Utilizing Customers' Purchase and Contract Renewal Details to Predict Defection in the Cloud Software Industry

Niken Prasasti Martono[1,2], Katsutoshi Kanamori[1], and Hayato Ohwada[1]

[1] Department of Industrial Administration, Tokyo University of Science, Japan
[2] Graduate School of Business and Management, Bandung Institute of Technology, Indonesia
niken.prasasti@sbm-itb.ac.id, {katsu,ohwada}@rs.tus.ac.jp

Abstract. This study aims to predict customer defection in the growing market of the cloud software industry. Using the original unstructured data of a company, we propose a procedure to identify the actual defection condition (i.e., whether the customer is defecting from the company or merely stopped using a current product to up/downgrade it) and to produce a measure of customer loyalty by compiling the number of customers' purchases and renewals. Based on the results, we investigated important variables for classifying defecting customers using a random forest and built a prediction model using a decision tree. The final results indicate that defecting customers are mainly characterized by their loyalty and their number of total payments.

Keywords: Customer defection, Cloud software industry, Machine learning, Decision tree, Random forest.

1 Introduction

With the recent increase of Internet use, the cloud software industry has exhibited some very real trends. The cloud software market's 36% compound annual growth is predicted to continue through 2016 [1]. This increasing growth of software is supported by its convenience: it can be used everywhere as long as the users' devices are connected to the Internet. Examples of widely used cloud software are web-based file hosting, social networking, office applications, and security software.

Predicting customer defection is especially important for a fast-growing business with contractual models in order to improve marketing decision-making. Defection refers to a customer's decision to stop using the service or product provided by the company. Defection prediction has been a concern in research and industry, as it is an important measure used to retain customers [2]. In making predictions, most companies collect useful customer data in order to create a predictive model of defection using predictive analytic methods such as data mining and machine learning.

Here, we focus on defection in the cloud software industry. In this study, our case is a security software company. Though we are able to obtain customer data from the company's e-commerce site, predicting customer defection is not a simple task for

Y.S. Kim et al (Eds.): PKAW 2014, LNCS 8863, pp. 138–149, 2014.

four reasons. First, the data features are limited and include only a few customer attributes, unlike several previous works on defection prediction that use typical customer demographics, call logs, and usage details. Second, it is barely possible to gain more customer information by directly approaching each customer, since this company has many customers. Third, in particular, the available data contains simply the records of customer activity in opting-in (continuing) and opting-out (defecting) from one product, while in reality some customers are opting-out to upgrade/downgrade their product. Fourth, with the vast market growth, managing customer defection is an important issue.

This study seeks to tackle these challenges in managing customer defection in one security software company. First, we provide an algorithm in an effort to detect which customers are literally defecting from the company and which are not. Second, from the available data we produce a new feature that can be used as a measure of customer loyalty. Third, using a random forest, we analyze the most important variables that contribute to classifying defecting customers. In addition, we model customer defection using a decision tree, in order to have a visually interpretable result that can be useful for the company as the end user.

The remainder of this paper is organized as follows. Section 2 reviews former works that focus on defection prediction. Section 3 defines the data used in the study. Section 4 describes the data preparation procedures. Section 5 presents the machine-learning procedures in analyzing the important variables and predicting customer defection. Section 6 provides the results of the experiments. Finally, Section 7 discusses the conclusion and future work.

2 Related Works

In recent years, predicting customer defection has increasingly received the attention of researchers. Studies focus on the search for methods and features that most effectively predict defection. The most common methods used for defection prediction are decision tree, regression, Naïve Bayes, and neural network. Most former works focus on customer defection in the telecommunication industry.

Predicting customer defection involves searching for and identifying defecting indicators. Assuming that changes in call patterns may appear as defection warning signals, [3] used call details to extract the features that describe changes in customers' calling patterns. These features are then used as input into a decision tree to build classifiers. Using the same method, a decision tree, [4] discovered that the most significant differentiator between defecting and retained customers are age, tenure, gender, billing amount, number of payment, call duration, and amount of changing information. These findings were obtained using customer demographics, billing information, service status, and service change logs. Other useful features were explored by [5], using data containing customer complaints and service interactions with the operator to predict defection. They also compared the predicting performance of a neural network, a decision tree, and regression.

In [6] we reviewed the applicability of some machine-learning techniques to predicting customer defection using several common techniques such as a decision tree, a random forest, a neural network, and a support vector machine. In a complementary approach in [7], the result of predicting customer defection was applied to calculate the customer lifetime value, considering the strong relationship between customer defection/retention and the predicted customer lifetime value.

