IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

A Comparative study on Logit leaf model (LLM) and Support leaf model (SLM) for predicting the customer churn

Kunchaparthi Jyothsna Latha[1*], Markapudi Baburao[2], Chaduvula Kavitha[3]

[1]Student, M.Tech, CSE Dept, Gudlavalleru Engineering College, Gudlavalleru, India
[2]Professor & Mentor (AS&A), CSE Dept, Gudlavalleru Engineering College, Gudlavalleru, India
[3]Professor & HOD, IT Dept, Gudlavalleru Engineering College, Gudlavalleru, India

Corresponding Author: jyothsnalatha9@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i5.16281632 | Available online at: www.ijcseonline.org

Accepted: 23/May/2019, Published: 31/May/2019

***Abstract*—** Decision trees, logistic regression and support vector machine are very popular algorithms for predicting the customer churn with comprehensibility and well-built predictive performance and. Regardless of the strengths they are having flaws, decision trees having problem to handle the linear relations among the variables, logistic regression is having difficulties to handle interaction effects among the variables, and support vector machine performs marginally better than logistic regression. Consequently a new hybrid algorithm named as support leaf model (SLM) was proposed to classify the data. The idea following the support leaf (SLM) is that implementation of different models on segments of the data gives better predictive performance rather than on the entire dataset, the comprehensibility is maintained from the models which are constructed on the leaves. The SLM consists of two phases, one is segmentation phase and another one is prediction phase. In first stage by using decision tree the customer segments are identified and in the second stage a model is created for every leaf of the tree. To measure the predictive performance area under the receiver operating characteristics curve (AUC) and top decile lift (TDL) are used. Based on the performance metrics AUC and TDL, logit leaf model (LLM) works well when compared with support leaf model (SLM).

***Keywords*—** Customer churn prediction, Hybrid algorithm, Logit leaf model, Support leaf Model, Predictive analytics.

I. INTRODUCTION

In a time of increasingly saturated markets have an increased competition among different companies; loss of customers is the real problem [1], [2]. Therefore the companies have to get a clear idea about the past information of each and every customer. The models are created based on existing customer data, these are the important assets to predict the customer churn [3]. Identification of customers who shows high preference to move or leave the company or predicting the customer churn plays a major role [3], [4].

From previous research customer churn can be tackled in two different angles [5]. Firstly, researchers mainly focused on improving the models for predicting the customer churn, in which many complex models have been developed and proposed to boost the predictive performance [6]. Secondly, researchers want to recognize the important factors of customer churn [7].

In customer churn prediction the popular techniques are decision trees (DT), logistic regression (LR) and support vector machine (SVM) towards assess the churn probability as they combine good predictive performance with the good comprehensibility [8]. These techniques are having both strengths and flaws as well. DT can handle interaction effects between the variables very well but having problem to handle the linear relations between the variables. Logistic regression [10] can handle the linear relations between the variables, but it cannot accommodate and detect interaction effects between the variables.

In this paper, the support leaf model (SLM) was proposed as new hybrid classification algorithm which combines two individual models they are decision trees and support vector machine. Theoretically the decision tree is used to split the data into homogeneous subsets in SLM on which support vector machine is fit to every subset.

This paper was organized as follows. The next section was discussing about Existing methods. The 3rd section describes about the proposed system. The 4th Section represents experimental setup. The 5th section describes about the

results. The 6th section describes about conclusion and future work.

II. EXISTING SYSTEM

a. Decision tree (DT):

The fundamental thought of the decision tree in SLM is to divide the data recursively into subsets with the end goal that every subset contains approximately homogenous states of target variable. Decision Tree (CART) [11] was developed here by using gini as the main metric used to decide the root node attribute in the decision tree.

b. Logistic Regression (LR):

Logistic regression [10] is one of the powerful statistical methods used to analyze the datasets which have at least one independent variable for determining the outcome. By using logistic function, we can measure the relationship between one or more independent variables and a categorical dependent variable through estimating their probabilities. Outcome is two-valued which is in categorical nature. Furthermore it is used for predicting probability of non occurrence or occurrence of an event.

