Comparative Analysis of Machine Learning Techniques for Telecommunication Subscribers' Churn Prediction

Saad Ahmed Qureshi, AmmarSaleemRehman, Ali Mustafa Qamar, Aatif Kamal, Summaya Mumtaz, KhurramJaved
School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
{09bicseaqureshi, 09bicsearehman, mustafa.qamar, aatif.kamal, 12mscssmumtaz, 12mscskjaved}@seecs.edu.pk

ABSTRACT: *During the last two decades, the mobile communication has become a dominant medium of communication. In numerous countries, especially the developed ones, the market is saturated to the extent that each new customer must be won over from the competitors. Advancements in technology and rapid improvements in telecom industry have provided customers with many choices. Customer retention is one of the major tasks for the telecom industry. On the other hand, public policies and standardization of mobile communication now allow customers to easily switch over from one carrier to another, resulting in a highly fluid market. Churn refers to customers who will leave or turn to other service providers. Acquiring new customers is much more expensive as compared to retaining existing customers. Therefore, it is far more cost-effective for service providers to predict customers who will churn in future and customize services or packages according to the customer's demands. As a result, churn prediction has emerged as one of the most crucial Business Intelligence (BI) applications that aim at identifying customers who are about to transfer to a competitor. In this paper, we present commonly used data mining techniques for the identification of customers who are about to churn. Based on historical data, these methods try to find patterns which can identify possible churners. Some of the well-known algorithms used during this research are Regression analysis, Decision Trees and Artificial Neural Networks (ANNs). The data set used in this study was obtained from Customer DNA website. It contains traffic data of 106,000 customers and their usage behavior for 3 months. The data set comprises of 48 variables. Spearman's correlation coefficient is used to select the variables of high impact.In order to solve the problem of class imbalance in the data set, re-sampling is used.The results show that the decision treesisthe most accurate classifier algorithm while identifying potential churners.*

Keywords: Churn Prediction, Business Intelligence, Data Mining, ANN, Decision Trees

Received: 25 August 2013, Revised 29 September 2013, Accepted 4 October 2013

1. Introduction

In a competitive telecommunications market, customers are the ones who choose their service providers. Therefore, the customer becomes the central focus of the carriers' activities. Customer requirements not only determine service offerings, but they have also got an impact on the organizational structure of the company in order to focus on particular types of customers.

As we review the evolution of the telecommunications industry, it is evident that many cellular companies are aggressively moving (or have already moved) from a business model based on a product strategy to the one based on customer strategy. This

market is characterized by customer relationships, products customizations, and profitability. Telecommunications companies worldwide are exploring business intelligence solutions to gain competitive advantage over their competitors. The key solutions for which telecommunication companies are investing include customer retention, target marketing, campaign management, Customer Relationship Management (CRM) systems in order to streamline the network assets. In this paper, we present a churn prediction model which helps in identifying customers that are at the risk of churning and must be retained, while dealing with the problem of class imbalance through various re-sampling methods.

This paper is organized as follows. Section II describes the related work. The process of Data acquisition has been described in detail in Section III. Different evaluation methods are discussed in Section IV whereas feature selection is given in the next section. The problem of class imbalance along with its possible solutions is detailed in Section VI followed by experiments and results. The process of improving upon the basic methods using derived variables is provided next. The last section concluded the paper along with shedding light on future perspectives.

2. Related Work

During the last few years, there has been a lot of research in the field of churn prediction. Lazarov and Capota [1] stated in their study that customer retention is far more economical than customer acquisition. In their work, Artificial Neural Networks (ANNs) performed better ascompared to other conventional algorithms. Furthermore, they argued that a good prediction model has to be constantly updated and should use a combination of different datamining techniques. In another casestudy [2], churn prediction was done using regression models, where each model comprised of different sets of variables and coefficients. A total of 6 regression models were used over a specific time period. Two model shaving a churner to non-churner ratio of 1:1 and 2:3 for three different analysis periods of 4,6 and 8 months were used. The regression model with re-sampled churnertonon-churner ratio of 2:3, based on data over 8 months gave the best results during the testing phase. In that study, the authors concluded that due to the dynamic nature of a customer, the logistic regression model shad to be updated frequently in order to achieve higher accuracy.

Umayaparvathi and Iyakutti [3] performed churn prediction using ANNs and decision trees. They found that the decision treessur passed the former interm so faccuracy. They divided their project into five phases:data acquisition,data preparation, calculating derived variables, extracting variables, and model construction.

Jadhav and Pawar[4] claimed that the biggest revenue leakage is because of customer churning, which causes an unnecessary burden on telecom companies. In their work, they compared statistical techniques with different data mining techniques. Statistical techniques are most often limited in scope and capacity. The main objective of this study was to predict if a particular customer will churn, well before the actual churning of that customer. For that purpose, the data was collected from In-house customer database, external sources as well as a research survey. Many data mining tools are available for training decision support system to discriminate churning and non-churning customers but for this research back propagation algorithm was used.

Chandrasekhar [5] applied analytical customer relationship management techniques to get the better insight of the complete process of customer churn. In this research,many important variables were extracted which could help in customer management system. Churn prediction was performed using decision trees because of the fact that this technique provides rules that business users can easily understand.

In another research work, Shaaban et al. [6] introduced a simple model using all data mining techniques to keep record of customers and their behavior regarding churn. Decision trees,SVM and ANN were the different techniques used for classification whereas K-Means was used for clustering.

We can safely conclude from the existing research in the field of customer churn prediction, that there is not a single model that could give the highest accuracy in all of the cases. Instead, the performance of every algorithm will strongly depend on the characteristics of the data. In this study, we ran many of the conventional algorithms on our dataset. In addition to that, we

have also suggested some methods to deal with very commonly occurring problem in the telecommunication industry, known as the class imbalance problem. The problems along with its possible solutions are discussed in SectionV.

3. Data Acquistion

The data set used in this study was acquired from an online source. This data set is from a Telecom operator with approximately 106,000 customers (active and disconnected). Traffic type (outgoing, incoming, voice, SMS (Short Message Service), data); with traffic destination (on-net, off-net), rate plan, loyalty, traffic behavior etc. are some of the main attributes of this data set. The data set consists of a total of 48 variables. This data set is divided into two sub-data sets: the first one (churn data set1) with the traffic Figures for 3 months (approximately 300, 000 records) and the second one (churn data set2) with the profile variables for each customer (rate plan, contract renewal date, status, deactivation date, value segment etc.). The customer ID is the key variable for the two sub-data sets. The customers in the data set are classified by a dichotomous variable called *Status* (active or churn). A customer will be classified as *Active* if he/she continues to use the network. On the other hand, a customer will be classified as a Churner, in case the contract with the network is terminated. The list of all the variables used can be found at the site for the data source. Table 1 shows the descriptive statistics of the variables used in this research.

