NEW CUSTOMER CHURN PREDICTION MODEL FOR MOBILE TELECOMMUNICATION INDUSTRY

LLG Chathuranga[1], RMKT Rathnayaka[2], and HI Arumawadu[3]

[1]Department of Computing and Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka

[2]Department of Physical Sciences and Technology, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka

[3]Dialog Axiata PLC, Colombo, Sri Lanka

[1]*chathurangagihan39@gmail.com*

Abstract- The present Sri Lankan telecommunication industry remains extremely dynamic by constantly changing the landscape of new services, technologies, and carriers. Thus customers have more choices. So, predicting customer churn is one of the most challenging targets in the telecommunication industry today. The major aim of the study is to develop a novel customer churn prediction model for Sri Lankan Telecommunication Company by considering some soft factors for early identification of customers who leave the service provider. Three machine learning algorithms namely Logistic Regression, Naive Bayes and Decision Tree are used in this study. In fact, twenty attributes are mainly carried out to train these three algorithms. Furthermore, the Back Propagation Neural Network (BPNN) was trained to predict customer churn. In Artificial Neural Network (ANN) training; result of Logistic Regression, Naive Bayes and Decision Tree and eight attributes that mostly affecting the final result are used as inputs. The performances of the models are evaluated by using the confusion matrix using three different data samples. Final ANN model gives 96.7% accuracy in the testing process. Also it gives a high accuracy when comparing with the other data mining algorithms. Existing customer churn prediction models are designed using single algorithm. But the experimental results in this study show multiple algorithms for churn prediction that give higher performance than a single algorithm.

Keywords- machine learning; neural network; algorithm

I. INTRODUCTION

In the past few years, the telecommunication industry in Sri Lanka has shown significant growth compared with other industries. The focus of telecommunication companies has therefore shifted from building a large customer base into keeping customers in house.

Therefore, it is valuable to understand which customers may turn to competitors in the near future. Customer churn prediction was needed to identify these changes of customers. Customers are the major fractions to be focused in every industry as products and services are rendered to them. In fact, efficient business practitioners should be able to cater to the demands of business clients. For that customer churn prediction is very important in any industry. This is same to mobile telecommunication industry too. Because attracting new customers is costlier than protecting their existing customers. For that reason, companies always try to protect their existing customers. To protect their customers, they want to identify customers who have most risk to leave the service.

In this study, customers are post-paid users in Sri Lankan telecommunication Service Company. Post-paid telecom service providers can easily understand churn rate because they work with contracts. When a customer terminates their contract, the company knows they lost a customer. Customer churn occurs when customers or subscribers stop doing business with a company or service provider. It means that they leave the existing company. Customer churning is directly related to customer satisfaction.

There is an intense competition in telecommunication market resulting to introduce more sophisticated products and services. The income can be considerably affected if the company loses its faithful customers. This leads to reduced customer loyalty. Losing an existing high-volume customer means losing lots of revenue. Analysing customer data and customer behaviour is the basis for understanding the needs of any customers. It is necessary to identify customers who are willing to move to a competitor before they do so. The present Sri Lankan mobile industry is extremely dynamic, with new services, technologies, and carriers constantly altering the landscape. Then customers have more choices.

Mobile telecommunication industry generates a huge amount of data like billing information, call detail data and network data. This voluminous amount of data ensures the necessity for the application of data mining techniques in telecommunication database. In the information generated in the telecommunications industry there are hidden data and patterns not yet identified. Analysing this huge amount of data in various perspectives allows service providers to improve their business in various ways.

The major aim of the study is to develop a customer churn prediction model by considering some soft factors like monthly bill, billing complaints, promotions, hotline call time, arcade visit time, negative ratings sent, positive ratings sent, complaint resolve duration, total complaints, and coverage related complaints.

ANN based approaches are mostly used in churn prediction in many subscribers based industries. This study uses 3 machine learning algorithms namely Logistic Regression, Naïve Bayes and Decision Tree with BPNN.

Rest of these is structured as follows. Section 2 discusses the methodology part and describes the technologies used to develop this model. Section 3 describes experimental design. Section 4 describes the result and discussion of the study. This section shows the final result of the model, the prediction accuracy and other performance measurements. Finally, in section 5, the conclusions are given.

