing the new clients. This has been justified by Van Den Poel and Larivi_ere [1] who surveyed on the importance of the economic value of customer retention. Since the major source of profit are customers, so customer churn plays a significant role in the survival and development of telecommunication industry. Therefore tools have to be developed for predictive modeling and classification of various systematic tasks [2]. The association of Customer Relationship Management helps in capturing consumer information and the organization further use this information to satisfy customer needs [3]. In order to improve and analyze the customer acquisition and retention, CRM tools have been developed for increase in the profitability and help in the predictive modeling and classification of various logical tasks [2]. Data mining plays a very important role in the telecommunication companies to improve their marketing efforts, identify fraud, and better manage their telecommunication networks [4]. Data mining techniques are applied in telecommunications for CRM because of the rapid growth of the huge amount of data; high pace in the market competition and increase in the churn rate [5]. Customer acquisition and retention can be improved by applying CRM tools for increasing profit and for supporting analytical tasks [2]. A lot of scope has evolved its way for the researchers for analysis of the data and to present the complete information for promoting their business because of the hidden data in telecom industries.

The goals of the paper are as follows:

- To define and explain the related terms in churn prediction modeling
- To find the gaps in the existing literature.
- To explicate the applications of data mining techniques.
- To propose the novel framework that uses churn prediction in Data Mining
- To evaluate the techniques used in the churn prediction

The paper has been organized as follows. Section 1 of the paper introduces the concept of churn prediction. Section 2 discusses the background. Section 3 presents the literature survey. Section 4 discusses the data mining techniques. Section 5 explains the proposed architecture. Section

978-1-4673-7231-2/15/$31.00 ©2015 IEEE

6 shows the implementation and analyzes the results and the last section concludes

2. BACKGROUND

In this secondary study carried out by C. Wei, it was pointed that majority of papers published on the topic mainly employed classification analysis for the construction of churn prediction models but were limited in their scope of calling pattern changes. However, it has been observed that many organizations are consistent with their research in the field of churn prediction in telecommunication industry. This is one of the reasons of creation of high percentage of tools. The following section focus on the basics of churn prediction model. Some of the various machine learning methods are Decision trees [15], Logistic regression technique [16] and neural networks [18] [19] which we will use in our work.

DEFINITION OF CHURN

It was discovered by Berson et al. (2000) noted that 'customer churn' is defined as the process of subscribers (either prepaid or post paid) switching from one service provider. Churn can be active / deliberate, rotational / incidental, passive / non-voluntary [6]. With proper management of customers, we can minimize the susceptibility to churn and maximize the profitability of the company A mechanism needs to be established to analyze the attributed of profitability. Churn Prediction can also be described as a method which helps in identifying possible churners in advance [7].

DATA MINING

Data mining come under the process of KDD (Knowledge discovery process). It is used to extract useful knowledge in the form of patterns from the different web sources such as databases, files etc. Nowadays data mining tools are be used to answer questions business that were earlier too time consuming and difficult to answer. The techniques of data analysis and the tools that help in the extraction of interesting hidden patterns play a vital role in the decision making process [8] [9]. The model has six phases which is shown in the figure 1.

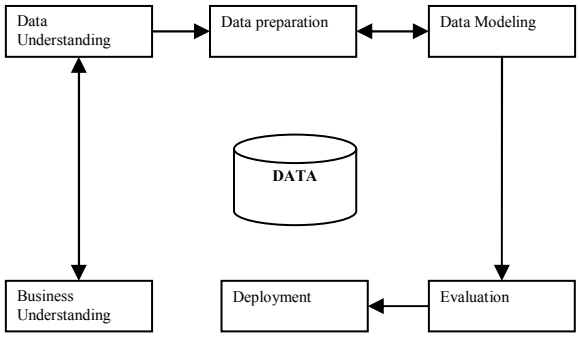

Figure 1: Data Mining Model

CUSTOMER RELATIONSHIP MANAGEMENT (CRM)

CRM is used to analyze the customer activities in order to raise the value of their customer portfolio. It gives a technological solution that helps in improving the targeting efforts by preparing databases and applying the automation tools bridge sales and marketing functions. CRM applications have complete information of each individual customer. During the process of registration, calls to support hotlines, etc, will bring outstanding results for advertising and marketing purposes if proper analysis is being done.