As previously mentioned, most former works relied on customer demographics, customer service logs, usage details, complaint data, bills, and payments. A relatively under-investigated source of input for predicting customer defection is the original purchase and renewal data of customers in a contract-based company, because the data often contains unstructured data that is difficult to analyze.

Our data are limited in the number of features and the structure is complicated, but in this study we provide a new system of customer defection management. We propose an algorithm to identify actual customer defection in order to make prediction more reliable and to produce from the available data a new feature that can measure customer loyalty. We subsequently use the results to analyze which variables are important in classifying defecting customers and to build a customer defection prediction model.

3 Data Set

The basic problem of predicting customer defection is finding a good model that can predict customer defection in a company. A quality model to predict defection can be constructed only if quality data is available. In this study, we use two types of data (purchase and auto-renewal data and web log data) and compare them for better prediction of customer defection.

Purchase and auto-renewal data contains six-year records (from 2007 up to 2013) of customer activity in purchasing and renewing their products. It includes the customer's contract ID, the latest status of the renewal flag, the latest date of auto-renewal contract, the total number of purchases and renewals, the type of product base that the customer purchased, the total payment by the customer, the warrant period of the product, whether or not an optional service is used, the type of customer (personal or commercial), and the status of e-mail delivery.

The web log data includes six-month log files (from January to June 2013) that contain total payment by the customer, use of optional service, type of customer (personal or commercial), type of operating system the customer uses, type of browser the customer uses to browse the Internet, number of website page views, number of website visits, number of product views and cart views, and number of orders the customer has made.

The data is originally used to record the details of "opting-in" and "opting-out" activities of each customer after receiving e-mail notification of auto-renewal. Customers who receive the e-mail will be automatically renewed for their current service unless they specifically choose to "opt-out." However, some customers may opt-out from one service of their product and opt-in for another service. Therefore, some new

features should be extracted from the original purchase and auto-renewal data for predicting actual defection from the company. In this case, actual defection means the customer does not renew or subscribe to any service from the company.

4 Data Preparation

Data preparation has one important rule that differentiates this study from former studies. The main purpose of data preparation in this study is to determine whether the original data may be used in developing a customer defection prediction model. If it is considered useful, it is necessary to determine which features can be extracted from it and may be useful for machine learning.

Table 1. Original table on the e-commerce site

CONTRACT_ID	NO_PURCHASE	$\mathbf{A_1} \ldots \boldsymbol{A_n}$	RENEW_ COUNT	RENEW_ FLAG
1	1		2	0
1	2		1	0
1	3	$v_1 \ldots v_n$	1	1
2	4		3	0
2	5	$u_1 \ldots u_n$	2	0

Table 1 illustrates the original contents of the table that contains historical records of customer activity collected from the company's e-commerce site, where $\mathbf{A_1} \ldots \boldsymbol{A_n}$ are the features previously mentioned in Section 2 for each type of data. It contains CONTRACT_ID, which is the ID number of a purchase or renewal that a customer makes. When one customer performs several actions, whether purchasing a new product or renewing the contract for a current product, the data is recorded by the e-commerce site under the same CONTRACT_ID. Thus, if we use the original data from the site without preparing it, the prediction model will not be reliable, since the site can only record data per activity. It does not provide a summary indicating whether the customer is truly defecting from the company or merely defecting from a current product.

To overcome this problem, we first detect actual defection by acquiring CLASS as the actual defection flag attribute of each customer. Second, we produce UPDATE_COUNT as a new measure of customer loyalty, defined as the length of time a customer has stayed with the company, by accumulating that customer's purchasing and renewing frequency. Equations 1 and 2 generally describe how we detect actual defection and calculate UPDATE_COUNT.

$$\text{CLASS} = \begin{cases} 0 & if \sum RENEW\_FLAG > 0 \\ 1 & if \sum RENEW\_FLAG = 0 \end{cases} \quad (1)$$

$$\text{UPDATE\_COUNT} = \begin{cases} \sum RENEW\_COUNT + (n_{rows} - 1) & if\ CLASS = 1 \\ \sum RENEW\_COUNT + (n_{rows} - 1) - 1 & if\ CLASS = 0 \end{cases} \quad (2)$$

In Eq. 2, UPDATE_COUNT is calculated by summing the frequency of renewals on each purchase and the total number of purchases excluding the first chase $(n_{rows} - 1)$. Since we are going to predict customer defection, we subtract one from the total length to yield data to be used for prediction.

