Syntax Literate: Jurnal Ilmiah Indonesia p–ISSN: 2541-0849
e-ISSN: 2548-1398
Vol. 6, Special Issue No. 2, Desember 2021

APPLYING SMOTE-NC ON CART ALGORITHM TO HANDLE IMBALANCED DATA IN CUSTOMER CHURN PREDICTION: A CASE STUDY OF TELECOMMUNICATIONS INDUSTRY

Ilma Amira Rahmayanti[1], Sediono[2], Toha Saifudin[3], Elly Ana[4]
[1,2,3,4] Statistics Study Program, Faculty of Science and Technology, University of Airlangga, Surabaya, Indonesia

Email: ilma.amira.rahmayanti-2018@fst.unair.ac.id, sediono@fst.unair.ac.id, tohasaifudin@fst.unair.ac.id, elly-a@fst.unair.ac.id

Abstract

These days, telecommunications is very much needed in all areas of life. This condition has made the competition among the company is extremely tense. One strategic way to protect the company is to retain existing customers. The retention program as a scheme to retain customers must be implemented precisely and efficiently so that the company can maintain as many customers as possible. In this case, customer churn prediction holds an essential role. However, the existence of imbalanced data can increase prediction errors and create problems. Hence, in order to overcome the issue, this study combined the Synthetic Minority Oversampling Technique – Nominal Continuous (SMOTE-NC) with Classification and Regression Trees (CART). SMOTE-NC was applied to balance classes on training data, while CART formed a classification tree from those balanced data. Then, this classification tree created by CART algorithm had become the basis for predicting customer churn. The data used in this study are from https://community.ibm.com/, where the variables are related to customer demographics, customer contracts, usage history, and customer status of one of the telecom companies. Based on the analysis of these data, SMOTE-NC and CART combination succeeded in reducing errors in predicting customer churn, which also led recall value to increase by approximately 19%. Moreover, the accuracy generated from this combination method was still in a pretty good range of over 75%. Therefore, this study proposes an excellent way to improve the performance of churn prediction, especially in the telecommunications industry.

Keywords: SMOTE; CART; decision tree; machine learning; customer churn prediction

Introduction

Information and Communications Technology (ICT) has progressed rapidly. People are now using telecommunications, such as the internet, as a way to communicate in a long-distance and in a short time (Stiawan et al., 2020). This increasing need for long-distance communications results in a tense business rivalry among the companies. In order to protect their existence, telecom companies tend to increase their promotional activities and develop more innovations nowadays; thus, they can prevent their consumers from switching to competitors (Mukaromah & Wijaya, 2020). According to Customer Relationship Management (CRM), retaining existing customers is a better marketing strategy than finding new ones (Ballings & den Poel, 2012). Finding new customers is more inefficient as it takes more time and costs (Almana et al., 2014).

Customer churn is a condition when the company loses its customer as the customer stops subscribing or purchasing products after a while. Many telecommunications companies suffered customer churn mainly because of mistargeting in implementing retention programs (Jain et al., 2020). Theoretically, the main goal of the retention programs is to prevent a potential churn, which means the retention programs are supposed to be targeted only to customers who have the potential to become disloyal. This action has to be done with the expectation that those customers will have several reasons to discourage their intention to leave the company. However, in real life, retention programs were often applied to all company's customers, including potentially loyal customers, so that the programs ran in vain as the company had already spent a lot of money implementing the programs. Therefore, a customer churn prediction, which is based on the classification of customer loyalty characteristics, is very beneficial in determining how the retention programs will be run precisely so that the company can save its resources and retain the customers as much as possible (Almana et al., 2014).

Classification is a way of learning the data characteristics in order to form a model, which then becomes the basis for the prediction process. The application of the classification method can be made with two approaches, namely the parametric approach and the nonparametric approach (Syaraswati et al., 2017). Nevertheless, in its implementation, the parametric approach requires many assumptions and is difficult to interpret (Lestawati et al., 2018). Meanwhile, the data owned by telecommunications companies are generally real-time and in large dimensions (Zahid et al., 2019). In this case, the nonparametric approach is much more flexible in dealing with complex data (Zhang, 2018). Hence, this study used the classification tree from Classification and Regression Trees (CART) algorithm as a nonparametric approach to classifying customer loyalty characteristics and predicting customer churn.

Classification and Regression Trees (CART) is a decision tree algorithm proposed by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone in 1984. CART utilizes the Gini diversity index to separate a node into two branches, commonly called binary splitting (Sumartini, 2015). The CART algorithm is

very flexible in dealing with numerical and categorical variables. The tree will be called a classification tree if its dependent variable is categorical; on the other hand, the tree will be called a regression tree if its dependent variable is numerical.