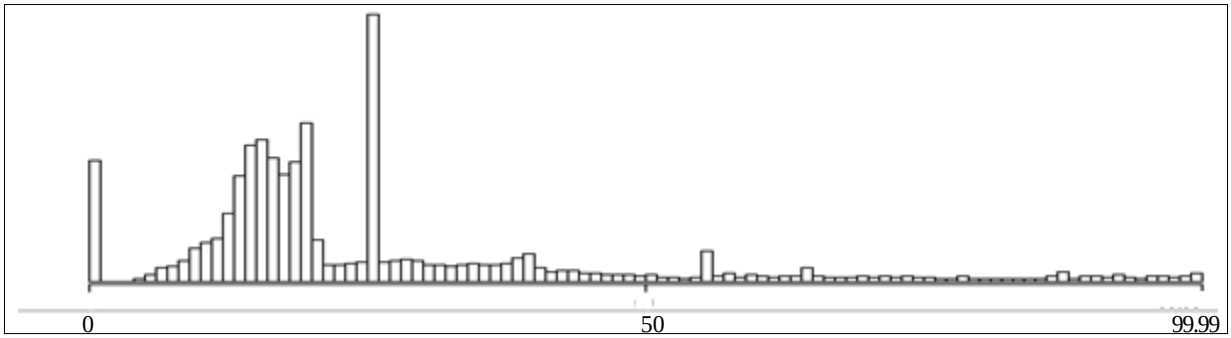

Figure 1. Distribution for Credit_Score

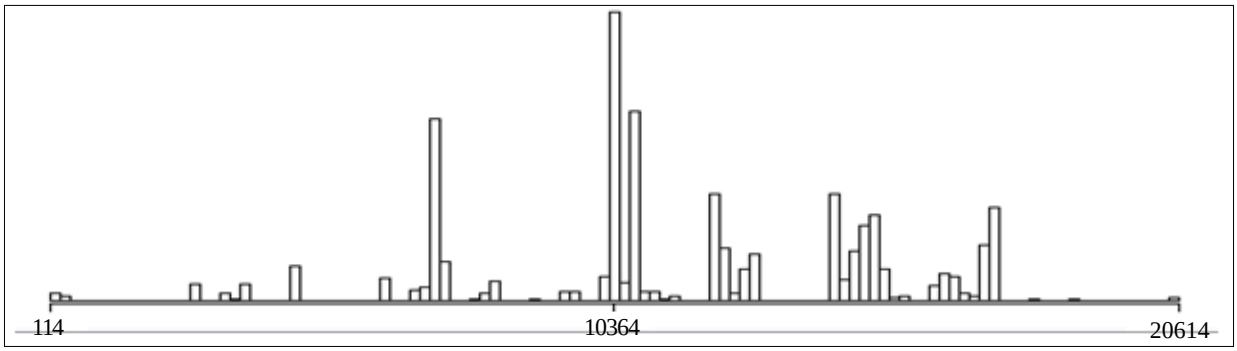

Figure 2. Distribution for Rate_Plan

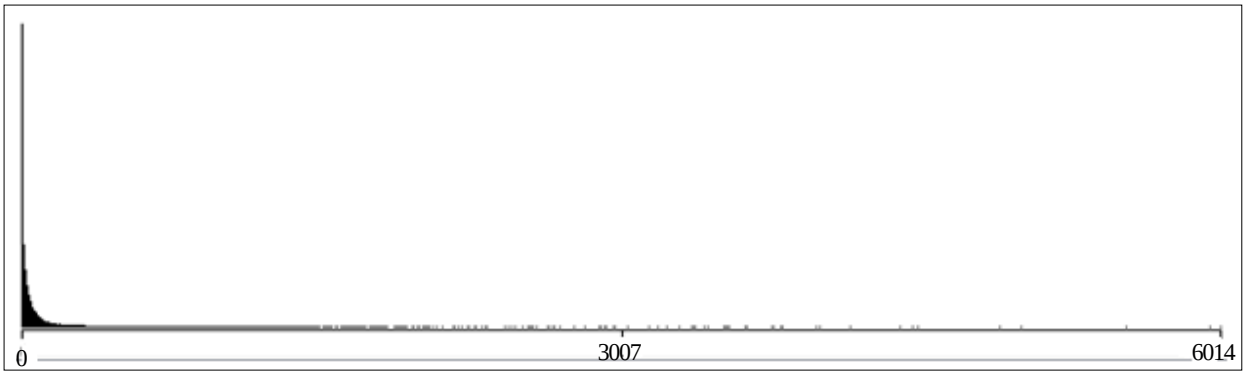

Figure 3. Distribution of duration for On-Net calls (Duration_Onnet_Inc)

