Handling Imbalanced Data in Customer Churn Prediction Using Combined Sampling and Weighted Random Forest

Veronikha Effendy
Telkom University
Bandung, Indonesia
veffendy@telkomuniversity.ac.id

Adiwijaya
Telkom University
Bandung, Indonesia
adiwijaya@ telkomuniversity.ac.id

Z.K.A. Baizal
Telkom University
Bandung, Indonesia
baizal@ telkomuniversity.ac.id

Abstract — **Customer churn is a major problem that is found in the telecommunications industry because it affects the company's revenue. At the time of the customer churn is taking place, the percentage of data that describes the customer churn is usually low. Unfortunately, the churn data is the data which have to be predicted earlier. The lack of data on customer churn led to the problem of imbalanced data. The imbalanced data caused difficulties in developing a good prediction model. This research applied a combination of sampling techniques and Weighted Random Forest (WRF) to improve the customer churn prediction model on a sample dataset from a telecommunication industry in Indonesia. WRF claimed can produce a prediction model which has a good performance on the imbalanced data problem. However, this research found that the performance of the prediction model developed by WRF using the dataset is still quite low. Sampling techniques were applied to overcome this problem. This research used the combination of simple under sampling and SMOTE. The result shown that the combined-sampling and WRF could produce a prediction model which had better performance than before.**

Keywords : Churn, Prediction, Weighted Random Forest, Combined-sampling, simple under sampling, SMOTE

I. INTRODUCTION

Nowadays, telecommunication industries have a big problem, that is customer churn, which means, they will suffer from customers and get the impact on the company's revenue [1]. To survive and win the market competition, some companies start trying to predict customer churn with data mining approach [2].

Data mining approaches can help a company for better understanding customer behavior from its own data, so that the company can implement the right CRM (Customer Relationship Management) strategies in order to save its revenue [2]. Unfortunately, churn is rare objects, but of great interest and great value in a company [3], so that we have only few churned customers in overall data. In the other word, we have an extreme imbalanced data set in this case.

There are two common approaches in handling imbalanced data. First is sampling approach and second is cost-sensitive approach [6].

The objective of this research is to handle the imbalanced data on the dataset used in this research, so that it can produce better performance in churn prediction model. The dataset was pre-processed by combining the sampling technique and WRF classifier. The dataset used in this research is a churn dataset of a telecommunication industry in Indonesia, which has 0.7% churn data. WRF implementation on this dataset still produce low performance. It was then solved by applying sampling techniques that are intended to reduce the imbalance between the two classes (churn and non-churn). SMOTE was used to generate the synthetics data from the churn class in order to increase the probability of drawing the churn data (churn is the minority data) [4], [7]. The addition of this data was certainly cause the direct impact on the computation time. To address this problem, a simple undersampling technique was also applied [5], [9].

II. RELATED WORK

Researchers have attempted to find methods to handle the imbalanced data in churn prediction. Some of them focuses on the data pre-processing, i.e. oversampling [4], and the other try to find the match classifier for this kind of problem, such as : logistic regression, linear classification, naïve Bayes, decision tree, multilayer perceptron neural networks, support vector machine, data mining evolutionary algorithm [1], and random forest [5], [6]. Some of those researches have resulted a good performance in churn prediction for their own data set, however every data set have their own characteristic and specific case [2].

There are two common approaches in handling imbalanced data. First is sampling approach and second is cost-sensitive approach [6]. There are continuous researches in improving prediction performance for handling imbalanced data using random forest, such as balanced random forest, weighted random forest [6], improve balanced random forest [1], weight random forest with under sampling [5], etc.

Weight random forest classifier claimed can handle the imbalanced data problem with cost-sensitive approach, that is to assign weight, so that it can reduce misclassified data [6] .

Sampling basic techniques are under sampling and oversampling. Each technique has its own benefits and

978-1-4799-3580-2/14/$31.00 ©2014 IEEE

drawbacks. Under sampling makes the model run faster, but this technique causes big loss of potential data from the majority class in imbalanced data, so that it reduces prediction performance. Oversampling create additional data (but not additional information), this causes slower running process. Although oversampling does not reduce the data record, but the additional data from the minority class causes over fitting (there are any possibilities that sampling makes any data in majority class moves to minority class) [4].

