ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Training and Deploying Churn Prediction Model using Machine Learning Algorithms

[1] Syed Muzamil Basha, [2] Avanti Khare, [3] Jyothi Gadipalli
[1] Assistant professor, SVCET, Chittoor ,[2][3] Student, SNIST, Hyderabad

Abstract: Now a days, in this competitive world, every organization has been investing their major resources like money and manpower towards predicting the customers behavior in maximizing their profitability. Understanding the public opinion in advance to their product release, helps the organizations in taking effective decisions on marketing strategies, Churn Prediction is an approach, through which one can easily attain the behavior of customer as Life Time Value. In our research, we focus more on building a model, that can train the input data records of about 2,666 having 21 attributes of different type and deploy the constructed rules of about 105 to generate reports on churn prediction using Decision tree algorithm. The records in the Dataset, considered in our experiment are classified as positive (2280) and negative (386) manually. In our research, we found that there is good correlation score(0.82) exist between Dates delay and days to settle attributes, upon implementing our proposed model, we achieved 85% of classification accuracy with 70.1% cohen's kappa value. Finally, we would like to conclude that much higher classification accuracy can be achieved using Decision tree algorithm by constructing a very low level Decision rules on the data with good domain knowledge.

Index Terms— Churn Prediction, Life Time Value, Decision Tree, Classification accuracy, cohen's kappa Value.

1. INTRODUCTION

Churn prediction is major distress in Mobile Telecommunication industry (MTI). It is termed as transfer of subscribers from one service provider to another service provider in the search of better services/rates. As per the Taxonomy of Churn Prediction Management, there are two categories of churn one is Voluntary and the other is involuntary churn. In our research we considered only voluntary churn. The average churn rate for MTI is 2.2% per month [1]. For good business support, Data mining have been used in discovering hidden pattern from huge dataset already available with the industry. Based on the knowledge, that to be extracted from databases, Data mining techniques are classified in to classification, clustering and data visualization. In classification any pre-determined model can be induced to classify the instances in to different classes. In [4] the author has focused on combining two DM models in constructing a hybrid data model, which result in better churn prediction. As per the literature review made, Decision tree is the widely used technique in classification process. In our research, the proposed model based on the inputs (Predictors) predicts the call details of the customers.

Our research, is organized as follows. In Section 2, the research work carried work in the area of churn prediction is discussed. Section 3 describes the design of our experiment using KNIME. Section 4 express the result achieved in our experiment. This paper is concluded in Section 5 with discussion on our contribution with limitations and future research direction.

II. LITERATURE SURVEY

In our Literature survey, we found that Ant Miner+ and Advanced Learning based approach (ALBA) has been mostly used. In which, Ant Miner+ uses Ant Colony Optimization (ACO) technique and ALBA uses Support Vector Machine (SVM) in inferring the rules from the huge available unstructured data. In addition to that C4.5 and Ripper are rule induction algorithms used to improve the prediction power.

The author in [8] have achieved 93.87% of classification accuracy with 0.21 standard deviation using Ripper and C4.5 rule induction algorithms. In [6], the author proposed a churn prediction model that uses social influence of players as predictors in the game and stated that an influence vector helps in modeling the real world more accurately. In [7], the author made different perspective on Generalized Additive models (GAM) models by making a detailed comparison of Logistic Regression model and GAM with respect to Classification in accuracy using relationships exists between the covariates of the variable and the logit, that results in better identification of customer risk.

TABLE 1. DESCRIPTION OF WORK CARRIED OUT IN CHURN PREDICTION.