Variables	Statistics						
	N	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance
NUM_MMS_OUT	344931	0	116	93608	.27	1.822	3.320
DURATION_VAS_OUT	344931	0	2522	232278	.67	13.433	180.440
NUM_CALLS_FIXED_OUT	344931	0	1128	6812796	19.27	28.093	789.204
DURATION_FIXED_OUT	344931	0	5954	995704	28.921	64.764	4194.330
NUM_CALLS-ONNET_OUT	344931	0	1131	10186555	29.53	43.840	1921.944
DURATION_ONNET_OUT	344931	0	9290	14656503	42.49	114.548	13121.287
NUM_SMS_INTER_OUT	344931	0	1077	355566	1.03	90542	91.057
NUM_CALLS_FIXED_INC	344931	0	1075	5111034	14.82	25.267	638.422
DURATION_FIXED_INC	344931	0	1888	7935027	23.00	42.765	1828.851
DURATION_CMP_INC	344931	0	3395	2511237	72.80	105.759	11184.903
NUM_CALLS_ONNET_INC	344931	0	1744	9850363	28.56	47.832	2287.856
DURATION_ONNET_INC	344931	0	6014	16881378	48.94	123.518	15256.609
NUM_CALLS_INTER_INC	344931	0	856	406115	1.18	60463	41.767
DURATION_INTER_INC	344931	0	1844	1113564	3.23	24.697	609.960
NUM_SMS_CMP_INC	344931	0	3077	7694462	22.31	68.335	4669.707
NUM_SMS_INTER_INC	344931	0	1269	519530	1.51	11.166	124.680
ACTIVITY_DAYS_OUT	344931	0	31	7293004	21.14	11.072	122.592
ACTIVITY_DAYS_INC	344931	0	31	6971898	20.21	11.487	131.951
DISTINCT_CALLERS_INC	344931	0	1074	8433185	24.45	27.206	740.149
NUM_CALLS_CMP_GSM_OUT	344931	0	2007	23672302	68.63	79.937	6389.970
NUM_SMS_CMP_OUT	344931	0	3086	6189322	17094	64.600	4173.134
DEACTIVATION_MONTHAREA	344931	0	199802	28845778569	83627.68	98548.516	9711810038.743
AREA	344931	1	52	14822359	42.97	13.671	186.904
FIRST_RENEWAL_DATE	344931	0	35734	11931367289	34590.53	1043.002	1087853.813
RATE_PLAN	344931	114	20614	3993971534	11579.04	3635.108	13214010.126
LAST_RENEWAL_DATE	344931	0	35970	6324213231	18334.72	17720.497	314016027.532
CREDIT_SCORE	344931	.0000	99.9900	8940339.9400	25.919213	20.0076655	400.307
DURATION_PER_FIXED_OUT	344931	.0000	80.5714	358049.1478	1.038031	1.1955681	1.429
DURATION_PER_ONNET_OUT	344931	.0000	122.0000	366427.8714	1.062322	1.2717266	1.617
DURATION_PER_FIXED_INC	344931	.0000	56.0000	405820.1492	1.176526	1.5107570	2.282
DURATION_PER_ONNET_INC	344931	.0000	121.0000	392387.1604	1.137582	1.4231439	2.025
DURATION-PER_INTER_INC	344931	.0000	89.0000	165448.0413	.479655	1.8030095	3.251
Valid N (listwise)	344931						

Table 1. Descriptive Statistics

Moreover, Figure 1-4 shows the distribution for some of the variables used in the churn prediction algorithms. The credit score for most of the customers is less than 25 as shown in Figure 1.

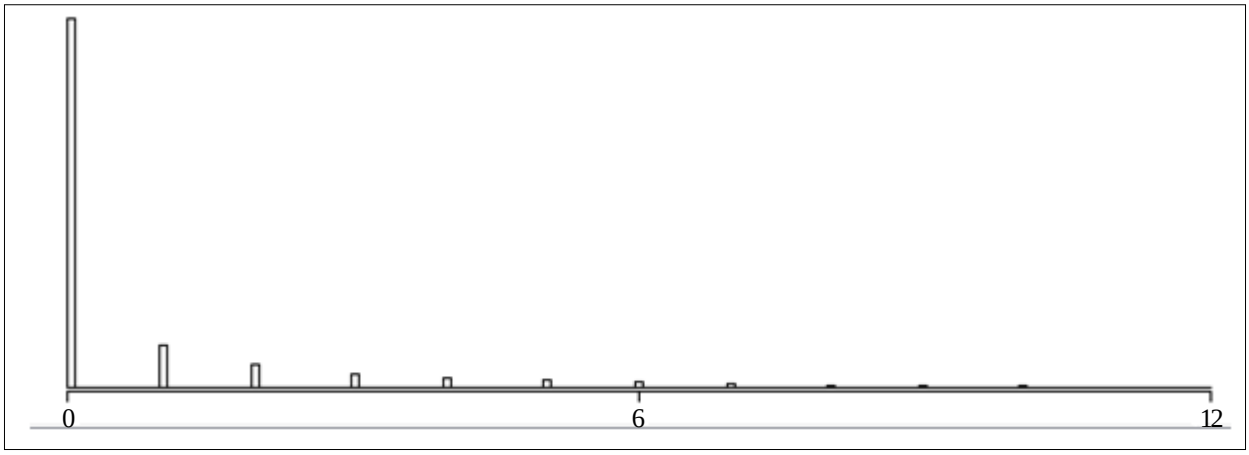

Figure 4. Distribution for Penalities_For_Non_Payment

4. Evalution Methods

In this paper, we consider precision, recall, and F-measure as the methods of evaluation to examine the performance of different prediction models. Table 2 shows the confusion matrix in order to calculate these evaluation measures.

	Predicted Class	
Actual Class	Active	Churn
Active	*a*	*b*
Churn	*c*	*d*

Table 2. Confusion Matrix (Churn Prediction)

Recall: It is the proportion of Active (or Churn) customers that were correctly identified [4]. It is calculated using Equation 1 and 2.

$$\text{Recall } (Churn) = \frac{d}{c + d} \tag{1}$$

$$Recall\ (Active) = \frac{a}{a + b} \tag{2}$$

Precision: It is the proportion of the predicted Active (or Churn) cases that were correct [4]. It can be calculated using Equation 3 and 4.

$$Precision\ (Churn) = \frac{d}{b + d} \tag{3}$$

$$Precision\ (Churn) = \frac{a}{a + c} \tag{4}$$

F-Measure: It the harmonic mean of recall and precision. It is calculated using Equation 5 [5]:

$$F = \frac{2\ .\ Recall\ .\ Precision}{Recall + Precision} \tag{5}$$

5. Feature Selection

Before training a model with the conventional machine learning algorithms, one of the essential steps is to select the right group of variables or features as predictors. In order to determine whether a variable has any predictive significance in an analysis, we calculate its p-value with respect to the target variable. The P-value is the probability that the sample data observed is by pure chance or in statistical terms, the probability that the *null hypothesis* is true. A general rule of thumb is to reject the null hypothesis

if the p-value is below 0.05 for a sample. Therefore, in order to obtain the best set of predictor variables for the analysis, all of the variables having p-values above 0.05 were discarded.

Spearman's Correlation: In statistics, correlation is an important measure to test any kind of dependence or relationship between two variables. The most commonly used correlation test is the Pearson correlation, which is most suited for continuous sets of normally distributed data. In case of the given data set, it was observed that most of the variables did not present a normal distribution. Therefore, another measure of statistical dependence known as the Spearman's Correlation was used in order to identify the variables that were closely correlated to the status of the customer. The top five variables with the highest spearman's correlation coefficient status are given in Table 3.