The data set used in this research is customer behavior profile data. It has been studied before in Indonesia. One of the results shows that SMOTE (Synthetic Minority Over-Sampling) algorithm has a good performance [4], however, the prediction measurement is not accurate. This research try to predict whether customers potentially churn or not-churn based on customer behavior profile with a reasonable accuracy. The basic idea in SMOTE is to create data synthetic from minority class [7].

The combination of simple under sampling and SMOTE algorithm may reduce the substantial loss of potential data from majority class and also reduce the probability of over fitting problem caused by the oversampling by SMOTE.

For those reasons, this research attempts to combine the two approaches (sampling and cost sensitive-learning) by applying dataset processed in combine sampling method (SMOTE for minority data and simple under sampling for the majority data) to the Weight Random Forest classifier [7], [6], [8].

III. MATERIALS AND PROPOSED METHOD

A. Proposed Method

We propose the combination of classification algoritma using WRF and combined-sampling. Based on the reason in the previous section, the combined-sampling was using simple undersampling and SMOTE.

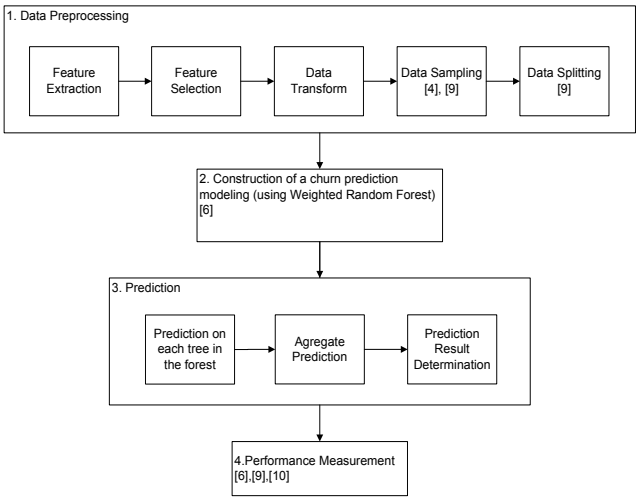

Figure 1. Block diagram of research design.

This study was divided into four main stages, (1) data preprocessing stage; (2) constructing prediction model stage; (3) prediction on data test (testing stage), and (4) performance measurement. Figure 1 shows the block diagram of the research design.

1) Data Preprocessing Stage

Data preprocessing stage is a preparation-stage before the data are processed. At this stage, first, attributes are selected. Attributes which has 80% of missing value and have redundant information was not use in the research. Then the data is converted into numerical type without changing the meaning of the data. This was done so that the data is readable by the system.

The last process in this stage is to generate a variety of datasets by using oversampling, undersampling, and a combination of both. The different datasets will become the input data for the system as part of the experimental scenario.

2) Constructing Prediction Model Stage

At this stage, the method used WRF (with input parameters: m, ntree and weight) to build a collection of many decision tree classifier which is formed from the bootstrap training data and the number of attributes m randomly selected by the system. Root node and splitting tree is determined by calculating the weighted Gini, while labeling at each node is determined by calculating the weighted majority vote based on pre-determined weight value on the input parameters. The growth of the decision tree is done repeatedly according ntree desired amount. Average weighted majority vote (i.e. number of case multiplied by the specified weight of the input parameters on each class) is then calculated based on the predicted outcome (i.e. labels and and weighted majority vote) of every tree that has been formed. Label the final prediction result is then determined based on the highest avergae of weighted majority vote.

3) Testing Stage

This testing stages is a stage where the formed model is tested using prepared data test. This stage used 10-fold cross validation in order to obtain reliable performance values. The division of the dataset into 10 parts performed by the system using a stratified sampling technique. This technique begins by dividing the minority data into 10 parts and majority data into 10 parts, then taking 1 part of minority data and 1 part of the major. This parts of data then being used as test data and the rest were used as training data. This technique is used so that the data input for WRF method has the same percentage of churn with the original dataset.

4) Performance Measurement

The performance of the prediction result on each scenario is then calculated by the system. The performance

was measured in F-measure, weighted accuracy and top-decile.

B. Data

The data input for the prediction system was a churn dataset with 24 attributes and 48,384 records. All the attributes was in categorical type.