The results can be used for prediction. Thus, the final features of the purchase and auto-renewal data that will be used for the prediction are UPDATE_COUNT, total payment by the customer (CC_PRODUCT_PRICE), use of optional service (OPT_FLAG), type of customer (personal or commercial) (ORG_FLAG), status of e-mail delivery (MAIL_STATUS), and the actual defection flag (CLASS).

The final web log data that will be used are UPDATE_COUNT, CC_PRODUCT_PRICE, OPT_FLAG, ORG_FLAG, MAIL_STATUS, operating system used in the gadget (OS), type of browser used (BROWSER), number of page views (PAGE_VIEW), product views (PRODUCT_VIEW), cart views (CART_VIEW), web visiting frequency (VISIT), number of orders the customer has made (ORDER), and CLASS.

5 Machine-Learning Process and Evaluation Criteria

Machine learning is executed in the form of a classifier using two algorithms: C4.5 Decision Tree and Random Forest. The advantage of classification using a decision tree is that it can be easily interpreted and is intuitively understandable. Moreover, it provides the ability to make prediction using very large data sets. The decision-tree algorithm selects the best feature for splitting a node based on a statistical measure. The widely used decision-tree algorithm ID3 uses information gain to select the attribute that will categorize the samples into individual classes [8]. However, ID3 does not allow attributes with continuous values, and there are some biases in measuring the information gain for attributes with many values. The successor to ID3, C4.5, overcomes these problems by creating a threshold to fit the continuous attributes and avoiding bias in the information gain by using normalization [9].

A random forest is a collection of unpruned decision trees with randomized selection at each split, and outputs the class that is the majority of classes output by individual trees [10]. Bagging enables random forest to improve prediction accuracy over a single decision tree. Moreover, random forest excels in characterizing and exploiting structure in high-dimensional data for classification and prediction [11]. However, the resulting model produced by random forest can be difficult to interpret. One key feature of the random-forest learning algorithm that is used in this study is a novel variable importance measure.

The machine-learning performance must be evaluated to ensure that the model was generated well. In order to assess classification performance, we calculate the following

performance criteria: accuracy, recall, precision, and F-measure. Based on the confusion matrix in Table 2, each evaluation criterion is calculated as follows.

- Overall accuracy is measured using the proportion of the total number of predictions that were correct, calculated by $\frac{a_{11}+a_{22}}{a_{11}+a_{12}+a_{21}+a_{22}}$.
- Precision or positive prediction value is calculated by $\frac{a_{11}}{a_{11}+a_{12}}$.
- Recall or true positive rate is calculated by $\frac{a_{11}}{a_{11}+a_{21}}$.
- F-measure or F-score is calculated by $\frac{2\,(Precision\ x\ Recall)}{Precision\ +Recall}$.

Table 2. Confusion matrix

		Predicted	
		Defect	Not defect
Actual	Defect	a_{11}	a_{12}
	Not defect	a_{21}	a_{22}

Table 3. Number of examples of the initial data set

	Purchase and renewal data		**Web log data**	
	Positive	Negative	Positive	Negative
Low price	273,339	117,748	5,694	4,401
Middle price	1,172,951	729,163	26,709	27,518
High price	386,872	274,437	8,947	13,867

Table 3 lists the total number of examples available from the initial data sets. Positive examples include customers who defect, and negative examples include those who remain. All the examples will be employed in building the prediction model using the C4.5 decision-tree algorithm, with 10-fold cross validation for data splitting to ensure that instances from the original dataset have the same chance of appearing in the training and testing set. However, when analyzing important variables using a random forest, only a subset of purchase and auto-renewal data samples with the same distribution as the initial data set are used due to its limitation in handling large amounts of data.

6 Experimental Results

6.1 Measuring Variable Importance Using Random Forest

For prediction, it is critical to determine the importance of variables in providing predictive accuracy. In this study, we used the variable-importance algorithm in a random forest to obtain the mean decrease in accuracy of each variable. The mean decrease in accuracy of a variable is the normalized difference in classification accuracy for out-of-bag data when the data for that variable is included as observed, and the

classification accuracy for out-of-bag data when the values of the variable in the out-of-bag data have been randomly permuted [11]. Higher values of mean decrease in accuracy indicate variables that are more important to the classification. Table 4 gives the number of samples used by a random forest to obtain the importance of each variable on each customer segment.