This study applied CART algorithm as the classification method due to its advantages, which are: 1) it does not require many pre-processing steps; 2) its model is extremely easy to interpret; 3) it is immune to multicollinearity; 4) it can eliminate insignificant variables by itself; and 5) it is accurate and fast in making classification (Anindya et al., 2018; Ghiasi et al., 2020; Singh & Gupta, 2014). Therefore, it is not surprising that many researchers have used the CART algorithm to make classifications. Previous studies showed that the CART algorithm had better performance than other methods. (Rai et al., 2020) compared CART and logistic regression in predicting customer churn of telecom sector, in which the result showed that CART produced better accuracy than the logistic regression. (Vafeiadis et al., 2015) also compared CART and other machine learning techniques, where the result showed that CART is considered as a good method to predict customer churn. Meanwhile, (Soeini & Rodpysh, 2012) stated that CART had better accuracy than the other data mining techniques in predicting customer churn of an insurance company.

However, the previous studies generally had not considered the existence of imbalanced data in churn prediction cases. In the meantime, a classification tree is extremely vulnerable to imbalanced data (Anindya et al., 2018). The existence of majority and minority classes can increase misclassification. Hence, one of the alternatives to improve the classification and prediction performances is to apply a resampling method, namely the Synthetic Minority Oversampling Technique (SMOTE). SMOTE is a resampling technique proposed by Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer in 2002. SMOTE has three different algorithms based on the variables' scale in the data. Because of the nominal and continuous scales of the independent variables, this study used the algorithm from Synthetic Minority Oversampling Technique – Nominal Continuous (SMOTE-NC). As reported by (Gök & Olgun, 2021), the SMOTE-NC algorithm was proven to increase the prediction performance in imbalanced data, specifically by increasing its recall value.

Under this justification, the objectives of this study were: 1) to find out how the CART algorithm can predict customer churn in the telecommunications industry; 2) to find out how SMOTE-NC influences CART's performance in making predictions. The methods used in this study are not limited to being applied in a company utilized as this research object, but they can also be applied to other companies in the telecommunications industry; or even in different industries. Hopefully, appropriate analytical methods can produce accurate customer churn predictions so that the results can be employed as the basis to conduct good retention programs. In turn, well-run retention programs can increase a company's income and improve national economic growth.

Research Method

1. Data and Variables

The data used in this study are customer data from a fixed-line telecommunications company, which were downloaded from https://community.ibm.com/. Those data include the information from 7043 customers, consisting of customers who were still subscribing to the company as of September 2017 and new customers who joined in September-December 2017. The information contained is related to customer demographics, customer contracts, usage history, and customer status. The data in this study then divided into training set and testing set with two different ratios, namely 80:20 and 90:10.

Meanwhile, the dependent variable used is *Churn*, in which it has two possible values, namely *Churn* and *Loyal*. The customer was identified as *Churn* if she/he voluntarily canceled subscriptions or is isolated for not making payments in the last month. On the other hand, the customer was identified as *Loyal* if she/he still subscribes or purchases products in that company.

(Maulana, 2016) stated that the company's service quality and price significantly influence customer loyalty. In addition, the identity of a customer, such as age, marital status, number of insured, and length of subscription, is also known to have a significant role (Suparto, 2008). Based on those facts, this study used 19 independent variables, which are: *Gender*, *Senior Citizen*, *Partner*, *Dependents*, *Tenure*, *Phone Service*, *Multiple Lines*, *Internet Service*, *Online Security*, *Online Backup*, *Device Protection*, *Tech Support*, *Streaming TV*, *Streaming Movies*, *Contract*, *Paperless Billing*, *Payment Method*, *Monthly Charges*, and *Total Charges*.

2. Analysis Steps

This study was conducted in four steps, which are: 1) data pre-processing; 2) classification tree construction; 3) customer churn prediction; and 4) performance assessment. All of these steps were carried with the cross validation approach of two iterations. Also, the steps were executed through programming in Python (for data pre-processing) and R (for tree construction, prediction process, and performance assessment).

2.1 Data Pre-Processing

The data pre-processing stages consist of: 1) missing value detection; 2) feature selection; and 3) SMOTE-NC resampling.

a. Missing Value Detection

The observation was detected as a missing value if it has the null value. The imputation of missing value must be done according to the situation at hand.

b. Feature Selection

Feature selection is a step to choose the significant independent variables.

- If the independent variables are on continuous scales, the feature selection is conducted by training a logistic regression based on different sets of independent variables. Then, the set that produces the largest accuracy will be used in the model.
- If the independent variables are on nominal scales, feature selection is done using the chi-square tests, where the independent variables are removed if they do not have significant relationship with the dependent variable.

c. SMOTE-NC Resampling

SMOTE-NC is a resampling technique that uses *k*-nearest neighbors, employing the modified-Euclidean distances to generate synthetic data, specifically the data on training set (Wijaya et al., 2018). In this study, SMOTE-NC was repeated with three different *k* values, namely *k* = 3, 4, and 5. The more detailed steps of SMOTE-NC algorithm are: 1) computing *k*-nearest neighbors; and 2) creating synthetic data.

