STM JOURNALS
Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720
Volume 6, Issue 3
www.stmjournals.com

Generalized Churn Prediction

Saima Jan**, *Afaq Alam Khan
Department of Information Technology, Central University of Kashmir, Ganderbal,
Jammu and Kashmir, India

Abstract
Customer churn evaluation and management are important to determine the loyalty of customers with the firm and thus to remain aware of each customer whether he will stay or leave the firm. This, in turn, will be used to determine the value of the firm and hence its place in the market. This paper includes existing literature which has been done in the field of churn analysis from time to time. It also provides a new methodology of churn prediction that will act as a general-purpose system for feature selection and hence can be used to evaluate churners of any firm.

Keywords: *Churn prediction, customer retention, profitable customers, automatic attribute selector, data mining, classification*

****Author for Correspondence*** E-mail: jansaima119@gmail.com

INTRODUCTION

As the markets are growing continuously so are the competitors. In this growing competition, it is easy for the company to lose its place in the market if proper care is not taken, i.e., to stay aware of the competition and competing with the competitors so that the position it has gained in the market is not lost. The company needs to remain aware not only in providing better services/products as compared to its competitors but also satisfying the existing customers so that they will not leave the company. Thus focus should be on each individual customer to satisfy him. If the company is not successful in satisfying any customer then that customer will leave the company. This loss of the customer from the company is called customer churn or customer attrition.

Churn prediction has become an important factor for firms in order to keep the population of customers in a firm up to remark. Different companies like telecommunication company, banks, ISP's, television companies, insurance companies, etc. need to remain up to date about each customer, i.e., they have to analyze each customer whether he is satisfied with the services provided by the company, their policies and the price the customer is paying. For this, the customer relationship management analysis is an important tool for the company to know the relationship of the customer with the company. If the relationship of the customer with the company is not good then the company needs to build a relationship with that customer and try to retain the customer.

There are several methods by which a firm can gain customers which come under the heading of customer acquisition like spreading awareness, advertisement, recommendations, etc. However, the cost of acquiring customers by this method is very high. Another method of maintaining the customer base is by means of customer retention which has proved cost-effective as compared to customer acquisition methods. It can be seen that applying retention actions to the already acquired customers is more beneficial for the firm if it takes loyalty into consideration. As the existing customers are already aware of the firm and are loyal as compared to new ones which are not only unaware but also the firm requires gaining their loyalty. Thus retention is better than acquisition as it is cost-effective and easy comparatively. Thus companies often try to win back defecting customers, as recovering existing customers is far better for the company than recruiting new ones.

The churn may be of the voluntary type or involuntary type depending on the type of attrition. In voluntary churn, customer attrition occurs when the customer himself makes a decision to switch to another company because of dissatisfaction with the existing company. However, in the case of involuntary churn customer attrition occurs due to natural conditions like the relocation of the customer to some other place or due to the death of the customer or some other personal reason. Thus involuntary reasons should not be included for our analysis. However, the focus should be on voluntary churners because it occurs due to the factors by which customer's loyalty with the company decreases like billing policy, the help which customer can get after purchasing or taking service of the company, service policy or better facilities which the customer can avail from some other company.

The company always looks for the efficient methods by which the customer behavior can be analyzed and his loyalty can be predicted. There are many business intelligence programs which are used for this purpose. These Software's can extract the information from the database by which the customer behavior can be analyzed and the factors of dissatisfaction like billing rivalry, technical support or dissatisfaction with the company's policies can be predicted and measured. There are many predictive analytic systems which employ churn prediction models to predict customer turn over by measuring their tendency of a threat to attire. These models help in depicting churners and hence a strategy can be used to perform retention actions on profitable churners.