c. Support vector machine (SVM):

Support Vector Machines (SVM) was one of the supervised learning algorithms. It was used for both regression and classification problem [9]. It is used for classifying both linear and nonlinear data.

d. Logit leaf model (LLM):

The Logit leaf model is a two-step hybrid approach, first step is used for identifying the homogeneous customer segments by constructing the decision tree and in the second step logistic regression were applied to each and every identified homogeneous segment. Figure1 shows the conceptual representation of LLM [14]. In this illustration, the complete customer set S was divided into S1, S2 and S3 as three different subsets from the decision tree. On the identified subsets logistic regression is applied separately, resulting their probabilities for each and every instance in every subset.

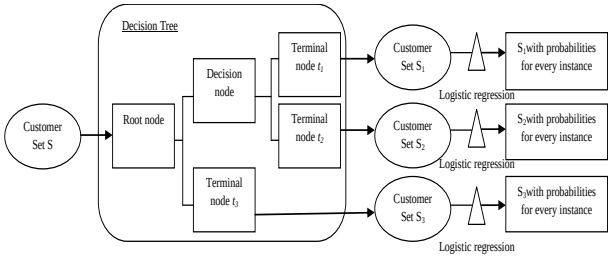

Figure1: Conceptual presentation of the logit leaf model.

Figure1: Conceptual

III. PROPOSED METHOD

By the inspiration of LLM, a new hybrid model support leaf model (SLM) is proposed, having two phases one is segmentation phase and another one is prediction phase.

Support leaf model (SLM):

It is a two-step hybrid approach which combines decision tree (DT) and support vector machine (SVM) [9], [15]. In the first step homogeneous segments are identified by constructing decision tree. In second step support vector machine (SVM) is applied to the identified homogeneous segments. Conceptual view of support leaf model is shown in Figure2.

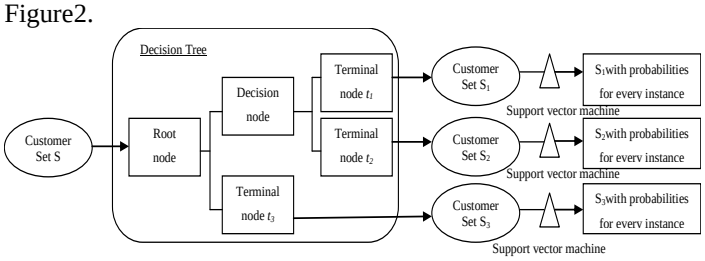

Figure2: Conceptual presentation of the support leaf model.

Algorithm:

Step 1: Selects the variables/features by using fisher score which is discussed in 4th section.
Step 2: Construct decision tree with selected variables.
Step 3: Extract the records from the terminal nodes of the decision tree.
Step 4: Apply support vector machine model on the homogeneous segments.
Step 5: Calculate the predictive performance using AUC and TDL.

IV. EXPERIMENTAL SET-UP

In the proposed system R programming is used to build the model for predicting the churn. R is freely available and a powerful statistical analysis tool.

The two customer churn datasets are used on which the SLM is compared against LLM. The information about datasets is shown in Table1.

Table3: Performance evaluation using Area under the receiver operating characteristics curve (AUC)

Area under the receiver operating characteristics curve		
Algorithm	**Datasets**	
	DS1	**DS2**
Decision Tree	0.68	0.70
Logistic Regression	0.83	0.85
Logit leaf model (LLM)	**0.79**	**0.80**
Support vector machine (SVM)	0.64	0.67
Support leaf Model (SLM)	**0.68**	**0.70**

a. Variable selection:

Variable selection is done by using fisher score [12], which is very simple and effective. It cans be defined by using below formula:

$$Fisher\ score = \frac{|\bar{X}_c - \bar{X}_{nc}|}{\sqrt{S^2{}_c + S^2{}_{nc}}}$$

Where

$\bar{X}_c$: Mean value of churners.
$\bar{X}_{nc}$: Mean value of non-churners.
$S^2{}_c$: Variance of independent variable with respect to churners.
$S^2{}_{nc}$: Variance of independent variables with respect to non-churners.

b. Evaluation criteria:

The predictive performance of different classifiers was assessed by using top decile lift (TDL) [13] and area under the receiver operating characteristics curve (AUC). From the confusion matrix AUC and TDL are derived. Table2 represents the confusion matrix for binary classification.

