COMPARING THE EFFECTIVENESS OF ARTIFICIAL NEURAL NETWORK WITH DECISION TREE IN CLASSIFICATION OF CUSTOMER CHURN

Nafisseh Heiat, College of Business, Montana State University-Billings, 1500 University Drive, Billings, MT 59101. 406-657-2224, nheiat@msubillings.edu

ABSTRACT

Customer turnover or churn is a very important concern in telecommunication businesses. In this study, Artificial Neural Network (ANN) and Decision Tree (C5) algorithms are applied and results are analyzed to determine the most efficient model for predicting customer churn in telecommunications.

INTRODUCTION

In many industries, especially telecommunication field, customer turnover or churn is an important concern. Churn is a term for subscribers switching from one telecommunication provider to another. In this study I am interested in voluntary churn. Voluntary churn is when the customer decides to leave the current provider for a variety of reasons like subscription costs, quality of service, more attractive incentives from competition etc. Customers switch from one provider to another results in loss of considerable profit. According to Lu Junxiang telecommunications industry experiences an average of 30-35 percent annual churn rate and it costs 5-10 times more to recruit a new customer than to retain an existing one.

In order to identify high risk customers that may switch to another provider and manage churning rate, we need to develop reliable models. Once a reliable and accurate churn model is developed and tested, companies may use the model for churn management. In this study, I am going to apply Artificial Neural Network (ANN) and Decision Tree (C5) algorithms to a telecommunication dataset and compare the results obtained from ANN with results obtained from C5 to determine the most efficient model in terms of classification accuracy.

DATA PROCESSING AND METHODOLOGY

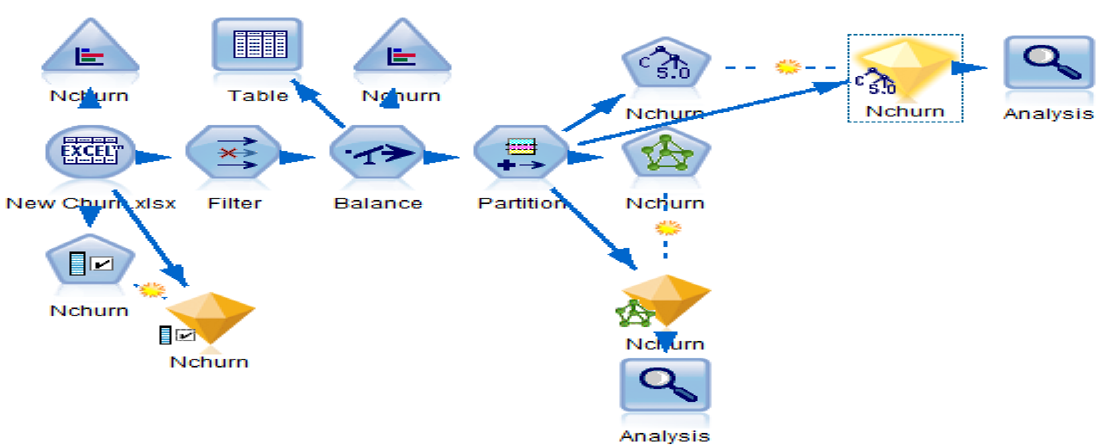
The dataset used in this study has 3333 entries and 21 attributes. The attributes of the dataset are shown in table 1.

Table 1. Attributes in telecommunication dataset

S.No.	Attribute name
1	State
2	Account. Length
3	Area. Code
4	Phone
5	Int .1 .Plan
6	VMail.Plan
7	VMail.Message
8	Day.Mins
9	Day.Calls
10	Day.Charge
11	Eve.Mins
12	Eve.Calls
13	Eve.Charge
14	Night.Mins
15	Night.Calls
16	Night.Charge
17	Intl.Mins
18	Intl.Calls
19	Intl.Charge
20	CustServ.Calls
21	Churn.

Figure 1 shows the model developed for this study.

Figure 1. Model developed for this study



I used the telecom churn dataset (available at http:/www.dataminingconsultant.com/data churn.txt) in this study. The original data set had a number of categorical variables, some of which have been transformed into a series of binary variables so that they could be appropriately handled by the data mining software.