IV. EXPERIMENT

A. Evaluation Criteria

As mention in the previous section, we use F-measure and Top-decile as performance evaluation measures. The calculation of F-measure need the value of precision and recall. The precision and recall will be calculated based on the confusion matrix as shown in TABLE I. Here, we take the minority class as positive class. The F-measure defined as (3), while the precision and recall defined as (1) and (2).

TABLE I. Confusion Matrix for Two-Class

COUNT		Actual Class	
		1	*0*
Prediction Class	*1*	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

$$\text{precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

$$\lambda = \frac{\beta}{\beta_0} \quad (4)$$

Another measurement we use is top-decile This is the percentage of the 10% of customers predicted to be most likely to churn who actually churned, divided by the baseline churn rate. The higher the top-decile, the more accurate the model is, and intuitively, the more profitable a targeted proactive churn management program. Top-decile was used in churn prediction due to the restriction of budget in a company. Top-decile defined as (4). As shown in (1) , indicates the percentage of churner on the 10% riskiest segment, while β0 indicates the percentage of churner on the whole customers (the actual), and indicates the top-decile [10].

B. Experiment Scenario

The whole scenario was run and tested using 10-fold cross-validation method to evaluate the performance of the churn prediction model. To obtain the best performance churn prediction model, the scenario was selected based on the value of the best F-measure. Based on the explanation in previous research by Breimenn [9], the best number of selected attributes used in the process of tree establishment is around int($\log_2$ M + 1) where M is the total number of attributes in the dataset. Since the dataset has 24 attributes, then the number of selected attributes which were used in the process of establishment tree is around 5.

The best input parameters and the best datasets will be selected by the best F-measure of the churn prediction in the experiments. While at the end of the experiment, the top-decile was calculated from the results of some classifier models. As described in the previous chapter, the top-decile was used to see how many churners included in 10% of riskiest-customer.

This research had three scenarios :

1) Scenario-1

The goal of the scenario was to find the best input parameter to produce the best performance result using WRF and original data (without sampling technique). This scenario involved multiple experiments involving variations WRF parameter values, i.e.: 1) weight, and 2) number of trees in the forest.

2) Scenario-2

The goal of scenario 2 was to find the best dataset which had the best performance result using sampling technique combined with WRF with standard input parameter. The scenario used the specified parameters: 1) the value of m = 5, referring to the number of attributes selected in the Random Forest algorithm, 2) the weight value = 0.9, referring of best weight of scenario-1, and 3) the value of ntree = 5, referring to the number of trees generated in the forest. This scenario involved multiple experiments involving variations of dataset. The original dataset was modified with sampling techniques to produce various dataset.

3) Scenario-3

The goal of the scenario was to find best performance result using WRF from few best datasets from the scenario-2 (WRF with sampling technique). The experiments were conducted for each of these datasets. Each experiment was used some variation of parameters number of trees (ntree) and weight. From the results of those experiments with different variations of these parameters, it was seen how the influence of the parameters weight and ntree to the performance of the classifier.

C. Experiment Result

In the presenting the results of experiments, *n* indicates the number of tree which is built on modeling, while *m* indicates the number of attributes that were randomly selected to form a decision tree. The following are the results of each experiment scenario:

1) Experiment Result of Scenario-1

TABLE II shows that an increase in weight value of 0.9 to 0.999 resulted in an increase in F-measure of the predictive models generated, but the F-measure values decreased when the weight is worth 0.995. This fact shows that the weight value 0.99 is the best weight parameters of the experiments carried out for the prediction model established by the WRF to the original data. The weight parameter values then become the fixed input parameters for the next part, where the results are shown in TABLE III.

TABLE II. THE PERFORMANCE RESULT OF SCENARIO 1.A

Data File	ntree	Weight of churn class	m	F-measure
Data telco	5	0.9	5	0
Data telco	5	0.99	5	**0.027273**
Data telco	5	0.999	5	0.015472
Data telco	5	0.995	5	0.018565

TABLE III. THE PERFORMANCE RESULT OF SCENARIO 1.B

Data File	ntree	Weight of churn class	m	F-measure
Data telco	5	0.99	5	**0.027273**
Data telco	10	0.99	5	0.025852
Data telco	50	0.99	5	0.01534

TABLE III is the result of the performance prediction model using the parameter values weight = 0.99 and several variations values of parameter ntree (number of tree constructed). TABLE III shows that the best performance of the experiments conducted, obtained when ntree worth 5. The best F-measure can be obtained in this scenario is 0.027273.