Whatever the method may be utilized for increasing the customer base, the value of the firm will increase if the customers in the firm are greater in number. The firm should thus employ those methods by which customer base can be increased. This not only involves expansion methods but also monitoring continuously the existing customer base to check whether the customers are satisfied with the company or not. This monitoring of existing customers can be used to determine the loyalty of these customers. Determining loyalty of each individual customer in a firm is very important as this may help in determining the relationship of customer with the firm which in turn can be used to detect whether the customer is satisfied or not and thus helps the firm to know whether the customer will continue with the firm or not, thus helping the firm to categorize the customers as either churner, if they are likely to leave or non-churners if they will stay. Determining churners can, therefore, help the firm to apply retention actions for such customers. Retention actions are important for maintaining the customer base and in turn the value of the firm.

RELATED WORK

In businesses, the customer is the key entity which should be given primary importance for the growth and sustenance of the company in the market. For this purpose, the need is to study the behavior of customer so that the methods which can satisfy him can be revealed and employed. It is also necessary to know whether the customer will stay with the company or not. For that purpose, many methods have been employed from time to time to determine this customer attrition. Some of the work performed by different researchers in this field is summarized below:

Wei et al. [1] exercised call information of a telecommunication provider for forecasting churners by using a data mining means in which recorded data and call particulars are utilized as attributes. To eliminate the extremely unbalanced characteristics of data in determining the effect of decision trees, a multiclassifier class combiner technique was applied. The outcome verified the enhanced effect of this approach.

Burez et al. [2] worked on managing class imbalance in churn prediction which occurs due to improper evaluation metrics, relative lack of data and noise. They employed six real-life datasets associated to churn by implementing better evaluation metrics, cost-sensitive learner, two sampling techniques and boosting. The outcome manifests that AUC (Area under the curve) and lift prove to be superior estimation measure on matching it with standard approaches like logistic regression and random forests. Also, under-sampling can bring enhanced prediction efficiency when measured using AUC. Though, advanced sampling method CUBE obstructs increase in predictive performance.

Glady et al. [3] employed CLV (Customer Lifetime Value) to classify customers as either churners or not. They exercised the dataset of a Belgian financial service corporation. They used a financial metric of accuracy, i.e., loss/profit caused by a wrong classification, examined using a CLV approach. Both traditional, as well as cost-sensitive methods were employed and the outcome demonstrates that cost-sensitive tactics accomplish superior results in terms of profit measure.

Khan et al. [4] employed dataset of Internet Service Provider Company utilizing the features like employment of billing, usage, and demographic data which revealed that there is very less influence of demographic features on the prediction of churners. Also, three classifiers, i.e., decision trees, logistic regression, and neural network was compared observing a little improved performance of neural networks compared to decision trees and logistic regression.

Tsai et al. [5] employed customer relationship management dataset of telecommunication company to forecast churners by means of the hybrid network. For this reason, they compared hybrid network ANN (artificial neural network) combined with ANN with that of SOM (self-organizing map) combined with ANN and baseline ANN model. The result indicated that the hybrid model ANN + ANN achieved better results than other representations.

Tamaddoni et al. [6] used telecommunication service Provider Company's dataset to calculate churning tendency in a noncontractual corporation in which there is not any bond of the client with the corporation which makes it hard to guess the conduct of client as related to a contractual one. They created a predictive model for this determination. But prior to this model creating phase, one more phase, i.e., clustering phase was scheduled to primarily split consumer base into four clusters, which was then employed in the model construction phase. The outcome presented satisfactory performance of this predictive model. Also for dealing with imbalanced data, the performance of cost-sensitive learning method was tested and its effectiveness was confirmed.

Liu et al. [7] castoff datasets of Samsung game launcher platform to predict churners in case of mobile games. For this determination, an innovative semi-supervised embedding model was offered which mutually acquires two functions, i.e., prediction function and edge-embedding function to map handcrafted features to latent features automatically. To illustrate the factors of an edge, a novel attributed random walk technique was used. Deep neural networks were used for modeling prediction and embedding function. The effectiveness of the proposed model was confirmed by comparing the proposed model with many advanced baseline techniques.