3. LITERATURE REVIEW

Web Chin-Ping Wei and I-Tang Chiu [10] proposed the churn prediction technique for customer retention analysis. The author used the decision tree approach C4.5 on customer call details. Yi-Fan wang, Ding-An chlang and Mei-Hua Hsu [11] discussed a Recommender system for customer churn by proposing a decision tree algorithm. Data used for the analysis has covered over 60,000 transactions and of more than 4000 members, over a period of three months. Jadhav and Pawar [12] designed a decision support system using data mining technique. The churn behavior of customers is predicted in advance using this technique. The authors have used Back propagation algorithm on a customer billing data. Tomas Philip Rúnarsson, Ólafur Magnússon, Birgis Hrafnkelsson [13] constructed a churn prediction model that can output the probabilities that customers will churn in the near future. In this paper the training data is used to build classifiers by using machine learning methods. N.Kamalraj and A.Malathi [14] focussed their research on the better understanding of churn prediction using data

mining techniques. Telecommunication industry can use this approach to customer retention activities within the context of their Customer Relationship Management efforts. The author uses the DM technique on customer details.

4. PROPOSED WORK

The study of predicting which persons are going to churn in advance will help the telecommunication industry and the CRM department to identify which persons are going to leave the network. The problem of our work discussed is the classification problem i.e. to classify each subscriber as potential churner or potential non churner. The framework discussed below is based on the Knowledge Discovery Data (KDD) process [19]. Our framework consists of the following five modules:

Data Acquisition: Acquiring data from the teleset industry is a big task because of the fear of misusing it. The data set for this study acquired from the KDD Cup 2009. It is used to analyze the marketing tendency of customers from the large databases from the French Telecom company Orange [20].

Data Preparation**:** Since the dataset acquired cannot be applied directly to the churn prediction models, so aggregation of data is required where new variables are added to the existing variables by viewing the periodic usage behavior of the customers. These variables are very important in predicting the behavior of customers in advance as they contain critical information used by the prediction models.

Data Preprocessing: Data preprocessing is the most important phase in prediction models as the data consists of ambiguities, errors, redundancy which needs to be cleaned beforehand. The data gathered from multiple sources first is aggregated and then cleaned as the complete data collected is not suitable for modeling purposes. The records with unique values do not have any significance as they do not contribute much in predictive modeling. Fields with too many null values also need to be discarded.

Data Extraction: The attributes are identified for classifying process. In our work, we have worked with numerical and categorical values.

Decision: The rule set will let the subscribers identify and classify in the different categories of churners and non churners by setting a particular threshold value.

The framework design used in our work is given in Figure 2. We have taken orange dataset which consists of total 18000 attributes.

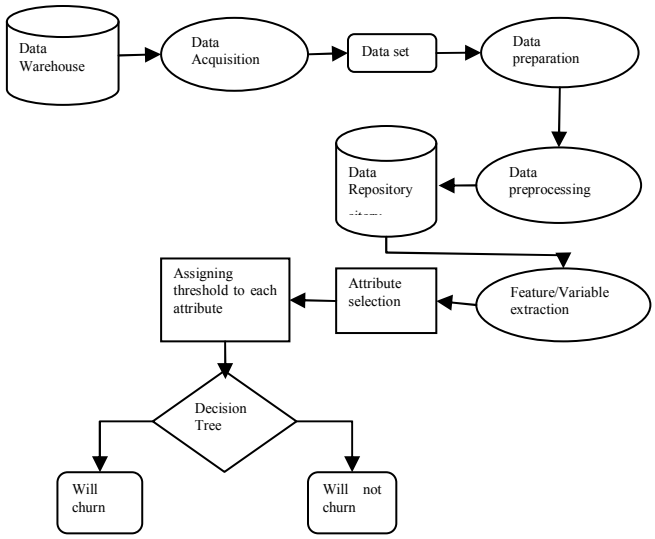

Figure 2: Churn Prediction Framework

5. IMPLEMENTATION

Decision Tree and Logistic Regression are used in our study for building the churn prediction model. The test records counts predicted correctly and incorrectly are evaluated on the performance of a classification model.

There are three datasets small, medium and large with varying attributes.

We have 10 numeric variables and 50 instances for the small dataset.

There are 50 numeric variables and 200 instances for medium dataset and for the large dataset, we have 100 variables (numeric variables and categorical variables) and 608 instances.