Variable	Correlation Coefficient
Credit score	0.731
No. of Penalities for Non-payment	0.500
No. of outgoing calls to rival networks	0.279
No. of Incoming SMS from rival networks	0.218
No. of days of Outgoing activity	0.208

Table 3. Spearman's Correlation Analysis

6. Class Imbalance

All kinds of data have different characteristics. Some of these characteristics might pose problems for data mining algorithms in order to extract the meaningful patterns in the data. For example, in the data we used, one of the major problems encountered was class imbalance. In case of class imbalance, the ratio of the output categories is one-sided to the extent that the learning algorithm only predicts the majority class [9]. For example in case of our data, there were 100, 264 active users (94.1%), whereas there were only 6231 churners (5.9%). This presented a typical case of class imbalance. As a result, logistic regression, decision tree as well as ANN made all of the predictions in favor of the majority class (active class in this case).

One of the methods to deal with the problem of class imbalance is re-sampling. There are two ways in which we can do that: over-sampling or under-sampling [10]. In the former case, we use only a subset of the majority class in order to train our data [11]. However, in our case, we removed a random selection of customers from the set of active ones, to the extent that the ratio of the churners and the number of users who would stay active would be roughly the same. Such a ratio would no longer present a case of class imbalance. On the other hand, random over-sampling increases the strength of minority class by replicating a random selection of the existing minority class. In the case of random sampling, we have to be careful that we do not over-sample our data to the extent that it leads to over-fitting. In the case of the given data, we will replicate the churn entries so that the data set no longer presents a case of either class imbalance or over-fitting. In our case, we kept the churners to active ratio to 40:60 approximately [12].

Algorithm	With Class Balance	Re-Sampled
Logistic Regression	0	110764
ANN's	0	80293
kNN's	9	163989
Decision Trees	377	133062

Table 4. Number of Customers predicted as Churns

The results in Table 4 show that compared to the data set with class imbalance, the one that was re-sampled gave unbiased results. With no class imbalance, different machine learning algorithms could now be executed and fairly evaluated on the given data set. Decision trees were able to predict some customers as churners even without resampling. On the other hand, other algorithms (logistic regression, ANN and kNN) were not able to predict even a single customer as churner without performing resampling.

7. Experiments and Results

This section presents the results of many data mining algorithms on the re-sampled data set, using the measures of recall and F-measure.

7.1 Regression Analysis

A brief overview of linear regression as well as logistic regression is next presented. Linear regression is used so as to predict a continuous dependent output variable, using one or more continuous independent input variables using a model based on the straight line equation [10].

$$Y = B_0 + B_1 X_1 + B_2 X_2 + .. \quad (6)$$

where Y represents the continuous target variable, X stands for the input variables playing the role of predictors, B_0 covers all of the errors as well as the noise and the factors that affect the output variable other than the predictors. The rest of the B values represent the coefficients to the predictors. Their values determine the weight and the impact of each predictor and the importance it has in predicting the output variable.

Figure 5 shows a scatter plot with simple linear regression having only one independent variable *DISTINCT_CALLERS_OUT*, representing the number of outgoing distinct callers and output target variable *NUM_CALLS_CMP_GSM_OUT* (the number of outgoing calls to competition networks). The regression line in Figure 5shows a direct correlation between the two variables, implying that with an increase in the number of outgoing distinct callers, the number of calls to rival networks is also likely to increase.

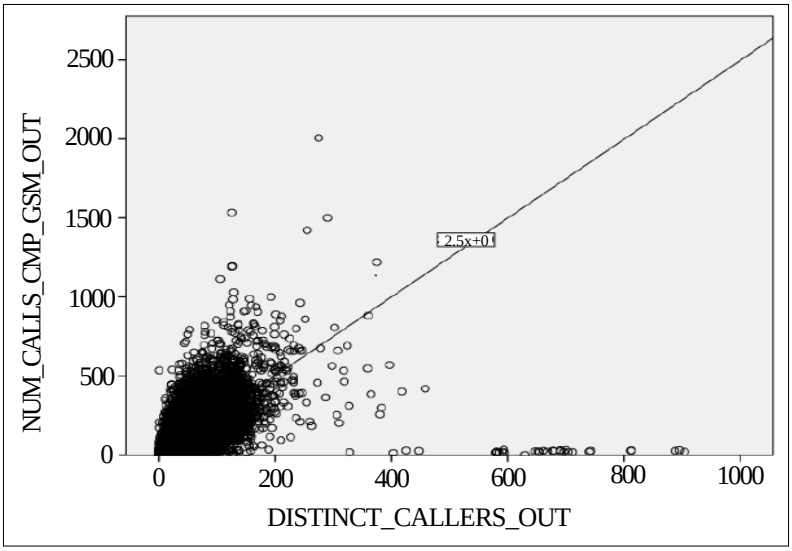

Figure 5. Simple Linear Regression

7.1.1 Logistic Regression

Linear regression is only applicable if we have got a continuous dependent variable and one or more independent variables. In case the target variable is categorical, we use a variant of regression known as logistic regression [1]. Since we have a dichotomous categorical outcome, and most of our independent variables were continuous in nature, logistic regression turned out to be the best choice. In the given data set, the status of the subscriber is a dichotomous variable. While performing the analysis, we model the conditional probability of our customer churning in the near future as a function of the given continuous variables. In order to obtain the conditional probability, we pass the straight line linear regression equation through the logistic function as shown in Equation 7.

$$P(Y \mid x) = \frac{1}{1 + e^{Y}} \quad (7)$$

where $P(Y \mid x)$ is the conditional probability obtained and Y is the simple linear regression equation.Based on whether the conditional probability of the customer is more than or less than the value of 0.5, they are classified as active or churn respectively. In this study, churn represents the target group. That's why a probability of more than 0.5 would classify the customer as a churner. Figure 6 shows a logistic function graph.

