Telecommunication Subscribers' Churn Prediction Model Using Machine Learning

Saad Ahmed Qureshi, Ammar Saleem Rehman, Ali Mustafa Qamar, Aatif Kamal
Department of Computing
School of Electrical Engineering and Computer Science
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
{*09bicseaqureshi, 09bicsearehman, mustafa.qamar, aatif.kamal*}*@seecs.edu.pk*

Ahsan Rehman
Business Analytic Consultant
IBM - Global Business Services
ahsanr@pk.ibm.com

***Abstract*—During the last two decades, we have seen mobile communication becoming the 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. At the same time, public policies and standardization of mobile communication now allow customers to easily switch over from one carrier to another, resulting in a fluid market. Since the cost of winning a new customer is far greater than the cost of retaining an existing one, mobile carriers have now shifted their focus from customer acquisition to customer retention. As a result, churn prediction has emerged as the most crucial Business Intelligence (BI) application that aims at identifying customers who are about to transfer their business to a competitor i.e. to churn. This paper aims to 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. We also discuss the use of re-sampling method in order to solve the problem of class imbalance. Our results show that in case of the data set used, decision trees is the most accurate classifier algorithm while identifying potential churners.**

***Keywords*-Churn prediction, Business Intelligence, Data Mining**

I. 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 also impact 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 telecommunications companies are investing include customer retention, target marketing, campaign management, and 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.

II. 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*) gave the best results as compared to other conventional algorithms. Furthermore, they argued that a good prediction model has to be constantly updated and should use a combination of different data mining techniques. In another case study [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 models each of 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 churner to non churner ratio of 2:3, based on data over 8 months got the best results during the testing phase. In that study, the authors concluded that due to the dynamic nature of the customer, the logistic regression models had 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 trees surpassed the former in terms of accuracy. They divided their project into five phases: Data Acquisition, Data preparation, Derived variables, Extracting Variables, and Model construction.

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 differ according to the characteristics of the data. In this study, we tested the conventional algorithms on our data set. In addition to that, we have also suggested some methods to deal with a

978-1-4799-0615-4/13/$31.00 ©2013 IEEE

TABLE I
CONFUSION MATRIX (CHURN PREDICTION)

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

very commonly occurring problem in the telecommunication industry, known as the class imbalance problem. The problem along with its possible solutions are discussed in Section V.

III. DATA ACQUISITION

The data set used in this study was acquired from an online source [1]. 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), traffic destination (on-net, competition), rate plan, loyalty, traffic behavior etc. are some of the main attributes of this data set. 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.). *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.

IV. EVALUATION 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 I shows the confusion matrix in order to calculate these evaluation measures.

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

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

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

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

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

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

TABLE II
SPEARMAN'S CORRELATION ANALYSIS

Variable	Correlation Coefficient
Credit Score	0.731
No. of penalties 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

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

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

V. IDENTIFYING IMPORTANT PREDICTORS

Before training a model with the conventional machine learning algorithms, one of the essential steps is to select the right group of variables 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. 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 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 best 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 to identify 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 II.

VI. 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 we faced 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 [6]. For example in case of our data, there were 100,264 active users, whereas there were only 6231 churners. In other words, there were 94.1% active users vs only 5.9% churners, presenting a typical case of class imbalance. As a result, whenever we ran algorithms such as logistic regression, decision tree, *ANN*; all of the predictions made were in favor of the majority class (active class in this case).

[1] http://www.customers-dna.com/

TABLE III
CASES PREDICTED CHURN

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

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: we can either over-sample or under-sample [7]. In under-sampling, we use only a subset of the majority class in order to train our data [8] . In the present case, we removed a random selection of entries from the set of active customers, 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 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 [9].

The results in Table III 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.

VII. Application of Machine Learning algorithms

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

A. *Regression Analysis:*

A brief overview of linear regression as well as logistic regression is next presented.

Linear Regression: Linear regression analysis is used 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].

$$Y = B_0 + B_1 X_1 + B_2 X_+ \cdots \tag{6}$$

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 all the factors that effect 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 impact of each predictor and the importance it has in predicting the output variable.