Table2 : Example of a confusion matrix for binary classification

		Actual		
		1	**0**	
predicted	**1**	True positive (TP)	False positive (FP)	Predicted positive (PP)
	0	False positive (FP)	True negative (TN)	Predicted negative (PN)
		Actual positives (AP)	Actual negatives (AN)	

Lift is a metric that expresses how the incidence in the 10% customers with the highest model predictions compares to the overall sample incidence. Lift reveals the specific cut off value for a classifier that predicts how much better (or worse) compared to random selection. Lift is defined by using confusion matrix, the formula is as follows:

$$Lift = \frac{TP/(TP + FP)}{AP/(AP + AN)}$$

Table1: Datasets description

	Dataset Name	**Source**	**No.of Records**	**No.of Attributes**
DS1	WA_Fn-UseC_-Telco-Customer-Churn	https://www.kaggle.com/blstchar/telco-customer-churn	7044	21
DS2	Cell2cell	https://www.kaggle.com/jpacse/datasets-for-churn-telecom	71049	58

V. EXPERIMENT RESULTS AND DISCUSSION

- Performance evaluation using Area under the receiver operating characteristics curve (AUC):

The proposed work is done on two churn datasets which is mentioned in Table1, individual methods and hybrid models are fitted on two datasets. Table3 represents the performance of individual and hybrid methods as well. From that comparison for churn datasets LLM fits well when compared with other methods. By visualizating the output using ROC curve, the performance of two hybrid methods is plotted for DS1as shown in Figure3 and for DS2 is shown in Figure4.

- Performance evaluation Top decile lift (TDL):

Table4: Performance evaluation Top decile lift (TDL)

Top decile lift		
Algorithm	**Datasets**	
	DS1	**DS2**
Decision Tree	2.4	2.4
Logistic Regression	2.7	2.6
Logit leaf model (LLM)	**2.8**	**2.7**
Support vector machine (SVM)	2.6	2.5
Support leaf Model (SLM)	**2.6**	**2.5**

Like AUC , TDL is one of the performance metric. Table4 shows the lift values for both the individual and hybrid methods.

Visualization of the output:

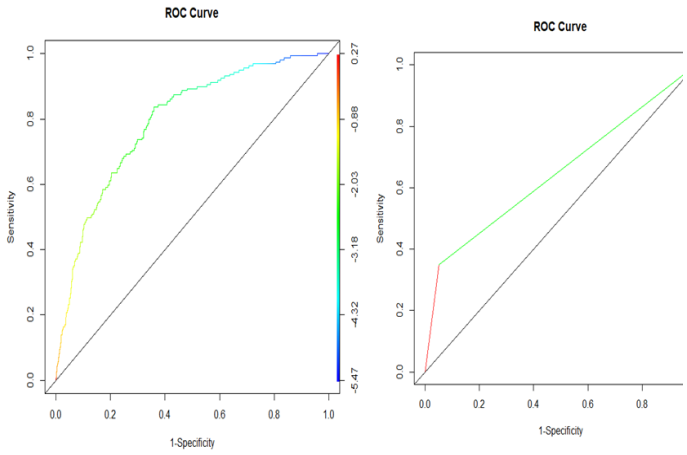

Figure3: ROC curve for DS1

As Figure3 shows the ROC curve for Dataset1for displaying the performance of LLM and SLM on churn datasets. Here only the hybrid methods are compared. LLM gives better performance rather than SLM.

