

Machine Learning Models for Customer Churn Prediction in Insurance Sector

Deepthi Das¹ and Dr. Raju Ramakrishna Gondkar²

¹⁻² CMR University, Bangalore, India

Email: deepthi.16phd@cmr.edu.in

Abstract—The Indian insurance industry is found to be dynamically changing with the global market, and the increase in the competition is observed based on the entry of several players post market liberalization. Therefore, it is imperative for the insurance companies to maintain the portfolios of the existing customers along with attaining new customers. Besides, it is noticed that a useful technique in retaining the current customers would make the insurance company more profitable by reducing the overall cost incurred in the marketing and advertisements (cost of customer acquisition) to offset the recruitment of new customer [1]. A recent study, indicate that the customer retention is a significant aspect of improving operational efficiencies and the diverse service sectors are required to focus better on retaining its existing customers. In the last decade, the customer relationship management is seen to have growing attention in terms with the prediction of customer churn prediction. In recent studies, several data mining techniques are employed to predict the customer churn by considering the characteristics of the datasets, validation, and evaluation of the behaviour of the customers. With the emergence of newer technologies in data mining like Big Data, Machine Learning, Neural Networks etc., such models are gaining dimensions which couldn't have been explored earlier. In this paper, a study on different data mining techniques and machine learning models of customer attrition prediction in the insurance sector are described. Generally, used prediction models are logistic regression, decision trees, artificial neural networks, support vector machines and random forests. Various authors/researchers have used these models in their study to predict the customer churn in service sectors. This compares the various prediction models used to predict the customer churn in the insurance and telecom sector.

Index Terms— Customer churn prediction, Data Mining, Machine Learning, Customer Relationship Management (CRM), Support Vector Machines (SVM), Artificial Neural Network (ANN), Kernel Extreme Learning Machine (KELM).

I. INTRODUCTION

The key source of income obtained by business and corporative enterprises are from their customers. Main focus for the enterprises is to retain the existing customers rather than obtaining new customers. It is much easier to explain the product features and offers to the existing customers and also this helps in building a strong relationship between the company and the existing customers. In the paper by Deepthi Das and Raju Ramakrishna Gondkar (2018) it is noticed that the customer loyalty is found to be lagging and customer churn is rapidly increasing because of the competitive world, price change and the increasing benefits from the competent company [2]. Presently, the insurance industry is observed to have rapid changes in the evolving market due to

the internationalization of activities, risks, and development of innovative approaches based on the requirements of customer service. However, the increase in the competition is found to develop numerous problems for the traditional insurance sectors. Besides, it is ascertained by KasemsapK (2018) that the growth of the insurance companies mainly depends on the services that are being provided to the customers as per their needs and requirements of the customer [3]. The customer retention in a service sector is found to be interrelated with diverse factors such as a set of various dimensions of parameters that influence the loyalty and satisfaction of customers based on the services that are provided. Thus, it is noticed by Nema J & Jatav S (2017) that in the present scenario there are several determinants such as innovation in products, modest pricing, prompt services, responsiveness, and reliability of services that affects the customer retention [4]. The customers are found to have numerous choices in terms of selecting the insurance policies from a wide range of life insurance providers among those contributing attractive benefits. Thus, the significant issues observed within the insurance sectors are found to be customer attrition. Consequently, discontinuing of payments by the policyholders resulted in the termination of insurance policies creating a lapse in the investments leading to the termination of the policy (Yang, C. G., Trimi, S., Lee, S. G., & Yang, J. S., 2017) [5]. In the motor insurance sector, recruitment of new customers is found to be extremely difficult leading to a higher expense in comparisons to retaining the customers. Besides, similar criteria are observed in life insurance services as the majority of the customers having bought the policy over more than 10 years are very high. Thus, termination of the premium over a period is a reason to cease the policies. Therefore, losing such customers is observed to affect the overall revenue of the insurance sectors (Goonetilleke, T. O., & Caldera, H. A., 2013) [6]. In addition, the possibility of detecting such terminations are found to be a rigorous process unless the customer contacts the company directly regarding their concerns for ceasing the policy. This indicates the necessity among the insurance sectors to perform exhaustive analysis such that processes could be defined to retain the existing customers (Ganeshan, H., & Nagarajan, S., 2017) [7].