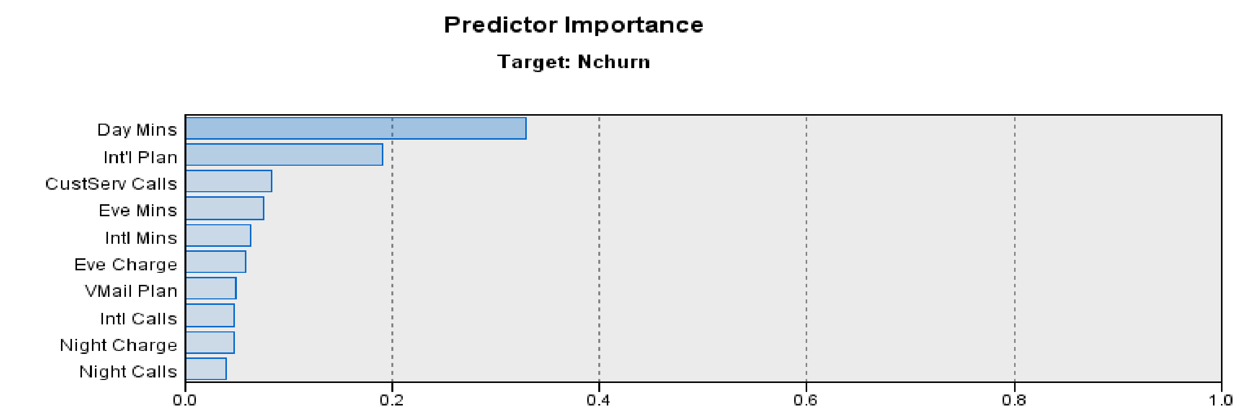
In Figure 1, my model starts with selecting the data set for the analysis. It follows with a feature selection node that identifies important variable for classifying the customer churn. Filter node includes the important variables and assigns the appropriate data type to the target/dependent and input/independent variables. Next, the dataset is partitioned to training and testing sets (70%, 30%). Since the original dataset has an un-even number of non-churned customers (85.1%) and churned customers (14.49%) a balance node is added to make the distribution of churned and non-churned customers almost equal.

Next C5 and ANN algorithms are applied to the dataset and analysis and evaluation node are added to analyze the results.

DECISION TREE (C5) ANALYSIS RESULTS

Figure 2 shows the important variables in classifying customer churn according to C5 algorithm.

Figure 2. Important Variables Determined By C5



In Figure 3, confusion matrix of C5 indicates the accuracy of the classification of the algorithm.

Figure 3. C5 Confusion Matrix

Coincidence Matrix for $C-Nchurn (rows show actuals)

'Partition' = 1_Training	1.000000	2.000000
1.000000	1,877	113
2.000000	23	1,867

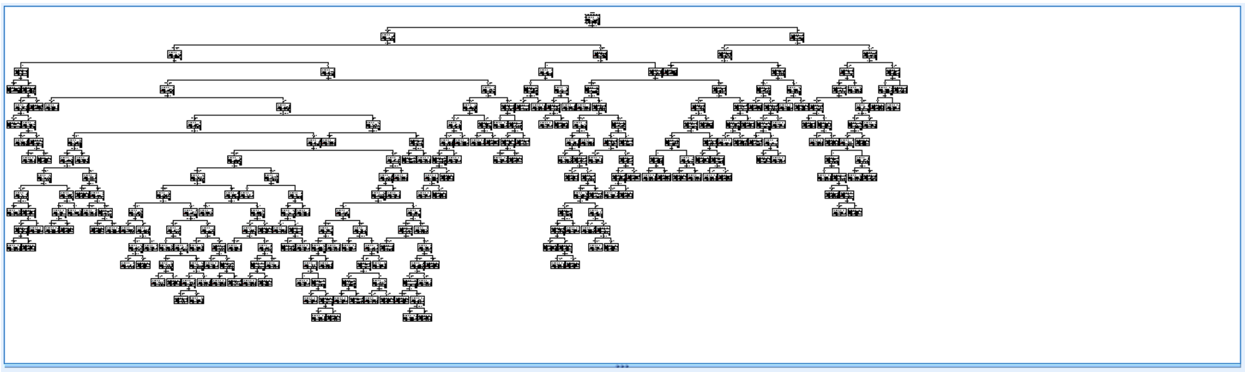
'Partition' = 2_Testing	1.000000	2.000000
1.000000	824	36
2.000000	11	810

Classification accuracy based on validation (testing) dataset for the non-attrition customers is 94.7% and for the attrition customers is 98.4%.