Author	Techniques	Classification Accuracy
Shaaban. E et al. 2012 [9]	DT	77.9%
	ANN	83.7%
	SVM	83.7%
Basha. SM et al. 2017 [10]	Weight Fuzzy rule based model	93.0%
Basha. SM et al. 2017 [11]	Holt-Winters, Auto Regressive Integrated Moving Average (Arima)	92.2% (P-Value)
Basha. SM et al. 2017 [12]	Generalized Linear Model	76.8%

In [13], the author used gradient ascent algorithm in finding out the exact weights of the terms used in determining the sentiment of tweet and used Boosting approach to improve the accuracy of linear classifier. In [14], the author provide a novel way of performing prediction on Breast cancer dataset, compared the performance of three different feature selection algorithm and proved that genetic algorithm is giving best result in selecting the best feature among all the available feature. SVM algorithms gives the best result in predicting the level of certainty in breast cancer. In [15], the author made an attempt to develop an recommender system, helping in searching the item, that might out found by themselves, In which precision and recall measures are used in measuring the performance of proposed model. In [16], the author made an research in solving the problem in Diabetic Retinopathy. In which, the author proposed a Model, which can capable of calculating the weights, that gives severity level of the patient's eye by using weighted Fuzzy C-means algorithm. In [17], the author proposed a build a model for airlines, that can performs sentiment analysis on customer feedback and achieved Vital accuracy. Where as in [18, 19], the author experimented on finding out the impact of feature selection on overall sentiment analysis and stated that Term frequency have greater impact on analyzing sentiments rather than bigram approach.

III. METHODOLOGY

In our experiment, we consider the work flow application using KNIME. In which, user have options to select the basic element called node. Each node in KNIME has its own functionality, based on the needs of the user, one can select the node and configure them as per the application. As the first step, we started focusing on constructing calls data having 21 attributes and 2666 instances of data using contract Data node available in KNIME. In second step, preprocessing stage take cares of missing data using last prediction value technique. Later, the data is partitioning in to 80% training and 20% testing data using partition node in KNIME. In order to train the model, decision tree algorithm is used. We aim to induced the user constructed rules into algorithm as discussed in TDCPK Model. Finally, Evaluation process aims to score the proposed model using the ROC and statistical nodes available in KNIME. For easy documentation, we have discussed the top most rules construction in TDCPK model.

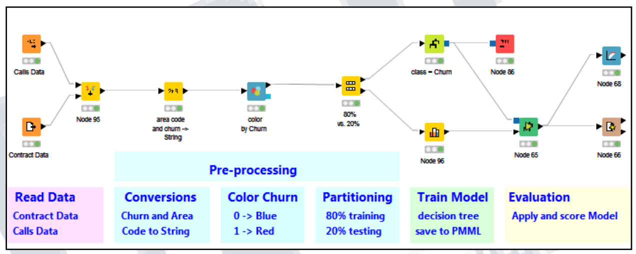

Fig 1. Design of Work Flow using KNIME.

TDCPK Model
Input: Number of fields 21 with names and datatypes "VMail_Message" of type "integer", "Day_Mins" of type "double", "Eve_Mins" of type "double", "Night_Mins" of type "double", "Intl _Mins" of type "double", "CustServ_Calls" of type "integer", "Day_Calls" of type "integer", "Day_Charge" of type "double", "Eve_Calls" of type "integer", "Eve_Charge" of type "double", "Night_Calls" of type "integer", "Night of type "double", "Intl_Calls" of type "integer", "Intl_Charge" of type "double", "Area _Code" of type ="string" (415, 408, 510), "Phone" of type "string", "Account_Length" of type "integer", "Churn" of type "string" (0,1),"Int'l_Plan" of type "integer", "VMail Plan" of type "integer", "State" optype="categorical" dataType="string" (KS,OH,NJ,OK,AL,MA,MO,LA,WV,IN,RI,IA,MT,NY,ID,VT,VA,TX,FL,CO,AZ,SC,NE,WY,HI,IL,NH,GA,AK,MD,AR,WI,OR,MI,DE,UT,CA,MN,SD,NC,WA,NM,NV,DC,KY,ME,MS,TN,PA,CT,ND)
Model: "DecisionTree" Category ="classification" Characteristic="binarySplit"
missingValueStrategy="lastPrediction"
noTrueChildStrategy = "returnNullPrediction"

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Number of Records: Total= 2666.0 (2280 having 0 score and 386 having 1 score)

Output: Confusion Matrix, Statistical Parameters, Classification Accuracy.