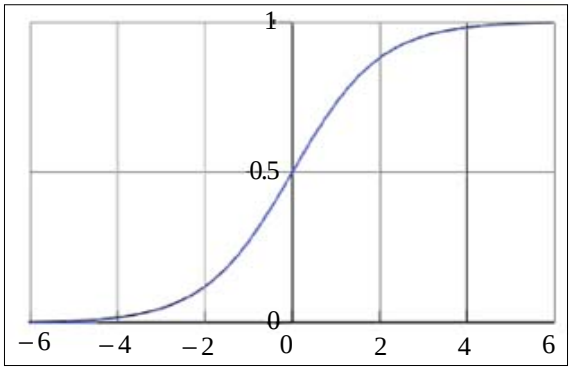

Table 6. Logistic Curve

The results in Table 5 are based on the subscriber data for two months trained with the regression model. The overall accuracy for the task of predicting the customer status was 62.9%, out of which 78.7% of the active users were correctly identified. Conversely, only 45% of the total churners were identified, which is too low. One could also note that although the algorithm did a good job in identifying the active cases, it failed to perform sufficiently well while identifying the churners.

	Recall	F-Measure
Active	0.787	0.720
Churn	0.450	0.515

Table 5. Logistic Regression Test

7.2 Artificial Neural Networks (ANNs)

We also implemented a feed-forward ANN also known as Multi-Layer Perceptron (MLP). It consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. The goal of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown.

In order to get an ANN work, we trigger an input node that in turn triggers the nodes to which it is connected. In neural networks, sets of input nodes are connected to the output nodes through connections, where each connection has a weight associated with it. Neural networks have zero or more hidden layers with arbitrary number of nodes between the input and output nodes, which makes it easier to regulate the weight of each node in order to satisfy the input and output relationships. ANN is applied on the given data set by varying the number of hidden layers from 1 to 10. The learning rate is set to 0.3 and the value of momentum was set as 0.2. The results obtained with ANN are described in Table 6 [14]. We can observe that the results are similar to the ones obtained by regression. Whereas the recall for active users has increased, the recall for churners (32.5%) is still way behind and is even lesser than the one observed with regression.

	Recall	F-Measure
Active	0.823	0.698
Churn	0.325	0.419

Table 6. Nueral Network Result

7.3 K-Means Clustering

K-means algorithm is the most simple and widely used algorithm for clustering. *K* refers to the number of clusters in which that

data set will be divided. The value of *K* has to be specified before the algorithm starts executing. Consequently, the first step in order to use thisalgorithm is to identify *K*,the number of seeds. This is done by taking *K* different observations and assigning them as seeds. This is followed by assigning the rest of the observations to one of these seeds based on their proximity. The proximity could be calculated using distance (such as Euclidean distance, Manhattan distance etc.) or similarity (e.g. cosine similarity). During eachepoch,the distance of the data points is calculated from all clusters and is associated with the cluster having the least distance. The algorithm stops when there is no change in cluster assignments of the data points. Table 7 shows the results obtained by applying the K-means algorithm. As we can see from the results, the recall values for active users as well as the churners are still too low.

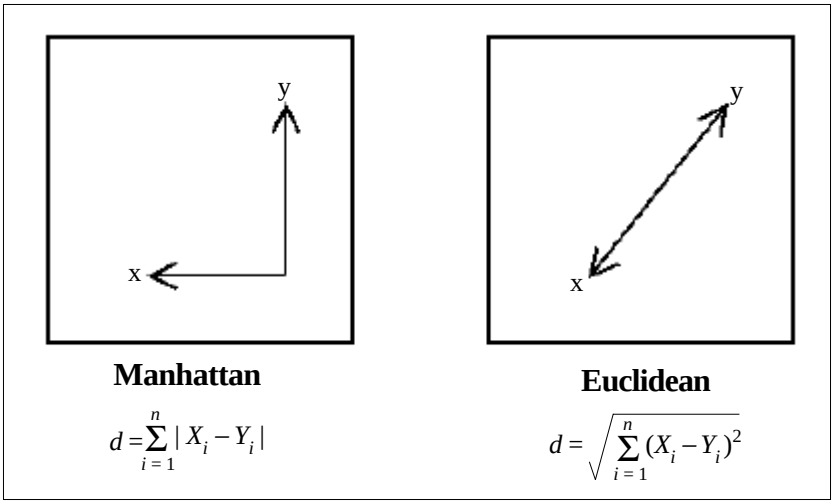

Figure 7. Different Distance Formulas

Table 8 shows the result of K-means clustering using Euclidean distance.

	Recall	F-Measure
Active	0.503	0.529
Churn	0.445	0.416

Table 7. K-means Clustering Result

7.4 Decision Trees

In decision trees, we have classification or regression models in the form of a tree structure. The variable for the root node is selected based on its predictive significance represented by its p-value. In the context of this study, each node represents one of the traffic usage attributes of the customer. Based on the customer's value for that attribute, it will branch out to further nodes until it reaches the leaf node, which will either be a churn node or an active node [15].

SPSS statistics software provides a possibility to use one of the four variations of decision trees, namely: CHAID, Exhaustive CHAID, CRT, and QUEST. We ran all of the aforementioned methods, and found that CHAID was the most successful among its counterparts.

7. 5 CHAID and Exhaustive CHAID

CHAID stands for CHi-squared Automatic Interaction Detector [16]. It uses the chi-square test to determine the next best split at each step. The three main steps involved are merging, splitting and stopping. However, first all of the continuous variables are converted to ordinal categorical variables by converting the continuous distribution into a finite number of categories Afterwards, the predictor categories are analyzed and if their tests are not statistically significant, the categories are merged This step is repeated with all of the predictors. This is followed by finding the most efficient way to split a set of cases into two child nodes based on their p-values. The tree stops expanding further when all the customers belong to the same class or when everyone