K-nearest neighbors are a set of *k* observations that have the nearest modified-Euclidean distances if calculated from the reference observation (which is randomly chosen from the original data). The formula of modified-Euclidean is almost similar to the original Euclidean, but it also adds a median value of the standard deviations of all minority class's continuous variables. This median value is used to penalize the difference of nominal variables. The mathematical formula of the modified-Euclidean distance on SMOTE-NC can be seen in Equation 1.

$$\Delta(x, y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2 + \sum_{j=1}^{q} Med^2} \quad (1)$$

In which, $\Delta(x, y)$ is the distance between observation x and observation y; p represents the number of continuous variables in each observation; q represents the number of nominal variables that differ between observations x and y; x_i is the value of the i-th variable on the observation x; y_i is the value of the i-th variable on the observation y (reference observation); and Med is the median of the standard deviations of all minority class's continuous variables.

Furthermore, after *k*-nearest neighbors have been obtained, synthetic data in SMOTE-NC are created by synthesizing a new observation for the minority class.

- If the independent variables are continuous, the synthesis is done based on the mathematical formula at Equation 2.

  $$\boldsymbol{y_{new}} = \boldsymbol{y} + (\boldsymbol{x} - \boldsymbol{y}) \times rand[0,1] \quad (2)$$

  Where, $\boldsymbol{y_{new}}$ is $1 \times p$ vector of new values for continuous independent variables (synthetic results); $\boldsymbol{y}$ is $1 \times p$ vector of values of continuous independent variables in the reference observation; $\boldsymbol{x}$ is $1 \times p$ vector of values of continuous independent variables in the observation taken at random from the *k*-nearest neighbors; and $rand[0,1]$ represents a scalar value that produced from the randomization between 0 and 1.

- If the independent variables are nominal, the synthesis is done by looking at the majority voting of the *k*-nearest neighbors vector, means that the category which appears the most will be chosen as the value of new observation.

A few steps of SMOTE-NC need to be repeated until the training set are balanced and can be used for model construction.

2.2 Classification Tree Construction

This study constructed the classification trees by classifying the customer loyalty characteristics with CART algorithm. CART algorithm consists of three stages, namely: 1) node splitting; 2) leaf node election; and 3) leaf node labelling. In this study, the classification trees were also simplified with pruning so that their forms are not too complex. This study used two pruning parameters, which are: minimum split parameter and complexity parameter (α). These pruning parameters were optimized with 10-fold cross validation.

a. Node Splitting

The node splitting in CART is done by looking at the ability of the splitter to reduce the class heterogeneity, that is, by finding the best splitter that produces the largest goodness of split. Thus, all possible splitters of an independent variable must be considered. Equations 3 to 5 state the number of possible splitters on an independent variable.

Numeric variable $= b - 1$ splits (3)

Ordinal variable $= L - 1$ splits (4)

Nominal variable $= 2^{L-1} - 1$ splits (5)

In which, b is the number of observations of a variable, and L is the number of categories of a variable.

After all splitters are identified, the step is continued into the calculation of Gini diversity index as the value of the heterogeneity function. The Gini diversity index is defined at Equations 6 and 7 as follows.

$$Gini(n) = 1 - \sum_{j=1}^{m} P^2(j|n) \quad (6)$$

$$Gini_{split}(n) = \frac{b_1}{b} Gini(D_1) + \frac{b_2}{b} Gini(D_2) \quad (7)$$

Where, $Gini(n)$ represents the value of the Gini diversity index at the n-th node; m represents the number of classes on the dependent variable; $P(j|n)$ is the conditional probability of the occurrence of class j at node n; $Gini_{split}(n)$ is the value of the Gini diversity index after splitting at node n; $Gini(D_1)$ is the value of the Gini diversity index in subset D_1; $Gini(D_2)$ is the value of the Gini diversity index in subset D_2; b represents the number of observations on a variable; b_1 represents the number of observations in subset D_1; and b_2 represents the number of observations in subset D_2.

As stated before, the best node splitter is chosen from the splitter that can generate the largest value of goodness of split. The goodness of split is the value that

shows the magnitude of heterogeneity reduction. Equation 8 shows the formula for calculating the goodness of split.

Goodness of split = $Gini(n) - Gini_{split}(n)$ (8)

b. Leaf Node Election

Node splitting in CART continues if the number of observations of a node is still not less than the minimum split parameter, and/or the observations in the node have not been gathered into one class. However, if the case is another way around, the node splitting can be stopped, and the node is elected as a leaf node.

c. Leaf Node Labelling

The label on the leaf node is chosen from the class with the highest number of members. Mathematically, the rule for the leaf node labelling is shown in Equation 9.

$$P(j_0|t) = max_j \frac{m_j(t)}{m(t)} \quad (9)$$

Where, $m_j(t)$ is the number of observations belonging to class j in leaf node t; and $m(t)$ is the number of observations in the leaf node t.