First 15 Rules are described for documentation purpose:
if (Day_Charge ≤ 44.96) then ScoreDistribution value="0"
else ScoreDistribution value ="1"
if(CustServ_Calls ≤ 3.5) then ScoreDistribution value="0"
else ScoreDistribution value="1"
if(Int'l_Plan ≤ 0.5) then ScoreDistribution value="0"
else ScoreDistribution value="1"
if(Day_Charge ≤ 38.185) then ScoreDistribution value="0"
else ScoreDistribution value="1"
if (State == "isIn") <Array n="13" type="string">"KS" "OH" "NJ" "AL" "LA" "MT" "VT" "VA" "CO" "WY" "ID" "CA" "SD"</Array> then ScoreDistribution value="0"
else ScoreDistribution value="1"
if (State == "isIn") <Array n="8" type="string">"KS" "NJ" "AL" "LA" "VT" "VA" "CO" "WY"</Array> then ScoreDistribution value="0"
else ScoreDistribution value="1"
if (State == "isIn") <Array n="43" type="string">"OH" "OK" "MA" "IN" "IA" "MT" "FL" "AZ" "NE" "MO" "HI" "NH" "AK" "MD" "AR" "ID" "WI" "OR" "MI" "DE" "CA" "MN" "SD" "WA" "UT" "TX" "NM" "NV" "DC" "NY" "WV" "GA" "KY" "ME" "MS" "RI" "SC" "TN" "PA" "IL" "NC" "CT" "ND"</Array> then ScoreDistribution value="0"
else ScoreDistribution value="1"
if (Day_Charge ≤ 37.655) then ScoreDistribution value="0"
else ScoreDistribution value="1"
if (Intl Charge ≤ 3.389) then ScoreDistribution value="0"
else ScoreDistribution value="1"
if (Intl_Charge ≤ 3.335) ScoreDistribution value="0"
else ScoreDistribution value="1"
if (Eve_Charge ≤ 17.31) then ScoreDistribution value="0"
else ScoreDistribution value="1"
if (Account_Length ≤ 132.0) then ScoreDistribution value="0"
else ScoreDistribution value="1"
if (Account Length ≥ 132.0) then ScoreDistribution value="0"
else ScoreDistribution value="1"
if (Night_Charge ≤ 9.415) then ScoreDistribution value="0"
else ScoreDistribution value="1"
if (Night_Charge ≥ 9.415) then ScoreDistribution value="0"
Stop

IV. RESULT

As stated in [2] SVM is better than traditional logistic regression, where random forests performs better than both. The Scorer node helps to visualize the confusion matrix as shown in figure 3. The proposed model able to achieve 85.05% of classification accuracy with 14.9% error in classification. The result obtained in our experiment can be improved by inducing additional rules in to our model based on subscriber choice. In figure 3 the statistical parameters (Precision, Recall, F-measure, sensitivity and specificity) obtained are plotted with upper and lower limit. Each point on the ROC curve represents a True positive rate and False positive rate (TP, FP) pair corresponding to a particular decision threshold. For interpretation the closer the ROC curve is to the upper left corner of figure 4, the higher the overall classification accuracy of the model. In [5], the author made ROC analysis, as accuracy should not emphasis over a majority classes used in churn prediction.