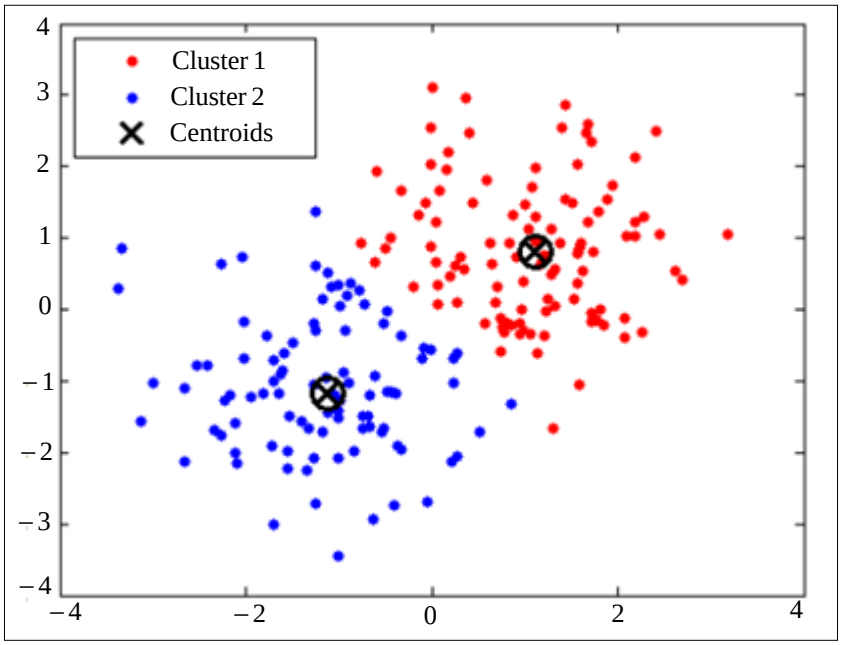

Figure 8. K-means Clustering with $K = 2$

hasgot similar attribute values. Exhaustive CHAID is a variant of CHAID, where the algorithm performs a more thorough merging and testing of predictors for similar pairs until only one pair remains. Therefore, it takes much more computing time. Figure 9 shows the result of SPSS for running decision trees with CHAID.

```
Clustered Instances

0      281001 (81%)
1       63930 (19%)
```

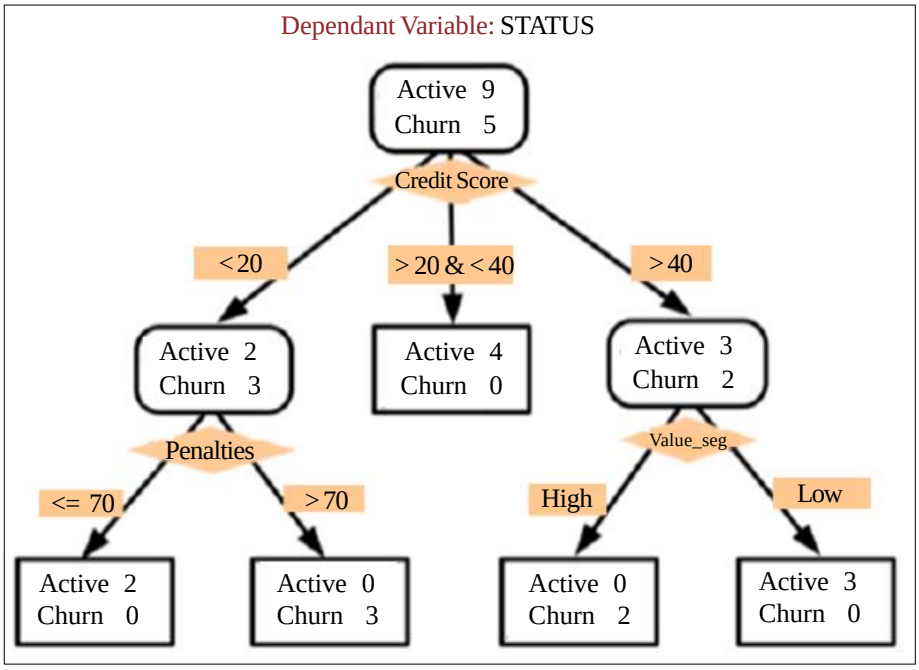

Figure 9. Decision Tree Example

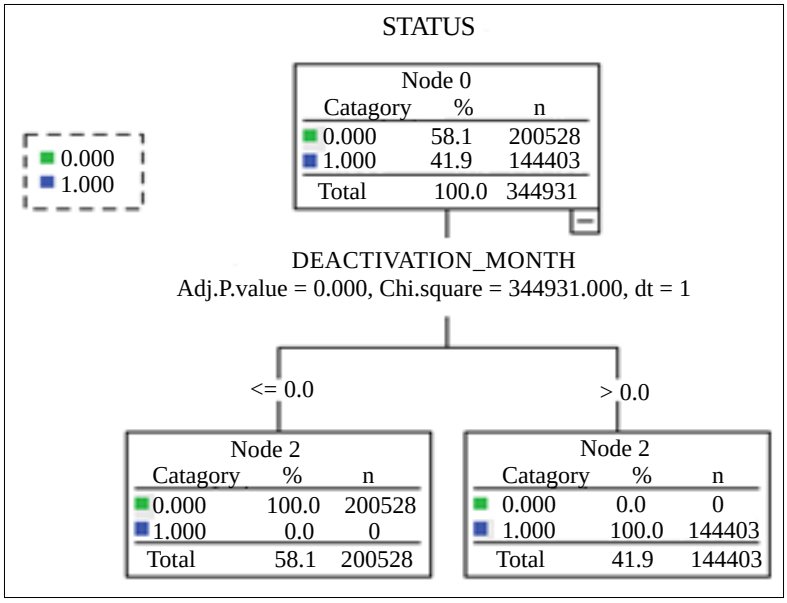

Figure 10. Result of Decision tree Using CHAID

	Recall	F-Measure
Active	0.792	0.743
Churn	0.531	0.583

Table 9. CHAID Results

	Recall	F-Measure
Active	0.773	0.750
Churn	0.603	0.628

Table 10. Exhaustive CHAID Results

7.5.1 CART

CART stands for *Classification And Regression Trees*. This method is more suitable for data supporting continuous dependent variable and categorical predictor variable [17]. In this method, the feature space is recursively split into non-overlapping regions. A classification tree is generated to predict the value of the dependent categorical variable. Moreover,regression trees are used to set conditions on variable valuesin order to predict the outcome of continuous dependentvariable [17].

	Recall	F-Measure
Active	0.897	0.740
Churn	0.271	0.383

Table 11. Results with CART

7.5.2 QUEST

QUEST (Quick, Unbiased and Efficient Statistical Tree) hasbeen known for its unbiased feature selection and handling ofcategorical variables with several categories. It uses ANNOVAF-statistical tests to choose the variable so as to split the node.The variable with the highest F-statistic is chosen first [18]. From the results shown in Table 7-10, we can observethat the decision trees in general and Exhaustive CHAID inparticular proved to be the most successful algorithm for churn prediction. As could be observed from Table 10, not only theoverall accuracy achieved while training the data, was the highest. But also the percentage percentage of correctly identified churnerswas the highest i.e. 60%. On the other hand, other decisiontrees variants supported by SPSS did not perform as well asExhaustive CHAID.