In the last decade, the customer relationship management is seen to have growing attention in terms with the prediction of customer churn prediction. Recent studies, several data mining techniques are employed to predict the customer churn by considering the characteristics of the datasets, validation, and evaluation of the behavior of the customers (Bharti, A., 2017) [8]. Generally, machine learning algorithms such as logistic regression, decision trees, neural networks, support vector machines, random forests, and regression forests are applied for the prediction of the customer attrition. Besides, the lucidity of models used for the prediction of customer attrition is not focused in the existing literature (Xia, G. E., & Jin, W. D., 2008) [5].

The analysis of the behaviour of the customers illustrates and provides an insight towards the reasons for the cause of retention. Thus, it is necessary to develop modelling techniques that could be used for predicting the churn among the existing customers. Therefore, the process of churn management is a pre-requisite for the growth of the company and helps to access the current situation of the company and setting future plans for the development of the company. The rate of customer churn in an organisation is effectively managed by employing Customer Relationship Management (CRM) persons.

Customer Relationship Management is a method to manage a company's relation with current and potential customers. It analyses the customer data available and uses this data to improve the relationship with customers. The early recognition of customer churn enables the organization to concentrate more on such customers and develop certain specific retention actions to know the reason and retain the profitable and loyal customers. Fig. 1, shows the different components of CRM. CRM people tries to recognize the customers who are probable to churn before they leave and avoid so by providing extra benefits and incentives to customers. However, the manual prediction is found to be inaccurate and the organization is losing its money by providing benefits to the customers who is not going to churn [9].

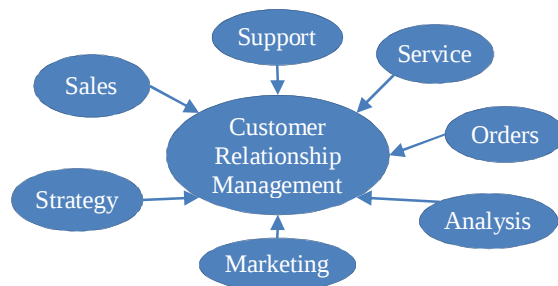


Figure 1. Components of CRM

Several models comprising many data mining methods and machine learning algorithms are being developed to analyse the behaviour of customer churn in telecom, industrial, and insurance sectors. These predictive model plays a vital role during the design of efficient customer relationship management programs and still researches are going to explore more accurate ways of determining the churn rate in various real time applications such as telecommunication, insurance, banking and industries. In this paper, an overview of various churn prediction models along with machine learning techniques deployed by several authors for early detection of churn rate is presented. This paper compares the different machine learning algorithms namely Support vector machine, Logistic regression, neural networks and Kernel Extreme Machine algorithm which are used for customer churn prediction in various sectors [2].

A. Data Mining

Data mining methods help to analyse large amounts of structured or unstructured data and summarise them into meaningful information and/or to discover patterns. Data mining techniques uses a combination of machine learning, statistics, and database systems methods to process such data sets. By using data mining techniques, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. The different Data mining Process is shown in Fig 2. The next section gives a brief description of few machine learning algorithms which are considered for this study.

B. Logistic Regression

Logistic Regression is a method for binary classification. It gives a discrete binary outcome between 0 and 1 (Logistic regression, n.d.). The usage of logistic regression has become common to find out the probability of customer churn and the probability is calculated as a function consisting of a set of customer attributes. Logistic Regression considers the dependent variable and one or more independent variable and finds the relationship between them. The relationship is calculated by estimating probabilities using the underlying logistic function. Then these probabilities are converted into binary values for making predictions.

C. Decision Tree

Decision Tree (DT) is a tree-like model which represents a set of decisions (DT, n.d.). It is a supervised learning algorithm that is commonly used in classification problems. The methodology is to split the sample data set into two or more subsets which are homogeneous based on the most significant variables. It is composed of root node and sub-nodes which is terminated by leaf nodes. The leaf nodes represent the class labels. It can handle both categorical and numerical values. It is easy to understand and very useful in data analysis. Hence this is a very popular model used for customer churn predictions.

D. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be applied for both classification and regression problems (SVM n.d.). A hyper plane is used to classify the items which are represented as a point in n-dimensional space. This algorithm is very effective in high dimensional space and during implementation this method uses memory efficiently as it uses the subset of training data known as the support vectors. It is found to be giving more accurate in customer churn prediction.