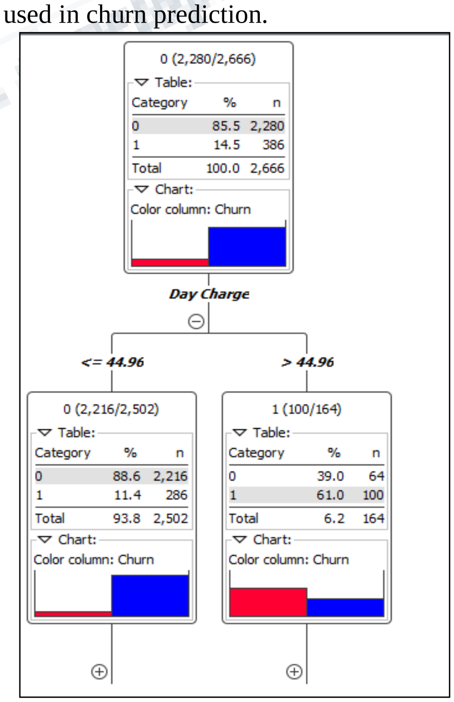

Fig 2. Two Level View of Decision Tree using KNIME.

IFERP
connecting engineers... developing research

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

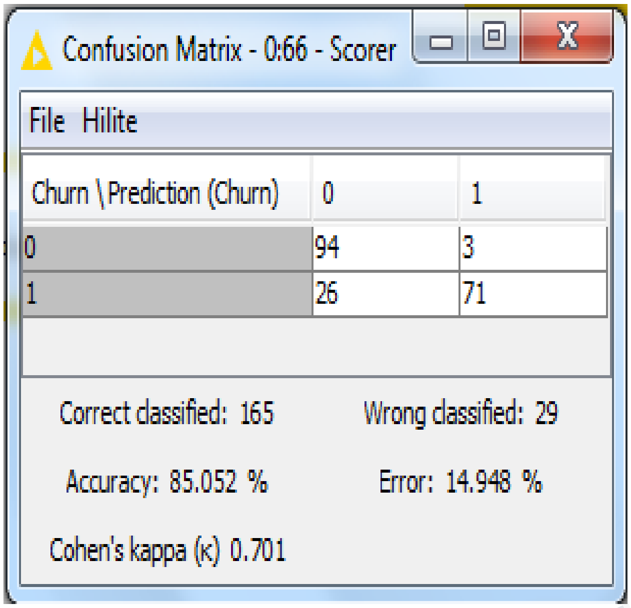

Fig 3.Confusion Matrix with Accuracy and Cohen's Kappa.

Columns: 11	Column Type	Column Index
TruePositives	Number (integer)	0
FalsePositives	Number (integer)	1
TrueNegatives	Number (integer)	2
FalseNegatives	Number (integer)	3
Recall	Number (double)	4
Precision	Number (double)	5
Sensitivity	Number (double)	6
Specifity	Number (double)	7
F-measure	Number (double)	8
Accuracy	Number (double)	9
Cohen's kappa	Number (double)	10

Lower Bound	Upper Bound
71	94
3	26
71	94
3	26
0.732	0.969
0.783	0.959
0.732	0.969
0.732	0.969
0.83	0.866
0.851	0.851
0.701	0.701

Fig. 4. Statistical Parameters obtained using KNIME.

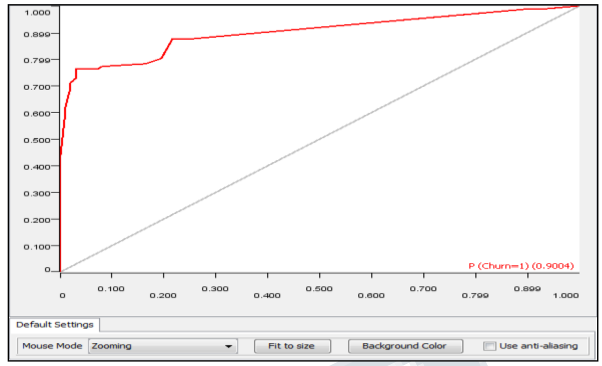

Fig. 5. Plot of ROC Curve.