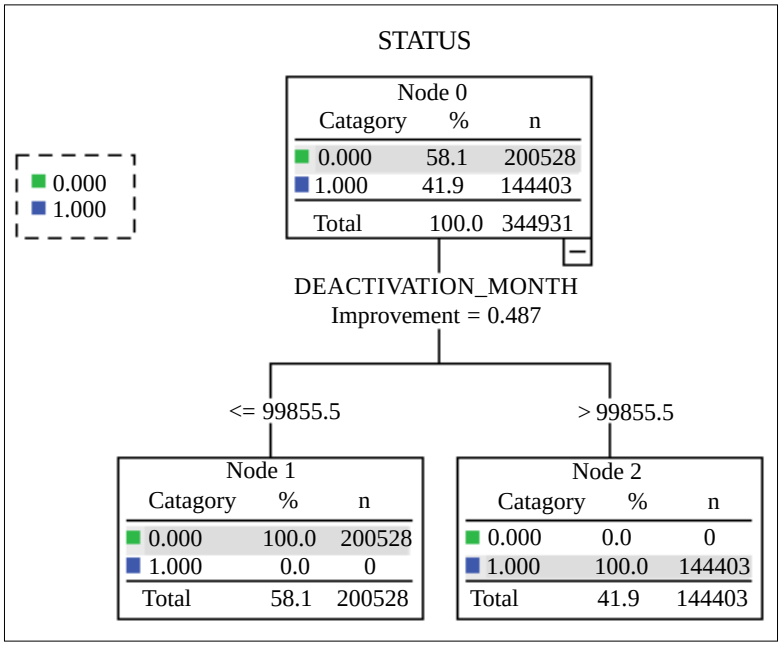

Figure 11. SPSS Result for CART

	Recall	F-Measure
Active	0.821	0.727
Churn	0.393	0.478

Table 12. Quest Result

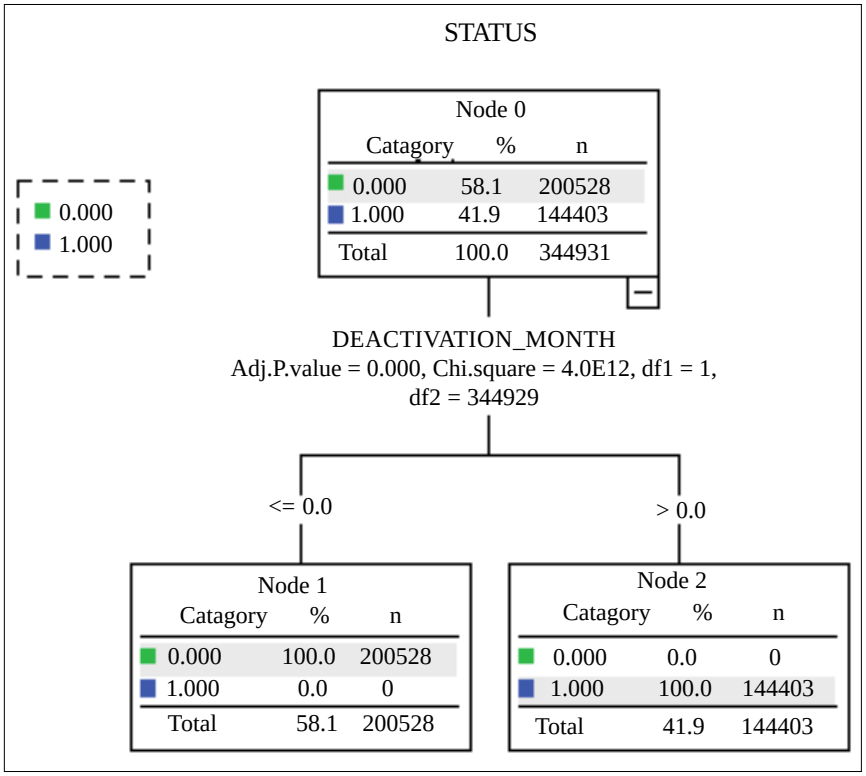

Table 12. SPSS Result for Quest

8. Improving Results Using Derived Variables

After analyzing the results of all algorithms, we found that *Exhaustive CHAID* was the most accuratevariant of decision trees for our data. So far the accuracyachieved was 70%. In order to build upon that result, wedecided to introduce some variables of our own in thedata set and see if this could further boost our accuracy.Five new variables were added to our data set [19]. Theywere derived from some of the existing variables. For example,one of the derived variable,*Duration_Per_Fixed_Out*was calculated by dividing the total outgoing call timeto fixed lines by the number of calls made to the fixedlines. The remaining four variables were also calculatedin a similar fashion. They were *Duration_Per_OnNet_Out*,*Duration_Per_Fixed_Inc*, D*uration_Per_OnNet_Inc*and *Duration_Per_Inter_Inc*.