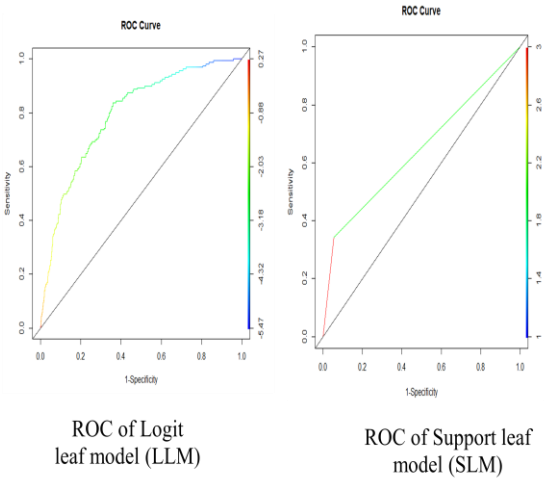

Figure4: ROC curve for DS2

As Figure4 shows the ROC curve for Dataset1for displaying the performance of LLM and SLM on churn datasets. Here only the hybrid methods are compared. LLM gives better performance rather than SLM.

The main observation that can be made from the above results is that SLM does not give better results on churn datasets when compared with LLM because logistic regression gives better results rather than support vector machine on churn datasets. Support vector machine does not work well will skewed data when compared with logistic regression. That's why LLM is better than SLM.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a comparative study was conducted on LLM and a new hybrid classification algorithm SLM with respect to two churn datasets. The performance of these two algorithms was compared with two metrics AUC and TDL. LLM gets highest score with AUC and TDL when compare with both the individual and hybrid algorithms. After comparing the results, SLM does not give best results for churn data when compared with LLM. SVM does not work well with skewed datasets that's why LLM gives better results when compared with SLM.

In future work, there are many opportunities in model variations. Firstly, models that are already used in segmentation phase can be changed to improve the performance of the model. Secondly, selection measures can be changed. Thirdly, it is possible to change the performance metrics. Lastly, there is a chance of replacement of supervised techniques with unsupervised techniques.

REFERENCES

[1] Michael C. Mozer, Richard Wolniewicz, David B. Grimes,"Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry," IEEE Transactions On Neural Networks, Vol. 11, pp. 690-696, September 2000.

[2] Colgate, M., Stewart, K., & Kinsella, R. (1996). Customer defection: A study of the student market in Ireland. International Journal of Bank Marketing, 14(3), 23–29.

[3] Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the customer base of service providers: An examination of the differences between switchers and stayers. Journal of Marketing, 64(3), 65–87.

[4] Shaffer, G., & Zhang, Z. J. (2002). Competitive one-to-one promotions. Management Science, , 48(9), 1143–1160.

[5] Blattberg, R. C., Kim, B. D., & Neslin, S. A. (2010). Database marketing: Analyzing and managing customers. New York, NY: Springer.

[6] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211–229.

[7] Gustafsson, A., Johnson, M. D., & Roos, I. (2005). The effects of customer satisfaction, relationship commitment dimension, and triggers on customer retention. Journal of Marketing, 69(4), 210–218.

[8] Seret, A., Verbraken, T., Versailles, S., & Baesens, B. (2012). A new SOM-based method for profile generation: Theory and application in direct marketing. European Journal of Operational Research, 220, 199–209.

[9] Ali Dehghan, Theodore B. Trafalis, "Examining Churn and Loyalty Using Support Vector Machine", Scienceedu Press, Vol. 1, (4), pp. 153- 161, December 2012.
[10] Marie Fernandes, "Data mining: A Comparative Study of its various Techniques and its", IJSRCSE, Volume-5, Issue-1, pp.19-23, February (2017).
[11] G. Sathyadevi, "Application of CART Algorithm in hepatitis Disease Diagnosis", IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011.
[12] mi zhou, "A hybrid feature selection method based on fisher score and genetic algorithm" , Journal of Mathematical Sciences: Advances and Applications Volume 37, 2016, Pages 51-78
[13] Coussement, K., Lessman, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication. Decision Support Systems, 95, 27–36.
[14] Arno De Caigny, Kristof Coussement, Koen W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees", European journal of operational research 269 (2018) 760-772.
[15] S. JabeenBegum, B. Swaathi, "A Survey for identifying Parkinson's disease by Binary Bat Algorithm", IJSRCSE, Vol 7,Issue 2, pp.17-23, April (2019).