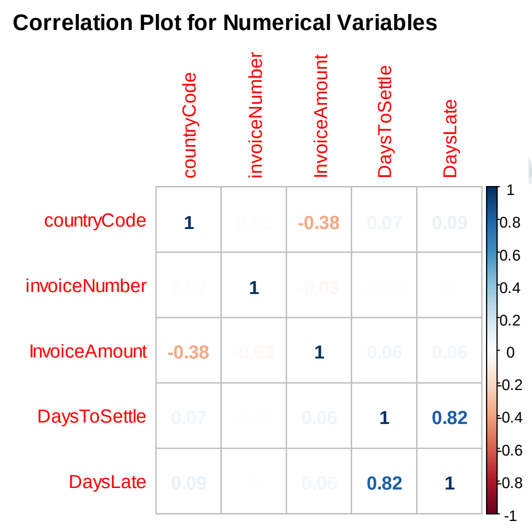

Fig. 6. Correlation Matrix.

Table 2. Comparison of other work with proposed work.

Parameters	Work 1	Work 2	TDCP K
Accuracy(A)	0.713	0.707	0.850
Misclassification Rate(MCR)	0.287	0.293	0.149
Recall(R)	0.586	0.523	0.732
Precision(P)	0.146	0.263	0.783
Prevalence(PV)	0.075	0.151	0.321
F Score(FS)	0.234	0.175	0.831

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Table 3. Data Report of First Row.

Name of the attribute	Value
VMail Message	0
Day Mins	204.9
Eve Mins	135.2
Night Mins	208.2
Intl Mins	10.4
CustServ Calls	5
Day Calls	107
Day Charge	34.83
Eve Calls	102
Eve Charge	11.49
Night Calls	106
Night Charge	9.37
Intl Calls	3
Intl Charge	2.81
Area Code	415
Phone	354-6960
Account Length	103
Int'l Plan	0
VMail Plan	0
State	WA
Prediction (Churn)	1
P (Churn=0)	0.153
P (Churn=1)	0.846

In [3], the author made a detailed comparison of SVM with Artificial Neural Network (ANN), decision tree (DT), logistic regression (LR), and naïve Bayesian (NB) classifier for customer churn prediction in mobile telecommunication and found that SVM have better precision value. The results obtained from our experiment are compared with our first and second work, that are proved to be better in terms of classification accuracy as described in Table 2.

V. CONCLUSION

In our finding, we would like to state that, there is a greater impact of rules construction on classification accuracy. To construct the rules, one should have complete understanding on the working domain. Thus, the organization have high bandwidth in changing their marketing strategies. From the complete dataset consider we have able to predict the one subscribe as churn and the details are given in table 3. In our Future work, we would like to experiment with other classification techniques on the same dataset and compare their performance with respect to classification accuracy.

REFERENCES

[1] Wei, Chih-Ping, and I-Tang Chiu. Turning telecommunications call details to churn prediction: a data mining approach, Expert systems with applications, vol. 23, pp. 103-112, (2002).
[2] Coussement, Kristof, and Dirk Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, Expert systems with applications, vol. 34, pp. 313-327, (2008).
[3] Xia, G.E. and Jin, W.D., 2008. Model of customer churn prediction on support vector machine. Systems Engineering-Theory & Practice, vol. 28 , pp. 71-77, (2008).
[4] Tsai, Chih-Fong, and Yu-Hsin Lu. Customer churn prediction by hybrid neural networks, Expert Systems with Applications, vol. 36, pp. 12547-12553, (2009).
[5] Burez, Jonathan, and Dirk Van den Poel. Handling class imbalance in customer churn prediction, Expert Systems with Applications, vol. 36, pp. 4626-4636, (2009).
[6] Kawale, Jaya, Aditya Pal, and Jaideep Srivastava. Churn prediction in MMORPGs: A social influence based approach. In Computational Science and Engineering, CSE'09. International Conference on, 4, pp. 423-428, (2009).
[7] Coussement, K., Benoit, D.F. and Van den Poel, D., Improved marketing decision making in a customer churn prediction context using generalized additive models. Expert Systems with Applications, vol. 37, pp. 2132-2143, (2010).
[8] Verbeke, Wouter, David Martens, Christophe Mues, and Bart Baesens. Building comprehensible customer churn prediction models with advanced rule induction techniques, Expert Systems with Applications, vol. 38, pp. 2354-2364 (2011).
[9] Shaaban, E., Helmy, Y., Khedr, A. and Nasr, M., 2012. A proposed churn prediction model. International Journal of Engineering Research and Applications, vol. 2, pp. 693-697, (2012).
[10] Basha, Syed Muzamil, Yang Zhenning, Dharmendra Singh Rajput, N. Iyengar, and Ronnie D. Caytiles. Weighted Fuzzy Rule Based Sentiment Prediction Analysis on Tweets, International Journal Of Grid And Distributed Computing, vol. 10, pp. 41-54,