2) Experiment Result of Scenario-2

TABLE IV shows the performance of the prediction results of dataset produced by original data, oversampling and undersampling using ntree=5 and m=5. TABLE IV shows some dataset which were processed by the under sampling produced very poor performance when the weight value to the value of 0.9. Therefore, the adjustment was done on the weight, and thus produced a better performance as shown in the TABLE IV.

It also seen that the implementation of the under sampling effect on the value of the F-measure. The improvement on the F-measure is not significant, almost the same. This was happen because the churn rate changes resulting from the under sampling is not very significant, so the problem of lack minority data was still exists there.

The oversampling also made significant change to the F-measure. The improvement of F-measure is quite significant. It happened because the problem of lack minority data can be reduced, and it can be seen by the change of churn rate shown in the TABLE IV.

The TABLE V shows that the combination of oversampling and undersampling gave significant improvement on the F-measure. The result was better than previous experiment part in scenario 2.

TABLE IV. THE PERFROMANCE RESULT OF SCENARIO 2 PART 1

Data File	RecNum	Churn pctg(%)	Weight of churn class	F-measure
Data telco	48384	0.77	0.9	0
Data telco	48384	0.77	0.99	0.027273
Data telco + SMOTE 5x	50258	4.47	0.9	0.19129
Data telco + SMOTE 10x	52133	7.91	0.9	**0.38923**
Data telco + SMOTE 50x	67133	28.48	0.9	**0.62203**
Data telco + US 8/10	38782	0.97	0.9	0
Data telco + US 8/10	38782	0.97	0.99	**0.02777**
Data telco + US ¾	36382	1.03	0.9	0
Data telco + US ¾	36382	1.03	0.99	**0.030539**
Data telco + US 9/10	43583	0.860427	0.9	0
Data telco + US 9/10	43583	0.860427	0.99	0.02992

TABLE V. THE PERFROMANCE RESULT OF SCENARIO 2 PART 2.

Data File	RecNum	Churn pctg(%)	Weight of churn class	F-measure
Data telco + US 8/10 + SMOTE 50x	57907	33.03	0.9	**0.66016**
Data telco + US ¾ + SMOTE 50x	55132	34.69	0.9	0.65763
Data telco + US 8/10 + SMOTE 10x	42532	9.70	0.8	**0.43729**
Data telco + US 8/10 + SMOTE 10x	42532	9.70	0.9	0.33363
Data telco + US ¾ + SMOTE 10x	40132	10.28	0.8	**0.50643**
Data telco + US ¾ + SMOTE 10x	40132	10.28	0.9	0.3784

3) Experiment Result of Scenario-3

This scenario used two datasets from the previous scenarios, i.e. data telco + US ¾ + SMOTE 10x and data telco + US 8/10 + SMOTE 50x. These datasets have been seen in the previous scenario has the F-measure improvement which is quite good when oversampling and under sampling were combined. Each of these dataset used in this scenario to make sure the influence of weight parameters and the amount of generated tree to the performance results of the prediction models.

Figure 2 shows that the value of F-measure began to decline when the weight of churn class above 0.8. Figure 3 shows that the value of F-measure began to decline when the weight of churn class above 0.6. This result shows that to produce a best F-measure, different weight was needed.

If the results compared to the original dataset with its best F-measure, the trend of the data looked like in Figure 4. Figure 4 shows the trend of a decrease in weight when there is an increase in the percentage of churn class.

Figure 5 and figure 6 show the trend of ntree change to F-measure on each dataset.

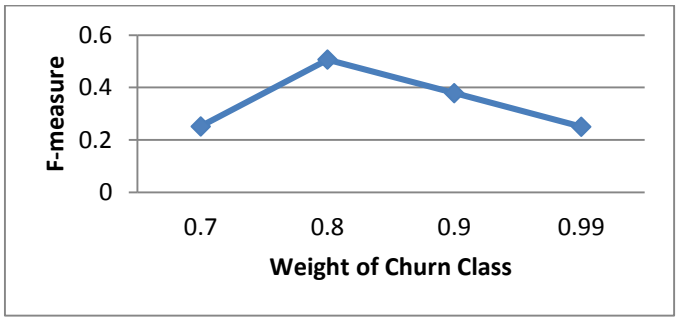

Figure 2. Trend of weight to F-measure (data telco + US ¾ + SMOTE 10x).