In this case, if $P(j|t)$ is equal to $P(j_0|t)$, then the class will be marked as a label on the leaf node t.

2.3 Customer Churn Prediction

In this study, customer churn prediction was conducted by referring to the classification trees that have been previously built.

2.4 Performance Assessment

The performance of customer churn predictions was evaluated by using evaluation metrics. The evaluation metric compares the predicted results with the actual data. A tool for calculating evaluation metric is the confusion matrix. Table 1 below shows the illustration of the confusion matrix.

Table 1. Confusion Matrix

CLASS		**ACTUAL CLASS**	
		Positive	**Negative**
PREDICTED CLASS	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Many evaluation metrics can be used to measure the performance of a classification tree in predicting classes. Some of these evaluation metrics can be written mathematically in Equations 10 to 13.

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ (10)

F-measure = $2(\frac{Precision \times Recall}{Precision + Recall})$ (11)

Recall = $\frac{TP}{TP+FN}$ (12)

Precision = $\frac{TP}{TP+FP}$ (13)

However, (Di Martino et al., 2013) stated that the f-measure is more suitable to measure the performance of a method on imbalanced data. So, in this study, the best classification model is chosen based on the largest f-measure.

Result and Discussion

1. Data Description

The first stage of this study is done by describing the customer characteristics based on their descriptive statistics. Customers in this research's object are, on average, potentially loyal. This is evidenced by the number of customer churn, which is much smaller than the number of loyal customers. Customers who stopped their subscriptions are 1869 customers, while customers who remained being loyal are 5174 customers. Based on these numbers, it can be seen that there is an imbalanced class in the data.

Furthermore, the characteristics of the company's customers can also be described through the frequency table. The frequency table for each categorical independent variable is shown in Table 2.

Table 2. Frequency Table

Variable Name	Frequency for Each Category
Gender	Female = 3488, Male = 3555
Senior Citizen	No = 5901, Yes = 1142
Partner	No = 3641, Yes = 3402
Dependents	No = 4933, Yes = 2110
Phone Service	No = 682, Yes = 6361
Multiple Lines	No = 3390, No phone service = 682, Yes = 2971
Internet Service	Yes (DSL) = 2421, Yes (fiber optic) = 3096, No = 1526
Online Security	No = 3498, No internet service = 1526, Yes = 2019
Online Backup	No = 3088, No internet service = 1526, Yes = 2429
Device Protection	No = 3095, No internet service = 1526, Yes = 2422
Tech Support	No = 3473, No internet service = 1526, Yes = 2044
Streaming TV	No = 2810, No internet service = 1526, Yes = 2707
Streaming Movies	No = 2785, No internet service = 1526, Yes = 2732
Contract	1 month = 3875, 1 year = 1473, 2 years = 1695
Paperless Billing	No = 2872, Yes = 4171
Payment Method	Bank transfer = 1544, Credit card = 1522, Electronic check = 2365, Mailed check = 1612

Based on Table 2, it can be seen that the characteristics which most customers commonly have:

- The customers are mostly young people with no dependents
- Most customers generally use fiber-optic, and they do not add any supporting services to their internet network
- The customers usually take 1-month length subscription
- The customers generally use paperless billing
- The customers are more likely to use the electronic check as their payment method

Meanwhile, when viewed from numerical independent variables, the customer characteristics can be shown by the boxplot in Figures 1(a) and 1(b).

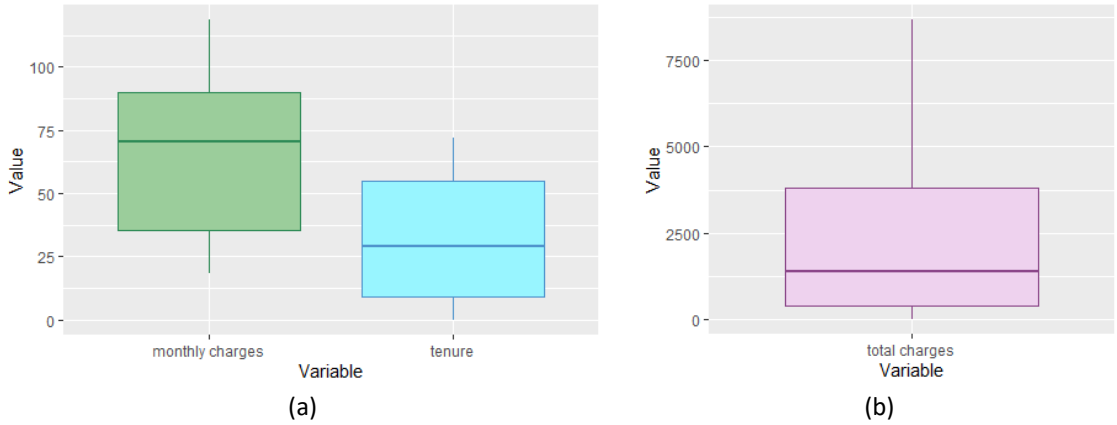

Figure 1
(a) Boxplot for Monthly Charges and Tenure (b) Boxplot for Total Charges