C5 algorithm created 77 rules set for classifying non-attrition customers and 58 rules set for classifying attrition customers. In case of developing a decision support system or an expert system, some of the rules that are redundant or are not used by decision makers should be removed or consolidated.

Figure 4 displays the created decision tree diagram.

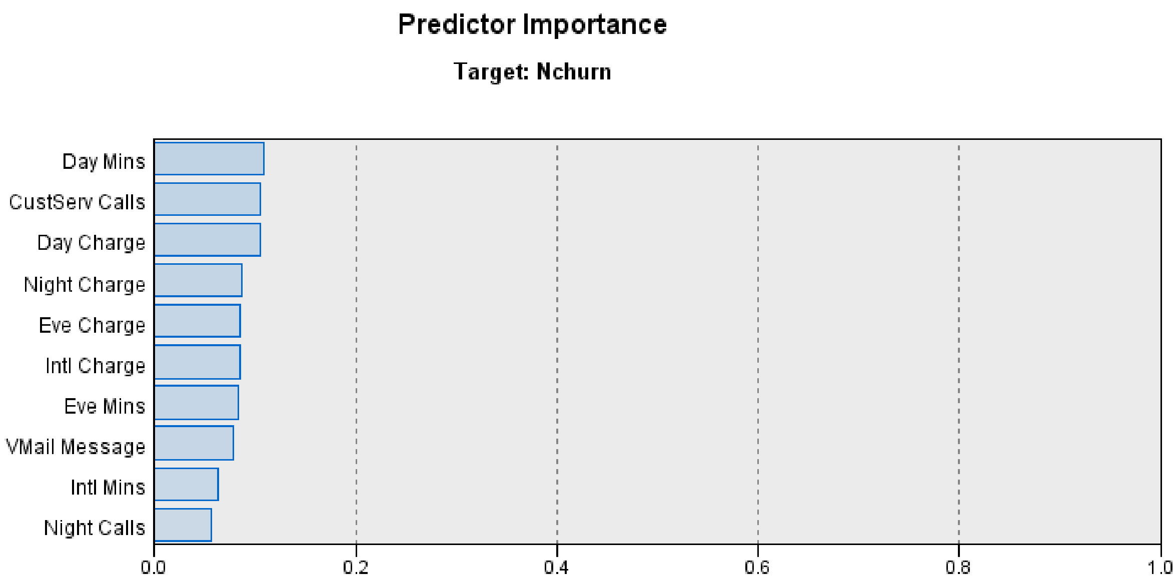
Figure 4. Decision Tree Diagram



ARTIFICIAL NEURAL NETWORK (ANN) RESULTS

Figure 5 displays the important variables in classifying customer churn according to ANN algorithm.

Figure 5. ANN Important Variables



In Figure 6, confusion matrix of ANN indicates the accuracy of the classification of the algorithm.

Figure 6. ANN Confusion Matrix

Coincidence Matrix for $N-Nchurn (rows show actuals)

'Partition' = 1_Training	1.000000	2.000000
1.000000	1,805	188
2.000000	284	1,603

'Partition' = 2_Testing	1.000000	2.000000
1.000000	764	93
2.000000	136	687

Classification accuracy based on validation (testing) dataset for the non-attrition customers is 89.46% and for the attrition customers is 85.35%.

CONCLUSION

The performance of Decision tree (C5) in classifying non-attrition customers and attrition customers (94.7%, 98.4%) is better than performance Artificial Neural (ANN) in classifying non-attrition customers and attrition customers (8.46%, 87.4%). In addition, Decision Tree algorithm creates a rule set that may be used to develop a decision support or an expert system.

REFRENCES

References are available upon request.