Fig. 1 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 the figure is showing 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.

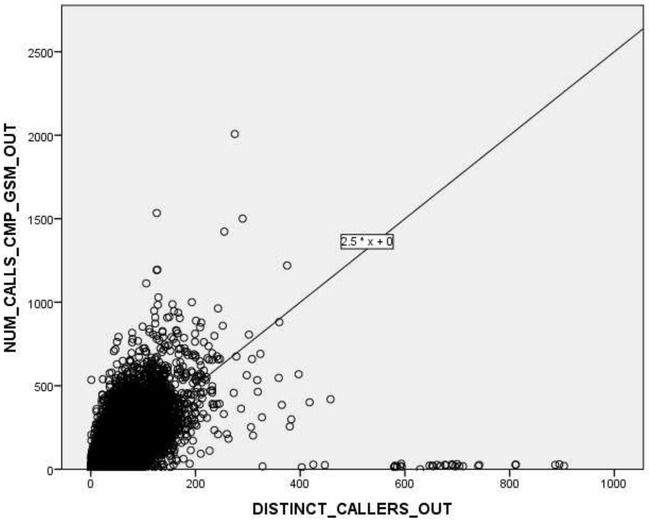

Fig. 1. Simple linear regression

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 seemed to be the best choice. In the given data set, the status of the subscriber is the 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.

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

where $P(Y|x)$ 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*. Fig. 2 shows a logistic function graph.

The results in Table IV 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% of the active users were correctly

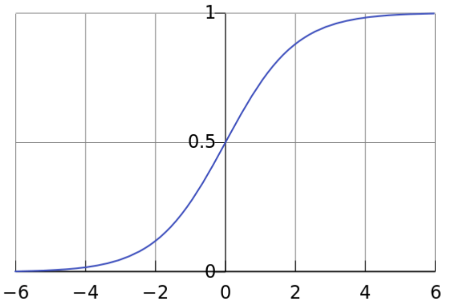

Fig. 2. Logistic Curve

TABLE IV
LOGISTIC REGRESSION RESULTS

	Recall	F-Measure
Active	0.787	0.720
Churn	0.450	0.515

identified. Conversely, only 45% of the total churners were identified, which is too low. One could also note here that although the algorithm did a good job in identifying the active cases, it failed to perform well while identifying the churn cases.

B. Artificial Neural Networks (ANNs)

We also implemented a feed-forward *ANN* also known as multilayer 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. The results obtained using neural networks on our data are described in Table V [11]. 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 is still way behind and is even lesser than the one observed with regression.

TABLE V
NEURAL NETWORKS RESULTS

	Recall	F-Measure
Active	0.823	0.698
Churn	0.325	0.419

TABLE VI
K MEANS CLUSTERING RESULTS

	Recall	F-Measure
Active	0.503	0.529
Churn	0.445	0.416

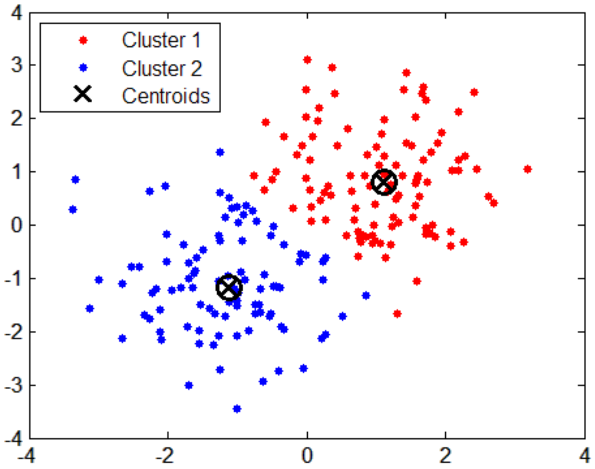

Fig. 3. K means clustering with K=2

C. K-Means Clustering

K-means algorithm is the most widely used clustering algorithm. The K in the name means that the algorithm will be looking to classify the data into K different clusters. The value of K has to be specified before the algorithm starts executing. Consequently, the first step in order to use this algorithm is to identify the K 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) or similarity (e.g. cosine similarity). In order to optimize the clusters, the algorithm goes through a number of iterations. In each iteration, the centroid of each cluster is recalculated and in the next iteration those points are assigned as seeds. Table VI 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.