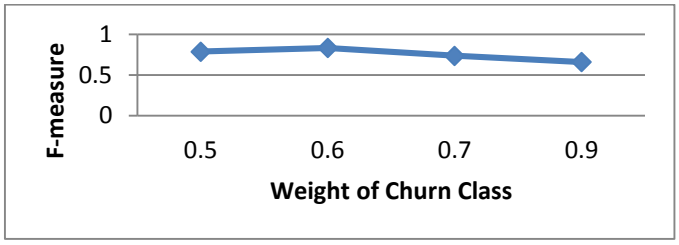

Figure 3. Trend of Weight Change to F-measure (data telco + US 8/10 + SMOTE 50x).

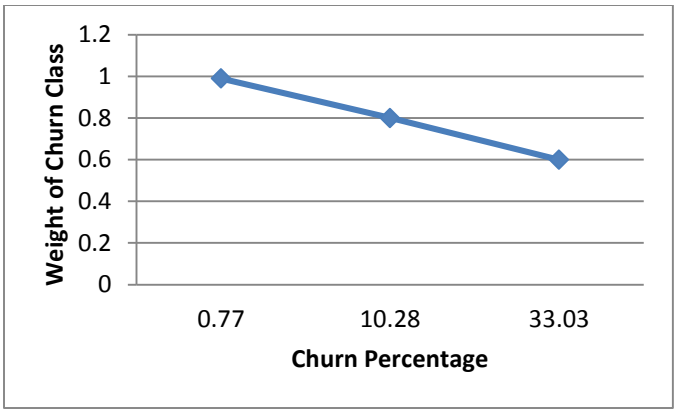

Figure 4. Trend of churn percentage to weight of churn class.

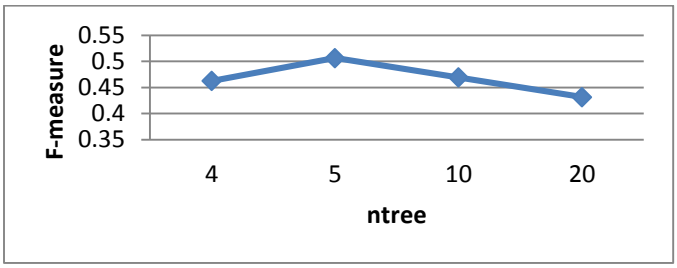

Figure 5. Trend of ntree Change to F-measure (data telco + US ¾ + SMOTE 10x).

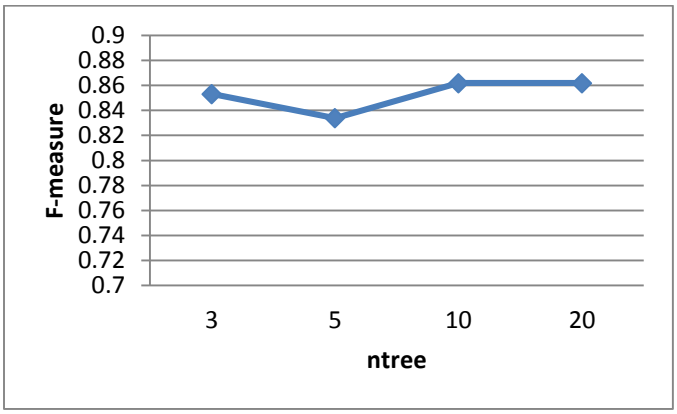

Figure 6. Trend of ntree Change to F-measure (data telco + US 8/10+ SMOTE 50x).

From those several conducted experiments, top-decile were calculated and data result were summarized in TABLE VI. Fluctuating value of the top-decile happened. This is because the value of the top-decile does not deal directly with the final label prediction results. Top-decile only use the results of prediction to determine the 10% of customers are most at risk of churn, then from these data, the actual churn will be counted.

TABLE VI. THE COMPARISON OF THE PERFORMANCE RESULTS.