According to Figures 1(a) and 1(b), it can be seen that *Monthly Charges*, *Tenure*, and *Total Charges* have a longer upper whisker line than the lower line. This situation shows that these variables have a positive skew, meaning that their values tend to be small. For further details, the average subscription time is 32 months, the average charge per month is 74 U.S. Dollars, and the average total charge is 2280 U.S. dollars. These low values of charges further prove that many customers do not use additional services on the telephone and/or internet network.

In the meantime, Figures 1(a) and 1(b) also show that most customers subscribed for a short time. Many things can cause this condition. For example, there are too many disloyal customers in a company; many new customers are joining; etc. The factors that cause this situation can be seen closely from the classification of customer characteristics. The classification model can show the characteristics which make customers more likely to be disloyal, starting from their demographic data, subscription contracts, and history of service usage.

2. Data Analysis and Discussion

From the pre-processing conducted before tree construction, it could be found that:

- The *Total Charges* has seven missing values. These missing values are mostly due to the post-paid payment regulation of the company. Because of this regulation, if the tenure is less than one month, then the company has not requested any payment from the customer, and the data detected it as a missing value. So, in this case, these missing values are imputed with the monthly charges.
- *Gender*, *Phone Service*, and *Total Charges* do not influence the dependent variable. Therefore, these variables were removed from the model.

After completing the pre-processing, the analysis was continued to the construction of classification trees. As stated before, the tree construction was carried with cross validation approach of two iterations. Hence, this study will discuss the

prediction performances based on the average values of evaluation metrics. This study used the *Churn* as a positive class, and *Loyal* as a negative class. Table 3 below shows the average values of f-measures that quantify the performances of the classification trees.

Table 3. The Average Values of F-Measures

SMOTE-NC TYPE	RATIO	
	80:20	**90:10**
Without SMOTE-NC	0.5826	0.5815
SMOTE-NC with *k* = 3	0.6177	0.6046
SMOTE-NC with *k* = 4	0.5976	0.5947
SMOTE-NC with *k* = 5	0.6270	0.5905

As reported by Table 3, the best prediction performance is produced by the classification tree that was trained using 80:20 data ratio and synthesized through SMOTE-NC with k = 5. Hence, in this case, the SMOTE-NC succeeded in increasing the f-measure by approximately 4%. This performance improvement can be illustrated and analyzed further in Figures 2 and 3 below.