D. Decision Trees

In decision trees, we have classification or regression models in the 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 customers value for that attribute, it will branch out to further nodes until the final leaf node is reached, which will either be a *churn* node or an *active* node [12].

SPSS statistics software provides a possibility to use one of the four variations of decision trees, namely: *CHAID*,

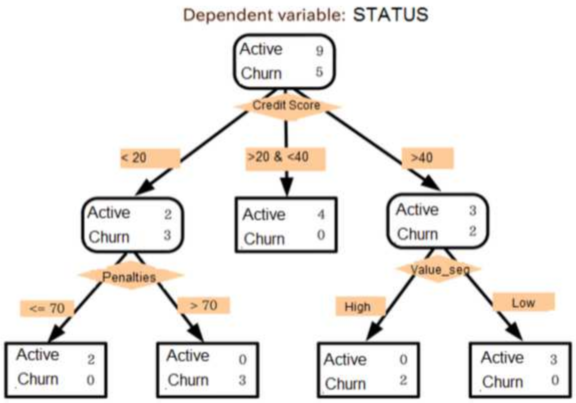

Fig. 4. An example of a decision tree

Exhaustive CHAID, *CRT*, and *QUEST*. We ran all of the aforementioned methods, and found that *CHAID* was the most successful among its counterparts.

CHAID and Exhaustive CHAID

CHAID stands for *CHi-squared Automatic Interaction Detector* [13]. 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. Then the predictor categories are analyzed and if their tests are not statistically significant, the categories are merged and 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 records belong to the same class or when all the records have 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.

CART

CART stands for *Classification And Regression Trees*. This method is more suitable for data supporting continuous dependent variable and categorical predictor variable [14]. In this method, the feature space is recursively split into non-overlapping regions. A classification tree is generated to predict the value of dependent categorical variable. Moreover, regression trees are used to set conditions on variable values in order to predict the outcome of continuous dependent variable [14].

TABLE VII
CHAID RESULTS

	Recall	F-Measure
Active	0.792	0.743
Churn	0.531	0.583

TABLE VIII
CART RESULTS

	Recall	F-Measure
Active	0.897	0.740
Churn	0.271	0.383

TABLE IX
QUEST RESULTS

	Recall	F-Measure
Active	0.821	0.727
Churn	0.393	0.478

TABLE X
EXHAUSTIVE CHAID RESULTS

	Recall	F-Measure
Active	0.773	0.750
Churn	0.603	0.628

QUEST

QUEST (Quick, Unbiased and Efficient Statistical Tree) has been known for its unbiased feature selection and handling of categorical variables with several categories. It uses *ANNOVA F-statistical* tests to choose the variable so as to split the node. The variable with the highest *F-statistic* is chosen first [15].

From the results shown in Table VII-X, we can observe that the decision trees in general and *Exhaustive CHAID* in particular proved to be the most successful algorithm for churn prediction. As could be observed from Table X, not only the overall accuracy achieved while training the data, was the highest. But also the percentage of correctly identified churners was the highest i.e. 60%. On the other hand, other decision trees variants supported by *SPSS* did not perform as well as *Exhaustive CHAID*.

VIII. IMPROVING RESULTS USING DERIVED VARIABLES

After analyzing the results of all algorithms, we reached the conclusion that *Exhaustive CHAID* was the most accurate variant of decision trees for our data. So far the accuracy achieved was 70%. In order to build upon that result, we decided to introduce some variables of our own in the data set and see if this could further boost our accuracy. Five new variables were added to our data set [16]. They were 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 time to fixed lines by the number of calls made to the fixed lines. The remaining four variables were also calculated in a similar fashion. They were *Duration_Per_OnNet_Out, Duration_Per_Fixed_Inc, Duration_Per_OnNet_Inc* and *Duration_Per_Inter_Inc*.