II. METHODOLOGY

Before use the dataset for analysing it is necessary to remove incomplete, noisy or inconsistent values in the dataset. In this study WEKA data mining tool was used for data pre-processing.

There are three machine learning algorithms namely Logistic Regression, Naive Bayes and Decision Tree and a BPNN were used in this model. Logistic Regression is a Machine Learning classification algorithm. Logistic Regression analyses studies the association between a categorical dependent variable and a set of independent (explanatory) variables (Arumawadu et al., 2015; Arumawadu et al., 2016). Naive Bayes is one of the speediest statistical classifier algorithm works on probability of all attribute contained in data sample individually and then classifies them accurately. It is used to predict class membership probabilities (Rathnayaka et al., 2012). Decision Tree is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees. Multilayer Perceptron is a nonlinear classifier based on the Perceptron. A Multilayer Perceptron (MLP) is a back propagation neural network with one or more layers between input and output layer.

Logistic Regression, Naive Bayes and Decision Tree algorithms are trained by applying 20 attributes until the output result of each algorithm is matched with the original resultant with minimum error rate for 3 sample datasets and then the training process is stopped. Once the algorithms are trained using these 3 data samples, algorithms with best performance showed dataset was saved.

An Artificial Neural Network with 11 inputs which are 8 input attributes that affect the customer churn mostly and three special inputs, which are prediction result from Logistic Regression, prediction result from Naive Bayes and prediction result from Decision Tree are used in the study. This neural network was trained and best performance model was selected as the final model. Training process was done by adjusting number of neurons in hidden layer1, number of neurons in hidden layer2 and value for the epoch.

Then test the model using testing datasets and the model which gives high performance was saved for use in future predictions. Performance measuring of the models is done

by using some of the confusion matrix-based measures like accuracy, precision, recall or sensitivity and F1 score. As well as used receiver operating characteristic (ROC) analysis.

Data set used in this study contains 3,334 subscribers, including 1,289 churners and 2,045 non-churners. Three data samples were used to train and test the model. Such as a data sample with 60% for training and 40% for testing, a data sample with 70% for training and 30% for testing and a data sample with 80% for training and 20% for testing.

III. EXPERIMENTAL DESIGN

When using neural networks to perform predictive modelling, the input layer contains all of the input fields or variables used to predict the outcome variable. The output layer contains an output field which is the target of the prediction. The input and output fields can be numeric or symbolic.

ANN was trained by changing number of hidden neurons and epoch until achieving better prediction accuracy. In this study first hidden layer neurons changed from one to fifteen. Second layer neurons changed from one to ten. Epoch was changed from 10 to 40. All these tests were done for all datasets such as a data sample with 60% for training and 40% for testing, a data sample with 70% for training and 30% for testing and a data sample with 80% for training and 20% for testing.

IV. RESULTS AND DISCUSSION

First the logistic regression algorithm was trained and tested using 3 data samples. After training and testing the logistic regression algorithm using these 3 data sets, the model that showed the best accuracy was selected to apply in the final prediction model. Table 1 shows the training and testing accuracy of the logistic regression algorithm.

Table 1. Training and testing accuracy score of the logistic regression algorithm.

	Training	Testing
Accuracy (60/40)	0.8689	0.8564
Accuracy (70/30)	0.8767	0.8622
Accuracy (80/20)	0.8794	0.8525

Then Naïve Bayes algorithm was trained and tested using 3 data samples and the model that showed was saved to use final prediction model.

Table 2. Training and testing accuracy score of the Naïve Bayes algorithm.

	Training	Testing
Accuracy (60/40)	0.9218	0.9058
Accuracy (70/30)	0.8565	0.9333
Accuracy (80/20)	0.9187	0.9203

Next the Decision Tree algorithm was trained and tested using 3 data samples. After training and testing the Decision Tree algorithm using these 3 data sets best accuracy shown model was selected to apply in final prediction model.

Table 3. Training and testing accuracy score of the Decision Tree algorithm.

	Training	Testing
Accuracy (60/40)	0.9218	0.9058
Accuracy (70/30)	0.8565	0.9333
Accuracy (80/20)	0.9187	0.9203

A modified ANN was built to perform customer churn prediction. Inputs for final model are 8 attributes most affect for churn and result of Logistic Regression, Naïve Bayes and Decision Tree algorithm. Trained ANN with 11 inputs in input layer, 2 hidden layers and one output layer is used in this model.