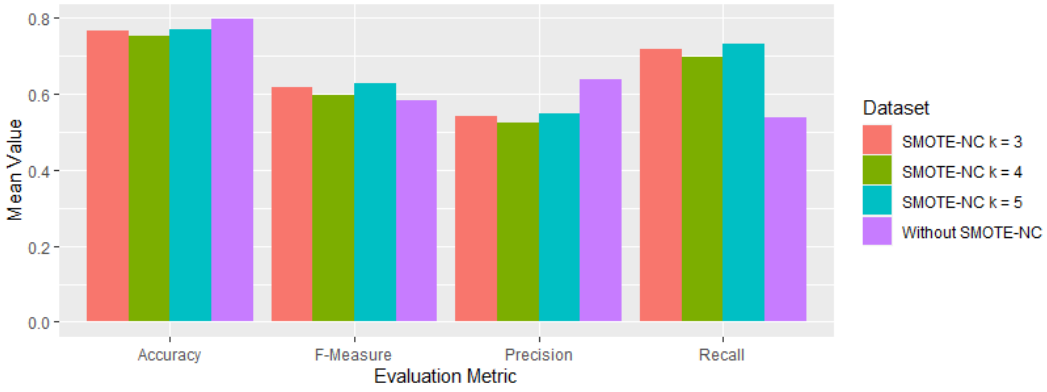

Figure 2
The Average Values of the Evaluation Metrics for 80:20 Data Ratio

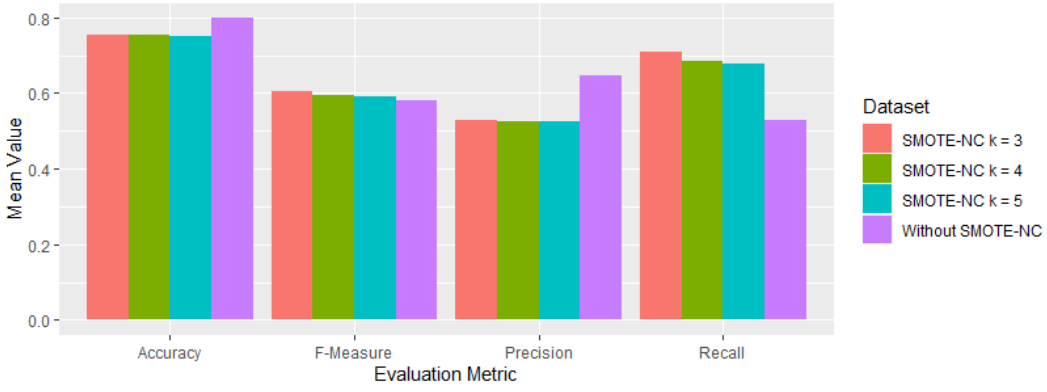

Figure 3
The Average Values of the Evaluation Metrics for 90:10 Data Ratio

Based on Figures 2 and 3, we can see that the dataset which does not use the SMOTE-NC procedure have a smaller average recall value than the one which uses the SMOTE-NC procedure. This situation happened as the SMOTE-NC method synthesizes many minority classes (which is the *Churn* classes) so that the model can learn more and predict more of them.

However, as illustrated in Figures 2 and 3, the use of SMOTE-NC can increase the "false positive" and reduce the precision values. This happened as the classification tree tended to predict more *Churn* classes, even though there was a possibility that some observations should be classified in negative classes instead. But, despite that, these degradations are not as significant as the improvements of recall, so the company can slightly ignore them as they are not too distracting.

In CRM, it must always be remembered that retention programs also cost money and time. Thus, from the company's point of view, it is expected that the company does not waste too many resources, but it can retain its customers as much as possible. Therefore, through churn prediction, it is hoped that the errors in predicting customer churn (when customers who turn out to be churn were at first predicted as loyal customers) can be minimized.

In line with previous justification, it can be assumed that "false negative" should be more considered rather than "false positive". Consequently, recall is the evaluation metric that should be improved more than other metrics. This study conducted efforts to increase recall by adding the SMOTE-NC procedure in the data pre-processing. This effort can be said to be successful because there is an increase in the average recall values. Besides, the improvement of recall also made the f-measure increase, in which it is good news for the predictions of imbalanced data.

Following Figures 2 and 3, it can be seen that there are no significant differences in the evaluation metrics between the three values of k used in SMOTE-NC. However, a smaller k value is recommended as it can create time efficiency, especially if the resampling is applied to data with an extremely imbalanced class.

In addition, referring to Figures 2 and 3, it can be calculated that the evaluation metrics in Figure 2 are averagely greater than the metrics in Figure 3. This maybe because the 80:20 data ratio has more observation members in testing data, so that they can better capture the characteristics described by training data. Thus, using the proportion of training data of 80% and testing data of 20% is more recommended.

For further details of explanation, a classification tree that had the best performance (classification tree which trained using 80:20 data ratio and synthesized through SMOTE-NC with $k = 5$) was created again. This classification tree consists of 1377 nodes; including one root node, 687 internal nodes, and 689 terminal nodes. However, due to the model complexity, the pruning was conducted with the minimum split parameter = 7 and complexity parameter = 1.047×10^{-3}. As a result, the classification tree changed into a tree with 101 nodes; consisting of one root node, 49 internal nodes, and 51 terminal nodes. In this tree, there are 19 terminal nodes indicate the customers will leave the company, and 32 terminal nodes indicate the customers will

be loyal. The performance of this tree in predicting churn, is quantified by: accuracy = 76.86%; f-measure = 62.7%; recall = 73.26%; and precision = 54.8%. According to this tree, the customer characteristics can be traced from its subtrees, whether they have the potential to be disloyal or loyal. Figures 4 and 5 below show some subtrees generated by this classification tree.