SNAPSHOTS

The snapshots of the Small dataset are shown from figure 3 to figure 5.

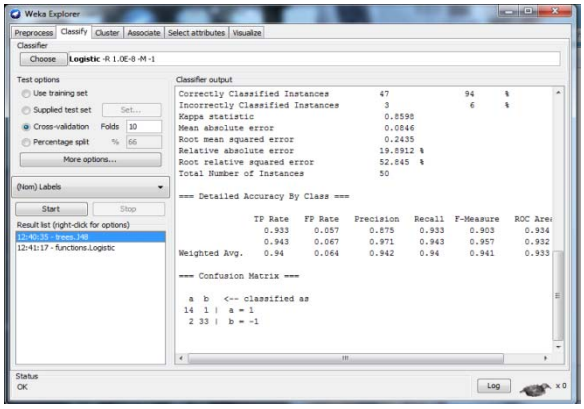

Figure 3: Weka J-48 Classifier

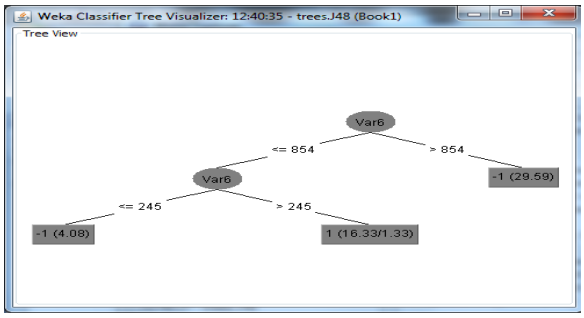

Figure 4: Weka J-48 Tree Visualizer

The Figure depicts the implementation of Decision tree J-48.

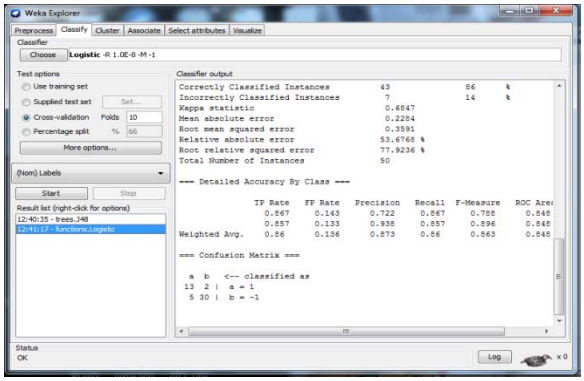

Figure 5: Weka Logistic Regression Classifier

The figure depicts the results with logistic regression.

The snapshots of the Medium Dataset are shown from figure 6 to figure 8.

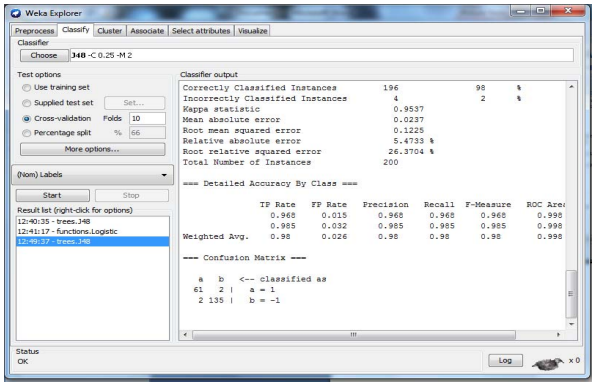

Figure 6: Weka J-48 Classifier

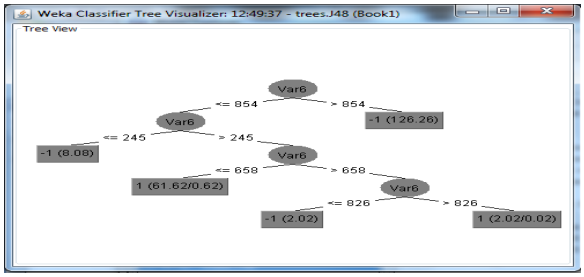

Figure 7: Weka J-48 Tree Visualizer

The Figure depicts the implementation of Decision treeJ-48.