Verbeke et al. [8] deployed datasets of wireless telecommunication operators. Several data mining algorithms were collated and it was established that the achievement of decision trees is superior. Evaluation measures like AUC and top decile lift were used alongside with novel profit-centric performance metric which revealed that statistic metrics can lead to suboptimal model selection and loss in profits. Good quality attribute selection is also essential for calculating churners with high correctness. Also, a process model involving four steps, i.e., data gathering and selecting variables from it, building a model and training it, implementing the model and reviewing the model continuously to monitor whether it still performs well is constructed to predict churners.

Wai-Ho et al. [9] suggested a novel evolutionary data mining method termed DMEL (data mining by evolutionary learning) to mine rules in databases. It works iteratively beginning with mining a set of first-order rules up to the abstraction of higher order rules. The act of DMEL is estimated by relating it to seven real-world databases and the outcome showed that DMEL delivers correct classification. Two additional classifiers, i.e., C4.5 and neural networks were also applied. However, the result shows that DMEL outperforms neural networks which in turn outperform C4.5. Also, DMEL verified to be vigorous in determining concealed rules from databases and predicting churners with different churn rates.

Hoppner et al. [10] showed that classical models of churn prediction remained unsuccessful to contemplate misclassification expenses and remunerations acquired after accurate classification. Therefore, they created a churn calculation model which is commercial and rather interpretable as well. For this, they incorporated a metric, i.e., expected maximum profit measure for consumer churn straight keen on the model production. The model is entitled ProfTree which practices an evolutionary algorithm for understanding profit focused decision trees. Datasets from various telecommunication service providers were castoff and the outcome shows that ProfTree accomplishes substantial profit progresses compared to traditional accurateness determined tree-based procedures.

Gupta et al. [11] assessed numerous CLV employed models, i.e., RFM model, probability model, econometric model, persistence model, computer science model and diffusion/growth model. In all these models CLV is evaluated and it is portrayed that CLV can be a good metric for computing the conduct of customer. It can be realized that customary metrics are not sufficient to create operational marketing. A disaggregate metric at the customer level is essential to define each consumer's nature in demand to calculate their profitability and therefore, whether to retain them in firm or not. This disaggregate metric is CLV which is used to decide forthcoming charge of each consumer and therefore, their involvement in the company.

METHODOLOGY

The overall process involved in our system is shown in Figure 1.