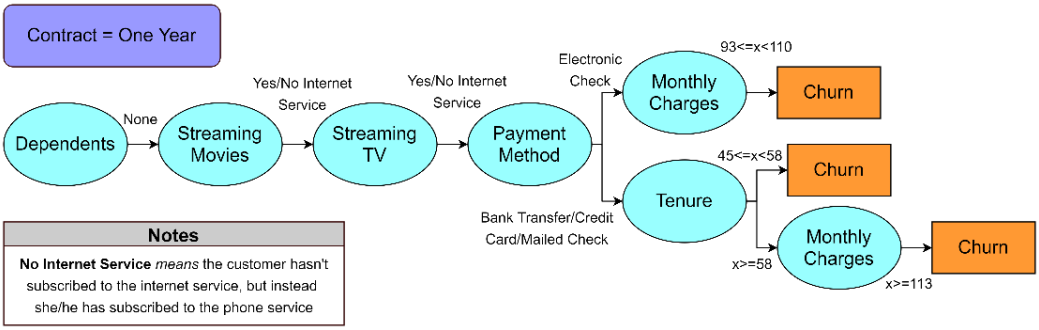

Figure 4
Subtree 1

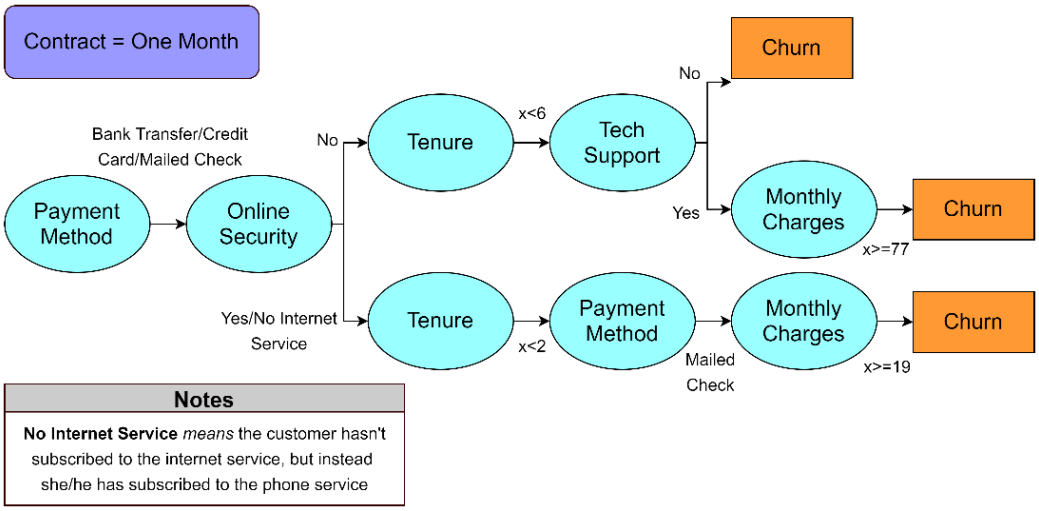

Figure 5
Subtree 2

By looking at Figure 4, Figure 5, and some other subtrees, the summary of the characteristics of potentially disloyal customers are:

- Customers whose contracts are short (one month and one year) are more likely to leave the company faster. This is assumed as the customer has a greater chance of not getting a penalty, in which this penalty is usually given when the customer leaves the company before the contract ends, where the customer has to pay some money to revoke the subscription.
- Customers who use electronic checks and mailed checks as their payment methods generally have the potential to leave the company faster. This may be due to the difficulties in payments via electronic checks and mailed checks.

- Customers with monthly charges that exceed 80 U.S. Dollars are usually leave the company faster. This is because of the tendency of customers to look for products or services at lower prices.
- Customers who subscribe for less than 29 months are more likely to leave faster. This is probably due to the customer's loyalty that have not yet been formed.
- Customers who do not have dependents generally have the potential to leave after three years from the contract's end. This may be due to the customers' desire to try the competitor's products that provide more special offers; or the customers no longer need long-distance communication as they are old, have no relatives, etc. Customers in this category generally do not care about the prices, but they pay more attention to their needs and convenience.
- Customers who use paperless billing are more likely to leave faster. This is possibly due to the inefficiency of the paperless billing.
- Customers who use technical support, online security, and/or online backup on their internet network are more likely to leave faster, as these services may cause the increases in the amount of bill.
- Customers who only subscribe to the phone service and do not subscribe to multiple lines are generally more likely to be disloyal. This is most likely because these customers use phone service for their personal consumption, and now they prefer other methods.

Conclusion

Based on the results, the CART algorithm is fairly quick in forming a classification model. This algorithm also does not require many pre-processing steps and is flexible to be applied to data with large dimensions. These factors make this algorithm suitable for use in the telecommunications industry, where the data are usually large and real-time. However, in the customer churn prediction case, the number of disloyal customers is generally much smaller than that of loyal customers. This imbalanced data may increase the errors in the prediction process. Therefore, this study uses the SMOTE-NC procedure to synthesize the minority class on the training data. This step is done before forming a classification tree with the CART algorithm. Referring to the outcome of this study, the SMOTE-NC procedure is proven to reduce errors in predicting churn, where the recall value increased by approximately 19%. The increase in recall value also causes an increase in f-measure, where it can be concluded that applying SMOTE-NC to the CART algorithm can make churn predictions more precise. In addition, the accuracy generated from this algorithm is still in a fairly good range of over 75%. So, in the future, the combination of SMOTE-NC and CART is recommended to apply in customer churn prediction cases, even in other telecommunication companies, or other industries. Besides, in order to improve prediction performance, further study can also be developed using other methods, such as the ensemble method.