	Recall	F-Measure
Active	0.769	0.770
Churn	0.685	0.682

Table 13. Modified Result based on Derived Variables

	Recall	F-Measure
Active	0.763	0.853
Churn	0.605	0.223

Table 14. Testing Results

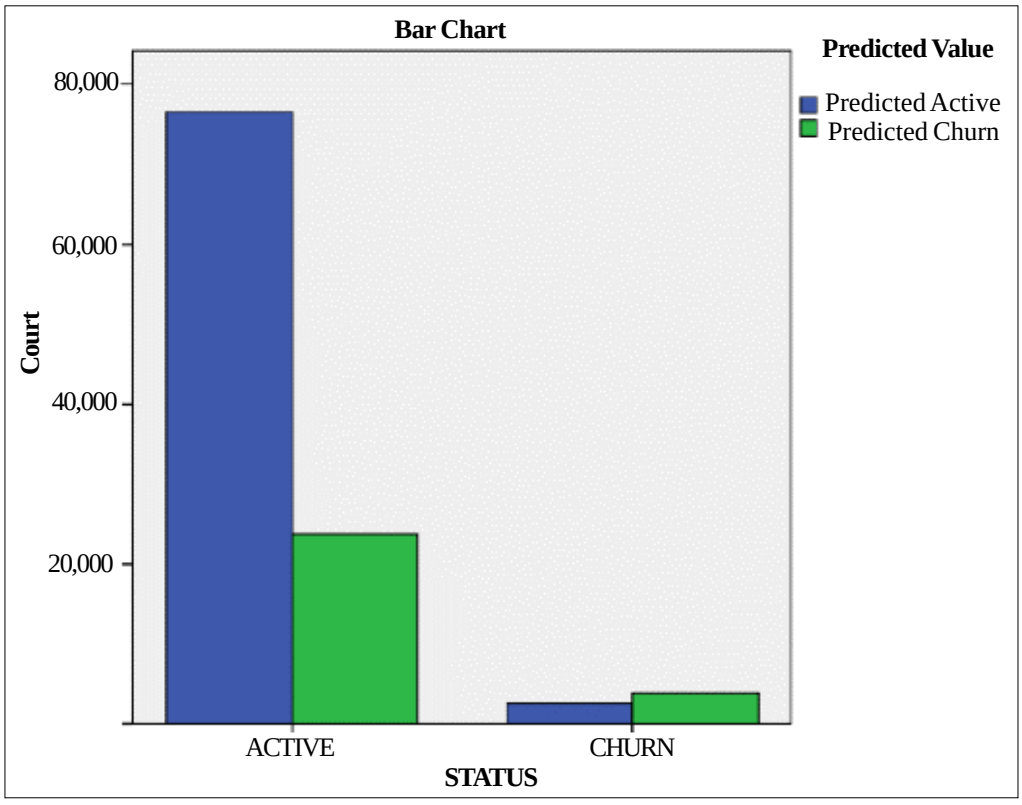

Table 13. Exhaustive CHAID Testing Results with Derived Variables

After including the derived variables in our analysis, the modified results are given in Table 13. One can observe that the value for recall for active users increased to 76.9%. More importantly, the recall for churners rose by a considerable margin of approximately 8.5% from the earlier best result to 68.5%.

Testing is used to verify the predictive relationship obtained in the training phase. The data was separated into training and

testing sets with a 70:30 ratio. In the case of the given data set, the data for the first two months were used for training while that for the third month was used for testing. In the testing phase, we achieved comparative results as well. The bar chartin Figure 13 shows the results for the test phase. It could be observedthat the recall for churners was 60.5%, whereas the recall foractive customers was 76.3%. The overall accuracy in this casewas 75.4%.

9. Conclusion and Future Work

In this paper, we applied different machine learning algorithms such as linear and logistic regression, Artificial Neural Networks, K-Means clustering, decision trees including CHAID, Exhaustive CHAID, CART and QUEST in order to classify churners and active customers. The data set contained telecommunication traffic data of 106,000 customers along with their usage behavior for 3 months. The results were compared based on the values of precision, recall and F-measure. We successfully resolved the problem of class imbalance. The best results were obtained with Exhaustive CHAID algorithm, a variant of the standard decision trees algorithm. In the future, we plan to test our approach on bigger data sets containing data over a longer period of time. Moreover, we plan to work on diverse data belonging to different countries and different telecommunication companies in the near future.

References

[1] Lazarov, V., Capota, M. (2007). Churn prediction, TechnischeUniversitatMunchen. Eighth ACM SIGKDD International Conference.

[2] Mutanen, T., Nousiainen, S., Ahola, J. (2010). Customer churn prediction –A case study in Retail Banking,Conference on Data *Mining forBusiness Applications*, Amsterdam, Netherlands.

[3] Umayaparvathi, V., Iyakutti, K. (2012). Applications of data mining techniquesin telecom churn prediction, *International Journal of Computer Applications*.

[4] Jadhav, R. J., Pawar, U. T. (2011). Churn Prediction in Telecommunication Using Data Mining Technology, *International Journal of Advanced Computer Science and Applications* (IJACSA), 2, February.

[5] Chandrasekhar, S. Predicting the Churn in Telecom Industry.

[6] Shaaban, E., Helmy, Y., Khedr, A., Nasr, M. (2012). A Proposed Churn Prediction Model, *International Journal of Engineering Research and Applications* (IJERA), 2 (4).

[7] Davis, J., Goadrich, M. (2006). The Relationship between Precision-Recalland ROC curves, *In*: Proceedings of the 23rd International Conference on Machine Learning (ICML).

[8] Davis, J., Goadrich, M . (2011). Evaluation: From Precision, Recall and Fmeasureto ROC, informedness, markedness and correlation, *Journal ofMachine Learning Technologies*, 2 (1).

[9] Drummond, C., Holtel, R. C. (2005). Severe class imbalance: Why betteralgorithms are not the answer, 16th European Conference *of Machine Learning* (ECML).

[10] Garcýa, V. , Sanchez, J. S . Mollineda, R. A . Alejo, R . Sotoca, J. M. (2007). The class imbalance problem in pattern classification and learning. In F. J. Ferrer-Troyano et al, editor, II CongresoEspanol de Informatica, p. 283–291, Zaragoza,. Thomson.

[11] Liu, X. -Y., Wu, J., Zhou, Z. -H. (2006). Exploratory under-sampling for classimbalancelearning. *In*: International Conference on Data Mining (ICDM), p. 965–969. IEEE Computer Society.

[12] Stahlbock, R., Crone, S. F., Lessmann. S. (2011). Data mining: Special issuein annals of Information Systems, chapter 8,

[13] Nguyen, D., Smith, N. A., Rose, C. P. (2011). Author age prediction fromtext using linear regression. Optimum Learning Rate for ClassificationProblem with MLP in Data MiningLaTeCH '11 Proceedings of the 5thACL-HLT Workshop on Language Technology for Cultural Heritage,Social Sciences, and Humanities.

[14] Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan, I., Keane. J. (2009). Comparingdata mining methods with logistic regression in childhood obesityprediction. Information Systems Frontiers, 11(4) 449–460.

[15] Rokach, L., Maimon, O. (2008). Data mining with decision trees: theory andapplications. World Scientific Pub Co Inc.

[16] Kass, G. V. (1980). An exploratory technique for investigating large quantitiesof categorical data. 32nd Annual Conference of the Gesellschaft Fur.

[17] Breiman, L. Friedman, J.. Olshen, R Stone, C. (1984). Classification andRegression Trees. Wadsworth and Brooks, Monterey, CA.

[18] Loh, W. Y., Shih, Y. S. (1997). Split Selection Methods for Classification Trees. Statistica Sinica.

[19] Alberts, L. J. S. M. (2006). Churn prediction in the mobile telecommunicationsindustry, Chapter 2. Master's thesis, Maastricht University.