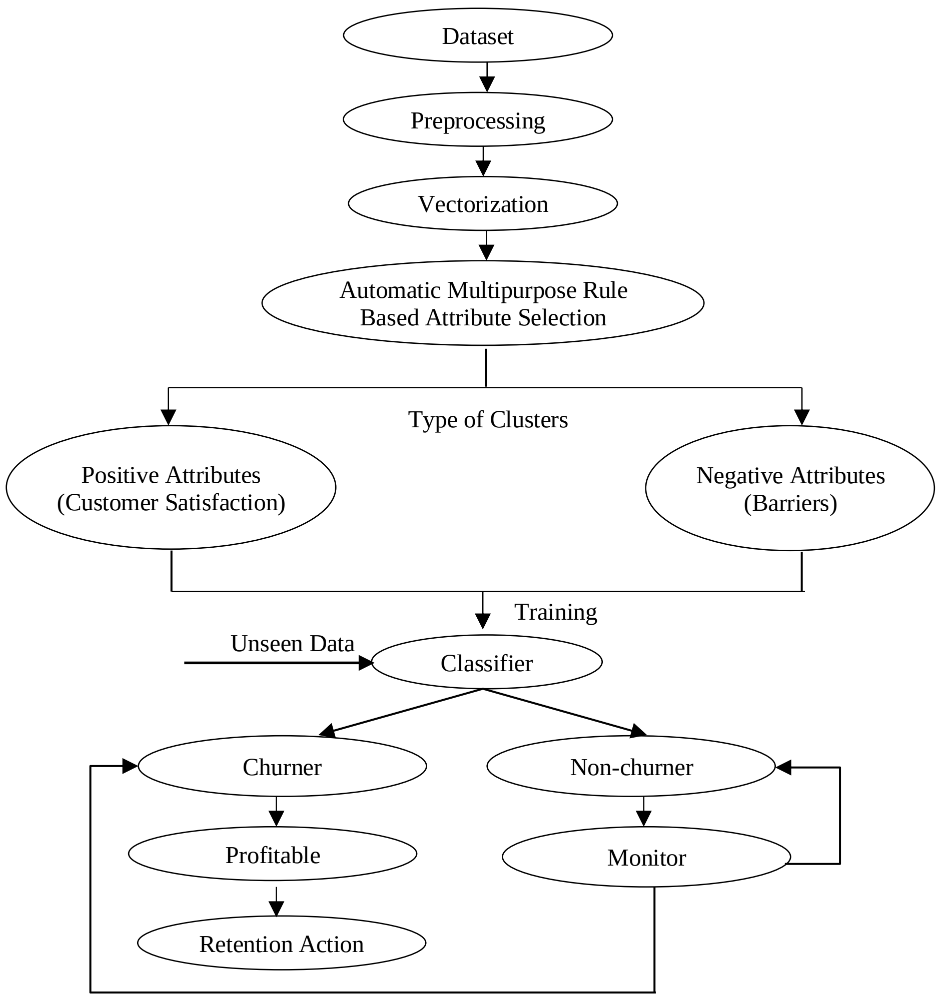

Fig. 1: *Overview of Churn Prediction.*

The first step in churn prediction is to select a dataset. For this purpose, any dataset can be used which may be of Telecommunication Company or Retail Company or Financial Company or any other company for which we want to provide the churner list so that the company can prevent them from leaving and thus maintaining its value in the market.

After the selection of dataset, the second step involves performing preprocessing of this dataset. The purpose of this step is to improve the quality of data, i.e., to convert the data to information and therefore, get it ready for the actual study. Preprocessing is thus the most important step and is divided into different substeps as shown in Figure 2.

After importing the data, the preprocessing step starts which involve data cleaning in which missing values are filled, noisy data are smoothened and outliers and inconsistencies are removed. After cleaning, integration of data is done in order to reduce redundancy and inconsistencies in data combined from different sources. It also improves the quality of data. Data integration is followed by data transformation in which steps like smoothening of data, aggregation, generalization, normalization and attribute construction are performed. After that, the final step in preprocessing involves data reduction which involves discarding faulty data and dimension reduction or feature selection.

After preprocessing, the next step is vectorization. The steps it involves are shown in Figure 3.

In this step, the categorical data are given to software which is a machine learning library called sci-kit learn, which is designed to interoperate with scientific and numerical libraries. From Scikit learn library a pre-built functionality, i.e., preprocessing is imported which can be used to convert this categorical data into a new more appropriate representation. For this purpose, a label encoder is used which converts nonnumerical labels by using a method called fit transform into numerical data.

The fourth step involves applying the vectorized data to an automatic multipurpose rule-based attribute selection classifier. The classifier used is an unsupervised learner which extracts the attributes from the data automatically. The classifier is called k-means clustering which will cluster the data into two clusters, i.e., positive cluster (containing the attributes which satisfy the customers) and negative clusters (barriers which can decrease customer loyalty).

The attributes hence obtained in the previous step are used to train the classifier. The classifier used is a supervised binary linear classifier called SVM (support vector machines) which is used to classify data as either belonging to one group or the other based on some common patterns from among one of the group. The trained classifier is then used to predict the unseen data and in the testing phase, it divides the customers on the basis of what it has learned in the training phase into two classes, i.e., churners and non-churners.