BIBLIOGRAPHY

Almana, A., Aksoy, M., & Alzahrani, R. (2014). A Survey on Data Mining Techniques in Customer Churn Analysis for Telecom Industry. *International Journal of Engineering Research and Applications*, *5*(6), 165–171. Google Scholar

Anindya, A., Indahwati, I., & Susetyo, B. (2018). Application of SMOTE on CART Method to Handle Imbalanced Data (Study Case: Labor Force Classification in Banten Province). *IOP Conference Series: Earth and Environmental Science*, *187*, 12055. Google Scholar

Ballings, M., & den Poel, D. (2012). Customer Event History for Churn Prediction: How Long is Long Enough? *Expert Systems with Applications*, *39*(18), 13517–13522. Google Scholar

Di Martino, M., Fernández, A., Iturralde, P., & Lecumberry, F. (2013). Novel Classifier Scheme for Imbalanced Problems. *Pattern Recognition Letters*, *34*(10), 1146–1151. Google Scholar

Ghiasi, M. M., Zendehboudi, S., & Mohsenipour, A. A. (2020). Decision Tree-Based Diagnosis of Coronary Artery Disease: CART Model. *Computer Methods and Programs in Biomedicine*, *192*, 105400. Google Scholar

Gök, E. C., & Olgun, M. O. (2021). SMOTE-NC and Gradient Boosting Imputation based Random Forest Classifier for Predicting Severity Level of Covid-19 Patients with Blood Samples. *Neural Computing & Applications*, 1–15. Google Scholar

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science*, *167*, 101–112. Google Scholar

Lestawati, R., Rais, R., & Utami, I. T. (2018). Perbandingan antara Metode CART (Classification and Regression Trees) dan Regresi Logistik (Logistic Regression) dalam Mengklasifikasikan Pasien Penderita DBD (Demam Berdarah Dengue). *Jurnal Ilmiah Matematika Dan Terapan*, *15*(1), 98–107. Google Scholar

Maulana, A. S. (2016). Pengaruh Kualitas Pelayanan dan Harga terhadap Kepuasan Pelanggan PT. TOI. *Jurnal Ekonomi*, *7*(2), 113–125. Google Scholar

Mukaromah, N. F., & Wijaya, T. (2020). Pasar Persaingan Sempurna dan Pasar Persaingan Tidak Sempurna dalam Perspektif Islam. *PROFIT: Jurnal Kajian Ekonomi Dan Perbankan*, *4*(2), 1–16. Google Scholar

Rai, S., Khandelwal, N., & Boghey, R. (2020). Analysis of Customer Churn Prediction

in Telecom Sector Using CART Algorithm. *First International Conference on Sustainable Technologies for Computational Intelligence*, 457–466. Singapore: Springer. Google Scholar

Singh, S., & Gupta, P. (2014). Comparative Study ID3, CART and C4.5 Decision Tree Algorithm: A Survey. *International Journal of Advanced Information Science and Technology (IJAIST)*, *27*(27). Google Scholar

Soeini, R. A., & Rodpysh, K. V. (2012). Evaluations of Data Mining Methods in Order to Provide the Optimum Method for Customer Churn Prediction: Case Study Insurance Industry. *International Conference on Information and Computer Applications*, 290–297. Singapore: IACSIT Press. Google Scholar

Stiawan, A., Baharuddin, H., & Amrozi, Y. (2020). Masa Depan Teknologi Komunikasi Data, Menebak Arah Perkembangannya. *Journal of Information Technology*, *5*(2), 1–5. Google Scholar

Sumartini, S. H. (2015). Penggunaan Metode Classification and Regression Trees (CART) untuk Klasifikasi Rekurensi Pasien Kanker Serviks di RSUD Dr. Soetomo Surabaya. *Jurnal Sains Dan Seni ITS*, *4*(2), 2337–3520. Google Scholar

Suparto. (2008). Perilaku dan Kepuasan Pelanggan Bank Muamalat Indonesia Cabang Surabaya dengan Menggunakan Analisis Regresi Logistik. *Jurnal Keuangan Dan Perbankan*, *12*(2), 331–341. Google Scholar

Syaraswati, R. A., Slamet, I., & Winarno, B. (2017). Classication of Status of the Region on Java Island using C4.5, CHAID, and CART Methods. *Journal of Physics: Conference Series*, *855*, 12053. Google Scholar

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A Comparison of Machine Learning Techniques for Customer Churn Prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9. Google Scholar

Wijaya, J., Soleh, A. M., & Rizki, A. (2018). Penanganan Data Tidak Seimbang pada Pemodelan Rotation Forest Keberhasilan Studi Mahasiswa Program Magister IPB. *Xplore*, *2*(2), 32–40. Google Scholar

Zahid, H., Mahmood, T., Morshed, A., & Sellis, T. (2019). Big Data Analytics in Telecommunications: Literature Review and Architecture Recommendations. *IEEE/CAA Journal of Automatica Sinica*, *7*(1), 18–38. https://doi.org/10.1109/JAS.2019.1911795 Google Scholar

Zhang, H. H. (2018). Nonparametric Methods for Big Data Analytics. In W. K. Härdle, H. H.-S. Lu, & X. Shen (Eds.), *Handbook of Big Data Analytics* (pp. 103–124). Cham: Springer International Publishing. Google Scholar