(2017).

[11] Basha, Syed Muzamil, Yang Zhenning, Dharmendra Singh Rajput, Ronnie D. Caytiles, and N. Iyengar. Comparative Study on Performance Analysis of Time Series Predictive Models. International Journal Of Grid And Distributed Computing, vol. 10, pp. 37-48, (2017).

[12] Basha, Syed Muzamil, H. Balaji1, N. Ch. S. N Iyengar and Ronnie D. Caytiles, A Soft Computing Approach to Provide Recommendation on PIMA Diabetes, International Journal of Advanced Science and Technology, vol. 106, pp. 19-32, (2017).

[13] Basha, Syed Muzamil, Dharmendra Singh Rajput, and Vishnu Vandhan. "Impact of Gradient Ascent and Boosting Algorithm in Classification."International Journal of Intelligent Engineering and Systems (IJIES) 11, no.1,(2018): 41-49. DOI: 10.22266/ijies2018.0228.05

[14] Basha, Syed Muzamil, Dharmendra Singh Rajput, N. Iyengar, and D. R. Caytiles. "A Novel Approach to Perform Analysis and Prediction on Breast Cancer Dataset using R." International Journal of Grid and Distributed Computing, Vol. 11, No. 2 (2018), pp.41-54, http://dx.doi.org/10.14257/ijgdc.2018.11.2.05

[15] Vivek P. Khadse, Basha, Syed Muzamil, N. Iyengar, and D. R. Caytiles," Recommendation Engine for Predicting Best Rated Movies " International Journal of Advanced Science and Technology, Vol.110 (2018), pp.65-76, http://dx.doi.org/10.14257/ijast.2018.110.07

[16] Suvajit Dutta, Basha, Syed Muzamil, N. Iyengar, and D. R. Caytiles. " Classification of Diabetic Retinopathy Images by Using DeepLearning Models." International Journal of Grid and Distributed Computing,Vol. 11, No. 1 (2018), pp.89-106, http://dx.doi.org/10.14257/ijgdc.2018.11.1.09

[17] Divisha khaturia, , Syed Muzamil, N. Iyengar, and D. R. Caytiles. " A Comparative study on Airline Recommendation System Using Sentimental Analysis on Customer Tweets." International Journal of Advanced Science and Technology, Vol.111 (2018), pp.107-114, http://dx.doi.org/10.14257/ijast.2018.111.10

[18] Basha, Syed Muzamil, and Dharmendra Singh Rajput. "Evaluating the Impact of Feature Selection on Overall Performance of Sentiment Analysis." In Proceedings of the 2017 International Conference on Information Technology, pp. 96-102. ACM, 2017.

[19] Basha, Syed Muzamil, and Dharmendra Singh Rajput. "Fitting a Neural Network Classification Model in MATLAB and R for Tweeter Data set." Proceedings of International Conference on Recent Advancement on Computer and Communication. Springer, Singapore, 2018.