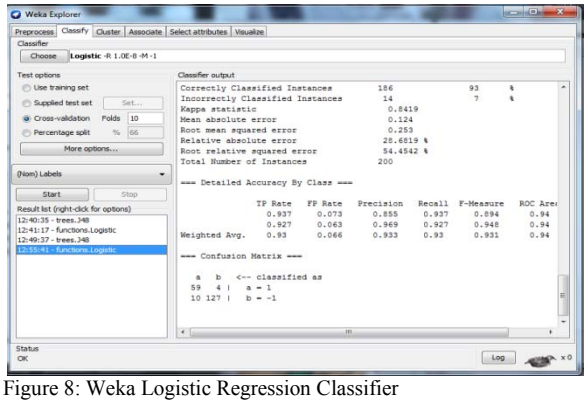

Figure 8: Weka Logistic Regression Classifier

The figure shows the results with logistic regression medium dataset.

The snapshots of the Large Dataset are shown from figure 9 to figure 12.

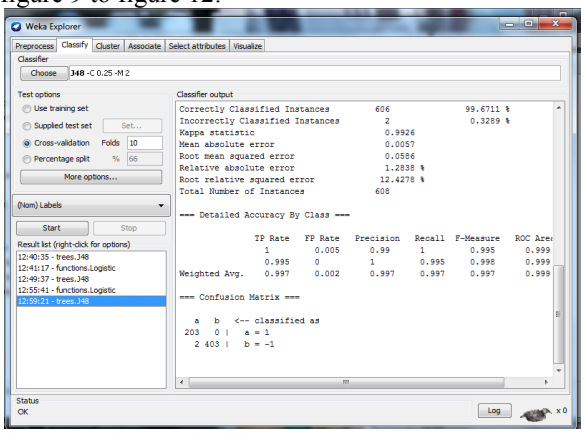

Figure 9: Weka J-48 Classifier

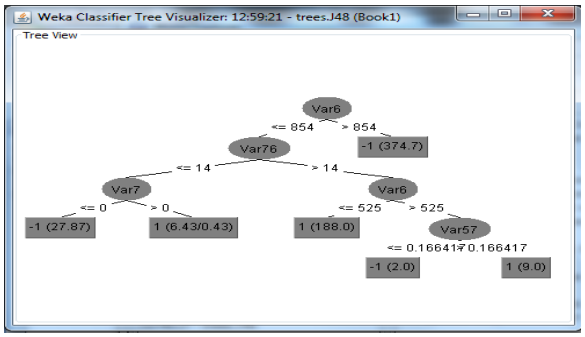

Figure 10: Weka J-48 Tree Visualizer

The Figure depicts the implementation of Decision tree J-48.

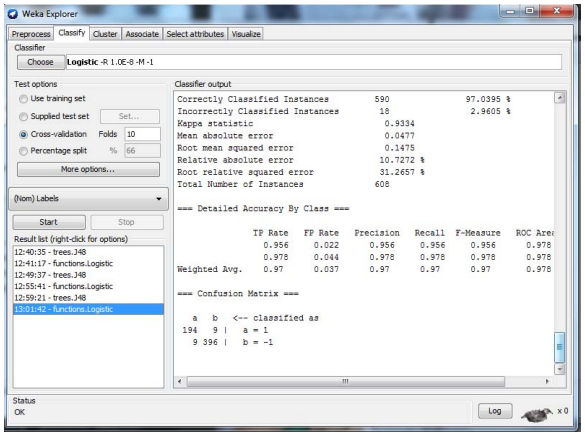

Figure 11: Weka Logistic Regression Classifier

The figure shows the results with logistic regression large dataset.

6. RESULTS

The results are shown in Figure 12.

Data Mining Techniques	Small Data set		Medium Data set		Large Data set	
	Churned	**Not Churned**	**Churned**	**Not Churned**	**Churned**	**Not Churned**
J-48 Decision Tree	47	3	96	4	606	2
Logistic Regression	43	7	86	14	590	18

Figure 12: Analysis of outcomes in data mining techniques

The performance of the above two techniques is evaluated by calculating the accuracy and error rate by using the given formula below:

Accuracy = Number of True Outcomes/Total Number of Predictions

Error Rate = Number of False Outcomes/Total Number of Predictions

On calculating the accuracy, we received the following results shown in the graph in Figure 13.

.