Data File + Method	Churn pctg (%)	ntree	Weight of churn class	F-measure	Top-decile
Data telco + WRF	0.77	5	0.99	0.027273	1.1132
Data telco + WRF + US 3/4 + SMOTE 10x	10.28	5	0.8	0.50643	3.9428
Data telco + WRF + US 8/10 + SMOTE 50x	33.03	10	0.6	0.86195	1.7379

V. CONCLUSSION

This research proposed the combination of combined-sampling and WRF. As shown in the previous section, we compared the performance of prediction models developed by WRF without any sampling technique, WRF with a sampling technique, and also WRF with combined-sampling.

The results from the scenario-1 shown that the best F-measure achieved by the implementation of WRF without any sampling technique was 0.027273. The result of scenario-2 shown that the implementation of any sampling technique combined with WRF could give a significant improvement in the value of F-measure. The best F-measure in the experiments achieved by the proposed method (i.e. combined-sampling and WRF). Combined-sampling (oversampling and undersampling) improved the F-measure value of the churn prediction model produced by WRF. Moreover, the implementation of undersampling reduced the number of data records in the dataset caused by the oversampling, so it minimized the computational cost.

The results of scenario 1 and 2 indicated that an increase in the performance (in this case, the performance is represented as the value of F-measure) is quite significant to the prediction model formed with WRF. From this results, it is proven that sampling technique was able to assist WRF algorithms to better predict the churn. By using the original data, the data churn was very difficult to find, although the weight value set to a very high value. The difficulties of getting the churn data was very influential in the learning process and model testing. In the process of bootstrapping the data, there was a possibility of data churn drawn very little or even nothing at all, so the formed classification tree cannot be used to predict the data churn. This difficulty was answered with the implementation of the sampling technique to increase the churn rate and WRF method that implemented greater weight to churn class.

TABLE VI shown that the high value of F-measure do not guarantee that the value of top-decile will be high anyway. In terms of pure classification, the best classifier is a classifier which has the highest F-measure and highest accuracy. But in the case of churn prediction, in terms of management, related to the limited budget, the classifier which has a high top-decile with a reasonable performance might be the choice.

Based on the results achieved in this research, the performance generated by the modeling is quite good, but it needs more research to be able to increase the value of top-decile. Churn prediction model which has a good performance and a good top-decile, can reduce the cost of doing treatment against potential customer churn. In order to help the management to determine the appropriate strategies, besides getting the right target (by increasing the value of top-decile), other work is necessary conducted to get what most attributes can influence customers to churn. Based on this information, management is expected to determine the appropriate action to be made to the appropriate target customers anyway. In the end, after the determined strategy is implemented, it is necessary to evaluate whether the big problems of churn can be resolved well or not.

REFERENCES

[1] Y. Xie, X. Li, E. Ngai dan W. Ying, "Customer churn prediction using improved balanced random forests," *Elsevier,Expert System with Application 36,* pp. 5445-5449, 2009.

[2] D. M. Maharaj, "Evaluating Customer Relations in The Cell phone Industry," *IJBSM (International Journal for Business, Srategy & Management) Vol 1 No1 ,* 2011.

[3] B. Huang, M. T. Kechadi dan B. Buckley, "Customer Churn Prediction in Telecomunications," *Elsevier, Expert Systems with Applications,* pp. 1414-1425, 2012.

[4] Z. A. Baizal dan A. S. Moch Arif Bijaksana, "Analisis Pengaruh Metode Over Sampling dalam Churn Prediction untuk perusahaan Telekomunikasi," *SNATI ISSN: 1907-5022,* 2009.

[5] J. Burez dan D. d. Poel, "Handling Class Imbalance in Customer Churn Prediction," *Elsevier.Expert Systems with Applications 36,* p. 4626–4636, 2009.

[6] C. Chen, A. Liaw and L. Breimenn, "Using Random Forest to Learn Imbalanced Data," Statistics Department of University of California, Berkeley, 2004.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[8] L. Breimann, Random Forest, Netherlands: Kluwer Academic Publishers, 2001.

[9] P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Boston: Person Education,Inc., 2006.

[10] S. A. Neslin, S. Gupta, W. kamakura, J. Lu and C. Mason, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," *Journal of Marketing Research,* pp. 204-211, 2006.

[11] R. Mattison, The Telco Churn Management Handbook, Oakwood Hills, Illinois: XiT Press, 2005.