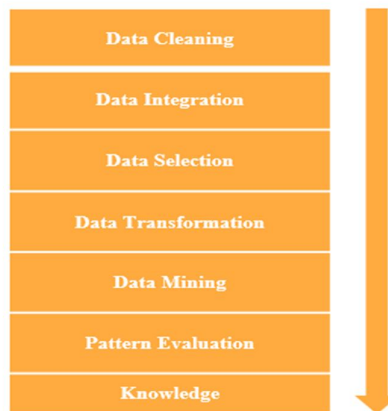


Figure 2. Data Mining Process

E. Artificial Neural Networks

Artificial Neural Network algorithm is based on the model of the neurons present in human brain (ANN, n.d.). This algorithm can be applied for classification as well as regression problems. Three layers namely Input layer, Hidden layer and Output layer are there in this model. The weighted links are used as the connection between the layers. ANN algorithms have shown better accuracy rate in customer churn prediction in various business sectors.

F. Kernel Extreme Learning Machine

Extreme learning machines are feedforward neural networks which can be used for feature learning with either single layer or multiple layers of hidden nodes. The single hidden layer feed –forward neural networks determines the initial parameters of input weights and hidden biases randomly with simple kernel function. In case if the function for feature mapping is unknown then kernel function can be used to improve the stability of the algorithm. This is known as kernel based extreme learning machine. But care should be taken to choose the kernel parameter for improving the performance of kernel learning machine algorithm.

II. SURVEY OF VARIOUS METHODS

The proposed model by XIA Guoe-EN, JIN (2008) [5] uses SVM with “structural risk minimization” to predict customer churn in telecommunication. The method assumes that if the customer data can be accurately separated by the hyperplane (closest to the data vector plane) then classification decision function can be derived. The authors tested the model on two datasets Dataset1- Machine learning UCI database of University of California and Dataset2- Home telecommunication carry. It is found that SVM has a good precision with an accuracy rate of 0.9088. Limitations of the proposed model is that the selection of kernel functions is depended on transcendent information and empiricism.

The authors Helen Treasa and Rupali Wagh (2017) [10] in their study have used Logistic regression to predict customer churn in telecommunication sector. The dataset used is the Telecom Dataset. The authors used the methodology where all variables are considered and backward stepwise regression is applied on the set which results in deletion of each variable with a certain criteria that is applied to it. The process is continued till there are no further variables that are available to be deleted without any statistical loss of fit. The proposed model has an Accuracy rate of 80.02 %.

The authors Anuj Sharma and Prabin Kumar (2011) [11] in their work propose an artificial neural network (ANN) based approach to predict customer churn in subscription of cellular wireless services. The churn data set from the UCI Repository of Machine Learning Databases at the University of California, Irvine is used for the study. The authors experimented with many hidden layers in the neural network. The model consists of three to seven neurons. It is found that the finest results are attained with one hidden layer consisting of three neurons. The output layer contains two neurons which represented the two output values (churn being true or false). Accuracy rate of the proposed model is found to be 92.35%

The authors S.Induja and Eswaramoorthy (2016) [12] in their paper have proposed a prediction model using KELM and Bat algorithm to predict the customer churn in telecom industry. For this work the authors have used data set from IBM Watson analytics Telco customer churn data. This data set consists details about mobile customers. The customer behaviour is analysed and categorised into four different types using Naïve Bayes classifier. The attribute selection is done using BAT algorithm and the churn prediction is done using kernel extreme learning machine algorithm. The proposed model in this work shows better accuracy than AdaBoost and hybrid support vector machines. The proposed model has an accuracy rate of 98.3%.

III. CONCLUSION

The primary aim of this research findings is to detect an efficient machine learning technique to predict the customer attrition in motor insurance industry. This paper focuses on different authors review on input data selection, pre-processing steps, selection of algorithms and their performance evaluation in terms of accuracy. The different methodologies done by various authors are consolidated in table 1. The literature review implies that many machine learning techniques are developed in terms of churn prediction for telecommunication sectors. Through the results obtained from the review, it is observed that the KELM and BAT algorithm gives a higher accuracy in customer churn prediction and can be enhanced to apply in prediction of customer churn in any other industry.