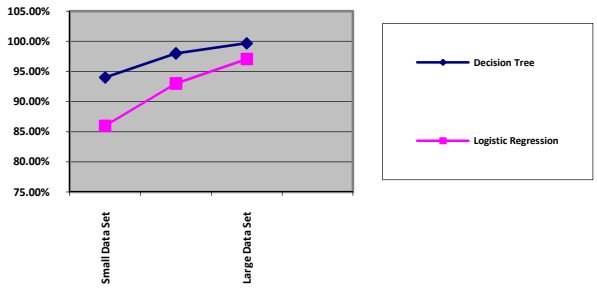

Figure 13: Performance matrix

We analyzed that selecting the correct grouping of attributes and setting up the proper threshold values can give more accurate results.

7. CONCLUSIONS

Telecommunication industry has suffered from high churn rates and immense churning loss. Although the business loss is unavoidable, but still churn can be managed and kept in an acceptable level. Good methods need to be developed and existing methods have to be enhanced to prevent the telecommunication industry to face challenges. In this paper we discussed the various prediction models and also compared the quality measures of pre-diction models like regression analysis, decision trees. We found that the accuracy achieved with decision tree is far much higher than the logistic regression technique which clearly states that decision tree is an efficient technique.

8. FUTURE SCOPE

The future scope of this paper will use hybrid classification techniques to point out existing association between churn prediction and customer lifetime value. The retention policies need to be

considered by selecting appropriate variables from the dataset. The passive and the dynamic nature of the industry ensure that data mining has become increasingly significant aspect in the telecommunication industry prospect.

REFERENCES

[1] D. V. Poel and B. Larivi. Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research, 157(1):196{217, 2004.

[2] V. Lazarov, M. Capota. Churn Prediction. Journal, Bus. Anal. Course. TUM Comput. Sci. 2007 publisher-Citeseer.

[3] R. Baran, Christopher, M. Zerres, Customer Relationship Management. Book, 2012.

[4] B V Chowdary, A. G. Raju, B. Anuradha, R.Changala, Decision Tree Induction Approach for Data Classification Using Peano Count Trees. Volume 2, Issue 4, April 2012 ISSN: 2277 128X.

[5] A. Ntoulas, P. Zerfos, J.Cho. Downloading Textual Hidden Web Content Through Keyword Queries. In: 5th ACM/IEEE Joint Conference on Digital Libraries (Denver, USA, Jun 2005) JCDL05, pp. 100-109.

[6] Shin-Yuan Hung , David C. Yen , H. Wang, Applying data mining to telecom, Expert Systems with Applications 31 (2006) 515–524, Elseiver.

[7] V. Umayaparvathi, K. Iyakutti, Applications of Data Mining Techniques in Telecom Churn Prediction, International Journal of Computer Applications (0975 – 8887) Volume 42– No.20, March 2012.

[8] O.R. Zaiane,Introduction to Data Mining, CMPUT^() Principles of Knowledge Discovery in Databases Chapter,pp.1-15,1999.

[9] Andrew H. Karp, Using logistic regression to predict customer retention

[10] C. Wei, I. Chiu, Turning telecommunication call details to churn prediction :a data mining Approach expert System with applications (2002).

[11] Y. Wang, D. Chlang, M. Hua Hsu ,A recommender system to avoid customer churn ,"Expert System with application, 2009.

[12] R. Jadhav and U. T. Pawar, Churn Prediction in Telecommunication Using Data Mining Technology," in Proc. The (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.2, February 2011.

[13] Emilía Huong Xuan Nguyen ,Customer Churn Prediction for the Icelandic Mobile Telephony Market , in proc. The Faculty of Industrial Engineering, Mechanical Engineering and Computer Science University of Iceland in September 2011.

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

[15] L. Yangi , C. Chiu , Subscriber Churn Prediction in Telecommunications,

[16]Logistic Regression at www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf]

[17] Maureen Caudill , Neural Network Primer: Part I, AI Expert, Feb. 1989].

[18] C. Stergiou, D. Siganos, Neural Networks by www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

[19] Jiawei Han and Micheline Kamber, Data mining, concept and techniques" http://www.cs.sfu.ca].

[20] I. Guyon, V. Lemaire, M. Boull´e, G. Dror , D. Vogel, Analysis of the KDD Cup 2009: Fast Scoring on a Large Orange Customer Database, JMLR: Workshop and Conference Proceedings 7: 1-22.