TABLE XI
MODIFIED RESULTS BASED ON DERIVED VARIABLES

	Recall	F-Measure
Active	0.769	0.770
Churn	0.685	0.682

TABLE XII
TESTING RESULTS

	Recall	F-Measure
Active	0.763	0.853
Churn	0.605	0.223

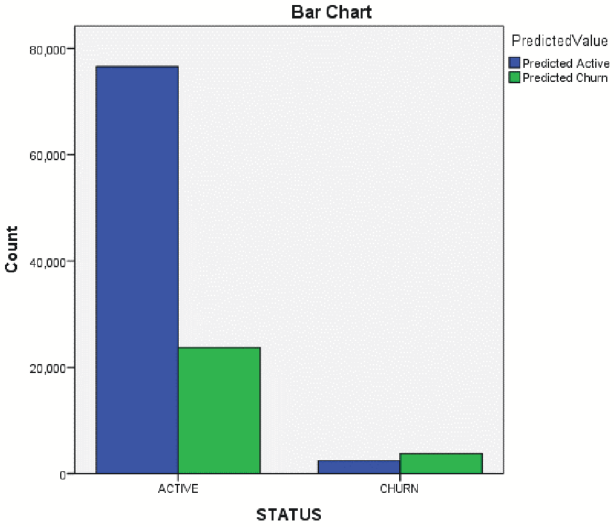

Fig. 5. Exhaustive CHAID Testing Results with Derived Variables

After including the derived variables in our analysis, the modified results are given in Table XI. 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 chart in Fig. 5 show the results for test phase. It could be observed that the recall for churners was 60.5%, whereas the recall for active customers was 76.3%. The overall accuracy in this case was 75.4%.

IX. 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] V. Lazarov and M. Capota. Churn prediction, Technische Universität München. *Eighth ACM SIGKDD International Conference*, 2007.

[2] T. Mutanen, S. Nousiainen, and J. Ahola. Customer churn prediction – A case study in Retail Banking. *2010 conference on Data Mining for Business Applications, Amsterdam, Netherlands*, 2010.

[3] V. Umayaparvathi and K. Iyakutti. Applications of data mining techniques in telecom churn prediction. *International Journal of Computer Applications*, 2012.

[4] J. Davis and M. Goadrich. The Relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[5] J. Davis and M. Goadrich. Evaluation: From Precision, Recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies,Volume 2, Issue 1*, 2011.

[6] C. Drummond and R. C. Holtel. Severe class imbalance: Why better algorithms are not the answer. *16th European Conference of Machine Learning*, 2005.

[7] V. García, J. S. Sánchez, R. A. Mollineda, R. Alejo, and J. M. Sotoca. The class imbalance problem in pattern classification and learning. In F. J. Ferrer-Troyano et al, editor, *II Congreso Español de Informática*, pages 283–291, Zaragoza, 2007. Thomson.

[8] X.-Y. Liu, J. Wu, and Z.-H. Zhou. Exploratory under-sampling for class-imbalance learning. In *ICDM*, pages 965–969. IEEE Computer Society, 2006.

[9] R. Stahlbock, S. F. Crone, and S. Lessmann. Data mining: Special issue in annals of information systems, chapter 8, 2009.

[10] D. Nguyen, N. A. Smith, and C. P. Rose. Author age prediction from text using linear regression. *Optimum Learning Rate for Classification Problem with MLP in Data MiningLaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 2011.

[11] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan, and J. Keane. Comparing data mining methods with logistic regression in childhood obesity prediction. *Information Systems Frontiers*, 11(4):449–460, September 2009.

[12] L. Rokach and O. Maimon. Data mining with decision trees: theory and applications. *World Scientific Pub Co Inc*, 2008.

[13] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *32nd Annual Conference of the Gesellschaft Fur*, 1980.

[14] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[15] W. Y. Loh and Y. S. Shih. Split Selection Methods for Classification Trees. *Statistica Sinica*, 1997.

[16] L. J. S. M. Alberts. Churn prediction in the mobile telecommunications industry, chapter 2. Master's thesis, Maastricht University, 2006.