Result of the study is a modified Artificial Neural Network with one input layer, two hidden layers and one output layer gives 96.7% accuracy score for mobile telecommunication customer churn prediction. Performance of the model was shown in table 4.

Table 4. Performance of the novel customer churn prediction model

	For Training data set	For Testing data set
Accuracy Score	0.9863	0.9677
Precision Score	0.9804	0.9788
Recall Score	0.9946	0.9820
f1 Score	0.9875	0.9804

The structure of the constructed ANN is illustrated in Figure 1. It consists 4 layers which are input layer with 11 input neurons, hidden layer 1 with 15 input neurons, hidden layer 2 with 10 input neurons and output layer with neuron. A training epoch is set to 40.

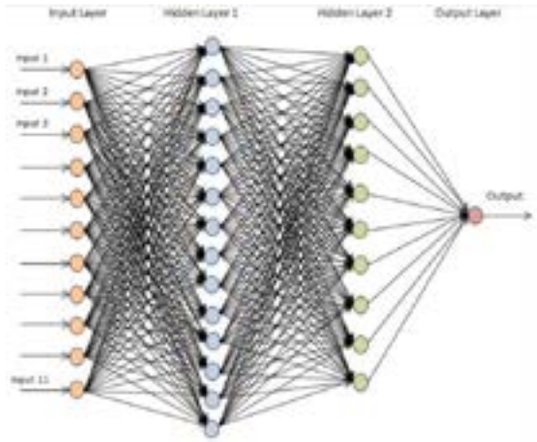

Figure 1. Structure of the constructed ANN model

V. CONCLUSION

A first logistic regression algorithm was trained using all 20 attributes. That trained model shows good performances. Accuracy score for this model is 84.7%. Then naive bayes algorithm was trained using same attributes. This model shows 91% accuracy in testing data set. Finally, decision tree algorithm was trained using same attributes. It shows 81.74% accuracy for testing data set.

Then novel model was trained using result of the above three algorithms and 8 attributes most affect for final result. The findings of this study confirm that churn can be predicted successfully with certain level of accuracy using this novel model. Developed model achieve above 95% overall accuracy on a testing set.

ANN with other three machine learning algorithms creates a strong customer churn prediction model. Performance of novel model is higher than using them separately.

REFERENCES

Almana A, Aksoy M, Alzahrani R, (2014) A survey on data mining techniques in customer churn analysis for telecom industry, *International Journal of Engineering Research and Applications*, 45, 165-171.

Eriksson K, Vaghult A, (2000) Effects on New Customer Acquisition by Retained Customers in Professional Services, In *16th IMP Conference*, Bath.

Arumawadu HI, Rathnayaka RMKT, Seneviratna K, (2016) New Proposed Mobile Telecommunication Customer Call Center Roster Scheduling Under the Graph Coloring Approach, *International Journal of Computer Applications Technology and Research*, 5(4), 234–237.

Arumawadu HI, Rathnayaka RMKT, Illangarathne SK, (2015) K-Means Clustering For Segment Web Search Results, *International Journal of Engineering Works*, 2(8), 79-83.

Arumawadu HI, Rathnayaka RMKT, Illangarathne SK, (2015) Mining Profitability of Telecommunication Customers Using K-Means Clustering, *Journal of Data Analysis and Information Processing*, 3(3), 63.

Rathnayaka RMKT, Zhong-jun W, (2012) Enhanced Greedy Optimization Algorithm with Data Warehousing for Automated Nurse Scheduling System, *E-Health Telecommunication Systems and Networks*, 1(4).

Hung S, Yen D, Wang H, (2006) Applying data mining to telecom churn management, *Expert Systems with Applications*, 31(3), 515-524.

Kirui C, Hong L, Cheruiyot W, Kirui H, (2013) Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining, *IJCSI International Journal of Computer Science Issues*, 10(2), 1694-0814.

Markopoulos A, Manolakos D, Vaxevanidis N, (2008) Artificial neural network models for the prediction of surface roughness in electrical discharge machining. *Journal of Intelligent Manufacturing*, 19(3), 283-292.

Sweeney J, Swait J, (2008) The effects of brand credibility on customer loyalty, *Journal of retailing and consumer services*, 15(3), 179-193.