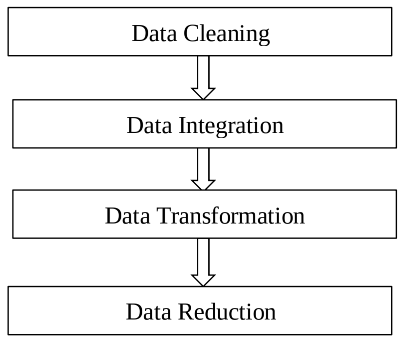

Fig. 2: *Preprocessing.*

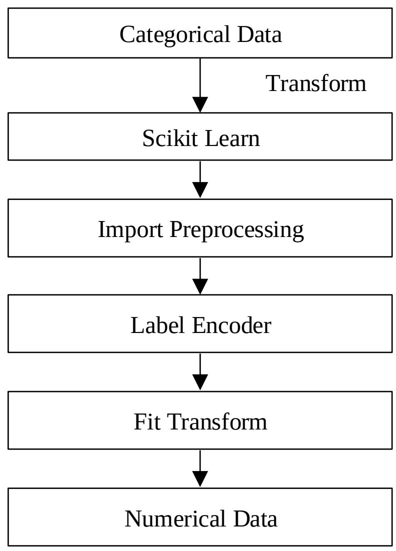

Fig. 3: *Vectorization.*

Non-churners obtained are continuously monitored whether they will remain loyal or not. If their loyalty is decreased, i.e., they are going to leave the company then they are added to the would-be churners group.

The would-be churners obtained from classification result are then checked for profitability. Since for nonprofitable customers applying retention actions is a waste of sources. Thus profitable customers are determined by setting a threshold value. If the value is below the threshold then the customers are not profitable and no retention actions will be taken on them. However, if the value is above the threshold then the customers are profitable and thus retention actions will be taken on them.

CONCLUSION

In this paper, we surveyed about how to predict churners by different methods presented by different researchers. We also presented our own methodology for determining churners. For this purpose, an automatic multipurpose rule-based attribute selection algorithm is used which will extract attributes automatically from any dataset. Thus the method we are presenting here will act as a general-purpose system for predicting churners of any firm. Also, a classifier is trained with these attributes and is then tested on unseen data to determine churners. From among these churners, profitable ones are detected for which retention actions are taken.

REFERENCES

1. Wei CP, Chiu IT. Turning telecommunications call details to churn prediction: a data mining approach. *Expert Syst Appl*. 2002 Aug 1; 23(2): 103–12p.
2. Burez J, Van den Poel D. Handling class imbalance in customer churn prediction. *Expert Syst Appl*. 2009 Apr 1; 36(3): 4626–36p.
3. Glady N, Baesens B, Croux C. Modeling churn using customer lifetime value. *Eur J Oper Res*. 2009 Aug 16; 197(1): 402–11p.
4. Khan AA, Jamwal S, Sepehri MM. Applying data mining to customer churn prediction in an internet service provider. *Int J Comput Appl*. 2010 Nov 21; 9(7): 8–14p.
5. Tsai CF, Lu YH. Customer churn prediction by hybrid neural networks. *Expert Syst Appl*. 2009 Dec 1; 36(10): 12547–53p.
6. Tamaddoni Jahromi A, Sepehri MM, Teimourpour B, Choobdar S. Modeling customer churn in a non-contractual setting: the case of telecommunications service providers. *J Strategic Market*. 2010 Dec 1; 18(7): 587–98p.
7. Liu X, Xie M, Wen X, Chen R, Ge Y, Duffield N, Wang N. A Semi-Supervised and Inductive Embedding Model for Churn Prediction of Large-Scale Mobile Games. *In2018 IEEE International Conference on Data Mining (ICDM)*, Singapore, 2018 Nov 17, 277–286p.
8. Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *Eur J Oper Res*. 2012 Apr 1; 218(1): 211–29p.
9. Au WH, Chan KC, Yao X. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE T Evolution Comput*. 2003 Dec; 7(6): 532–45p.
10. Höppner S, Stripling E, Baesens B, vanden Broucke S, Verdonck T. Profit driven decision trees for churn prediction. *Eur J Oper Res*. 2018 Nov 30. In press.
11. Gupta S, Hanssens D, Hardie B, Kahn W, Kumar V, Lin N, Ravishanker N, Sriram S. Modeling customer lifetime value. *J Service Res*. 2006 Nov; 9(2): 139–55p.

Cite this Article

Saima Jan, Afaq Alam Khan. Generalized Churn Prediction. *Journal of Artificial Intelligence Research & Advances*. 2019; 6(3): 19–24p.