DEVELOPMENT OF A PREDICTION MODEL IN IDENTIFYING THE CHURNERS AND NON-CHURNERS

Prasanth S.[1*] and Rathnayake R.M.K.T.[2]
[1]Department of Computing & Information Systems, Applied Science, Sabaragamuwa University of Sri Lanka, Sri Lanka
[2]Department of Physical Science and Technology, Applied Science, Sabaragamuwa University of Sri Lanka, Sri Lanka
*senthanprasanth007@gmail.com

Customers play a vital role in the overall function of the telecommunication industry. So, it is important for any industry to prevent the tendency of churning of the customers from an organization. In order to do so, an effective churn prediction model need to be developed in advance. In this view, the sole objective of this research is to develop an appropriate application with the intention of finding the churners and non-churners from any given number of customer data. During this research process, 10,000 post-paid subscriber details with 20 attributes were obtained from a local telecommunication company and a thorough analysis was executed. Among the above number, it was observed that 4888 were churners and the rest 5112 were non-churners. To find the best algorithm for the development of the final prediction model, several supervised machine learning techniques were incorporated and a proper comparison was done against certain evaluation metrics such as Accuracy, Mean Squared Error (MSE), Precision , Recall and so on. In fact, the following supervised machine learning techniques namely Random Forest, XGBoost, AdaBoost, Logistic Regression, Neural Network, Support Vector Machine, and Decision Tree were experimented with the given data set. As an initial step, the given data were pre-processed and feature engineering was performed with the help of correlation analysis. From the results obtained, 17 attributes out of 20 were identified as the most important aspects to cover the entire data. Consequently, the whole data were fed into aforementioned techniques for the purpose of finding the best one. In this process, more preferable results were obtained from the ensemble approaches such as Random Forest, XGBoost and AdaBoost. Eventually, it was found that XGBoost had the highest accuracy of 82.90% and lowest error rate was 17.1%. In addition to this, five (5) fold cross validation too had been performed for the purpose of ensuring the highest accurate results by the XGBoost incorporated with different percentages of training and testing data. Further, with the intention of getting an increase from the accuracy already obtained, hyper-parameter tuning was done with XGBoost and thus this attempt resulted an accuracy of 83.13%.

Keywords: *Churn, Machine learning, Prediction, XGBoost, Adaboost*