TABLE I. ACCURACY RATES OF MACHINE LEARNING MODELS FOR CUSTOMER ATTRITION PREDICTION

Author	Algorithm	Accuracy Rate	Dataset
Xia, G. E., & Jin, W. D	Support Vector Machine	90.88%	Machine learning UCI database of University of California
Helen Treasa & RupaliWagh	Logistic Regression	80.02%	Telecom Data Set
S.Induja & Eswaramoorthy	Kernel Extreme Machine Learning	98.3%	IBM Watson analytics Telco customer churn data set
Anuj Sharma, Dr. Prabin Kumar Panigrahi	Artificial Neural Network	92.35%	Machine learning UCI database of University of California.

REFERENCES

- [1] A. Amin, S. Shehzad, C. Khan, I. Ali & S. Anwar, "Churn prediction in telecommunication industry using rough set approach," *In New Trends in Computational Collective Intelligence*, Springer, pp. 83-95, 2015.
- [2] Deepthi Das, Raju Ramakrishna Gondkar, "Performance analysis of machine learning algorithms in customer churn prediction," *International Journal of Latest Trends in Engineering and Technology*, 29-34, 2018.
- [3] K. Kasemsap, *Customer Lifetime Value*, In Encyclopedia of Information Science and Technology, Fourth Edition, IGI Global, 1584-1593, 2018.
- [4] J. Nema and S. Jatav, "Determinants of Customer Retention in Health Insurance Sector," *IJRIS*, I(II), ISSN 2454-618, 2017.
- [5] G.E. Xia, and W.D. Jin, "Model of customer churn prediction on support vector machine," *Systems Engineering-Theory & Practice*, 28(1), 71-77, 2008.
- [6] Goonetilleke, T. O., & Caldera, H. A, "Mining life insurance data for customer attrition analysis," *Journal of Industrial and Intelligent Information*, 1(1), 2013.
- [7] H. Ganeshan and S. Nagarajan, "A study on customer perception on service quality with reference to public insurance sector in coimbatore district," *International Journal of Management Research & Review*, 7(4), pp. 513-520, 2017.
- [8] A. Bharti, "Customer churn management," *ACADEMICIA: An International Multidisciplinary Research Journal*, 7(5), 96-102, 2017.
- [9] E. Ascarza, S.A. Neslin, O. Netzer, Z. Anderson, P.S. Fader et. al., "In pursuit of enhanced customer retention management: Review, key issues, and future directions," *Customer Needs and Solutions*, 5(1-2), 65-81, 2018.
- [10] Helen Treasa, Rupali Wagh, "Churn Analysis in Telecommunication using Logistic Regression," *Oriental journal of computer science & technology*, 10 (1), 207-212, 2017.
- [11] Anuj Sharma, Dr. Prabin Kumar Panigrahi, "A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services," *International Journal of Computer Applications* (0975 – 8887) Volume 27– No.11, August 2011.
- [12] Induja, P Eswaramurthy, "Customers Churn Prediction and Attribute Selection in Telecom Industry Using Kernelized Extreme Learning Machine and Bat Algorithms," *International Journal of Science and Research (IJSR)*, Volume 5 Issue 12, December 2016.
- [13] C.G. Yang, S. Trimi, S.G. Lee, and J.S. Yang, "A Survival Analysis of Business Insolvency in ICT and Automobile Industries," *International Journal of Information Technology & Decision Making*, 16(06), 1523-1548. 2017.
- [14] C. Hung, and C.F. Tsai, "Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand," *Expert systems with applications*, 34(1), 780-787, 2008.
- [15] J. Burez and D. Van den Poel, "CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services," *Expert Systems with Applications*, 32(2), 277-288, 2007.
- [16] N. Delafrooz, and E. Farzanfar, "Determining the customer lifetime value based on the benefit clustering in the insurance industry," *Indian Journal of science and Technology*, 9(1). 2016.
- [17] ANN (n.d.) Retrieved from <https://www.datasciencecentral.com/profiles/blogs/artificial-neural-network-ann-in-machine-learning>
- [18] Decision Tree (n.d.) Retrieved <https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/#one>
- [19] Logistic Regression (n.d) Retrieved from <https://towardsdatascience.com/machine-learning-part-3-logistics-regression>
- [20] Random Forest(n.d.) Retrieved from <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>
- [21] SVM (n.d.) Retrieved from <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>
- [22] KELM <https://towardsdatascience.com/evolution-of-extreme-learning-machines-2c7caf08e76b>
- [23] KELM https://en.wikipedia.org/wiki/Extreme_learning_machine