Table 4. Number of samples used in obtaining the variable importance using random forest

	Purchase and renewal data		**Web log data**	
	Positive	Negative	Positive	Negative
Low price	48,693	21,037	5,694	4,401
Middle price	43,216	26,784	26,709	27,518
High price	40,632	29,368	8,947	13,867

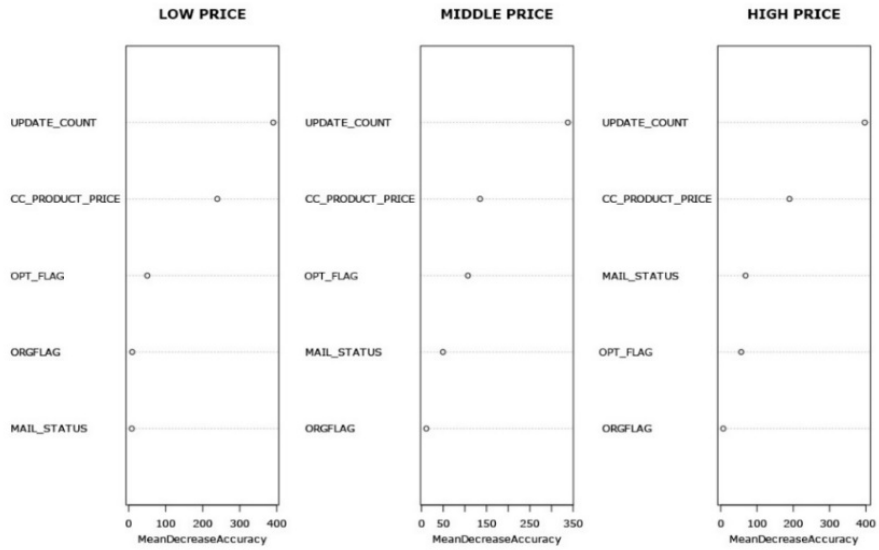

Fig. 1. Variable importance obtained using the purchase and auto-renewal data

For each of the three customer segments, UPDATE_COUNT was identified as the most important variable for classification using purchase and auto-renewal data (Fig. 1). Although we cannot say that variables identified as ''important'' are right or wrong, the results for a random forest coincide more closely with expectations based on understanding of customer loyalty. The more loyal the customer, described by their period of staying, the less probable their defection.

The results in Fig.2 indicate the variable importance obtained using web log data. Similar to the previous result, UPDATE_COUNT is the most important prediction variable in the Low Price customer segment. For the Middle Price and High Price customer segments, the total payment that each customer has made (CC_PRODUCT_PRICE)

appears to be the most important variable. However, there was consistency in the variables identified as the two most important using the entire data set: UPDATE_COUNT and CC_PRODUCT_PRICE.

Fig. 2. Variable importance for each customer segment obtained using the web log data

6.2 Prediction Model Using C4.5 Decision Tree

As previously mentioned, one advantage of using the decision tree classifier is convenience in interpreting the results. *R* package supports the process of interpretation by providing tree visualization and tree rules. Decision tree results make it easier for the company or other end user to determine the next action for retaining the customer based on defection prediction. We present an example of the visualization of customer defection prediction in the Low Price customer segment using purchase and auto-renewal data (Fig. 3) and an example of the rules of the defecting customer.

```
Rule number: 7 [RIHAN_FLAG=true cover=137787 (35%) prob=0.98]
   UPDATE_COUNT< 2.5
   CC_PRODUCT_PRICE< 4722
Rule number: 25 [RIHAN_FLAG=true cover=15697 (4%) prob=0.88]
   UPDATE_COUNT< 2.5
   CC_PRODUCT_PRICE>=4722
   UPDATE_COUNT>=0.5
   CC_PRODUCT_PRICE>=4972
Rule number: 13 [RIHAN_FLAG=true cover=70863 (18%) prob=0.84]
   UPDATE_COUNT< 2.5
   CC_PRODUCT_PRICE>=4722
   UPDATE_COUNT< 0.5
```

The rules on node 7 indicate that 35% of customers who have the attributes of UPDATE_COUNT less than 2.5 and make payment on CC_PRODUCT_PRICE less than 4,722 (JPY) have a 98% probability of defecting. Both visualization and rules indicate that the decision tree obtained a model that uses UPDATE_COUNT and total payment or CC_PRODUCT_PRICE is the most powerful predictor. Similarly, it occurs in all customer segments when we use purchase and auto-renewal data. Using web log data (Fig. 4), the status of e-mail delivery appears to be one of the three predictors resulting in predictive accuracy.

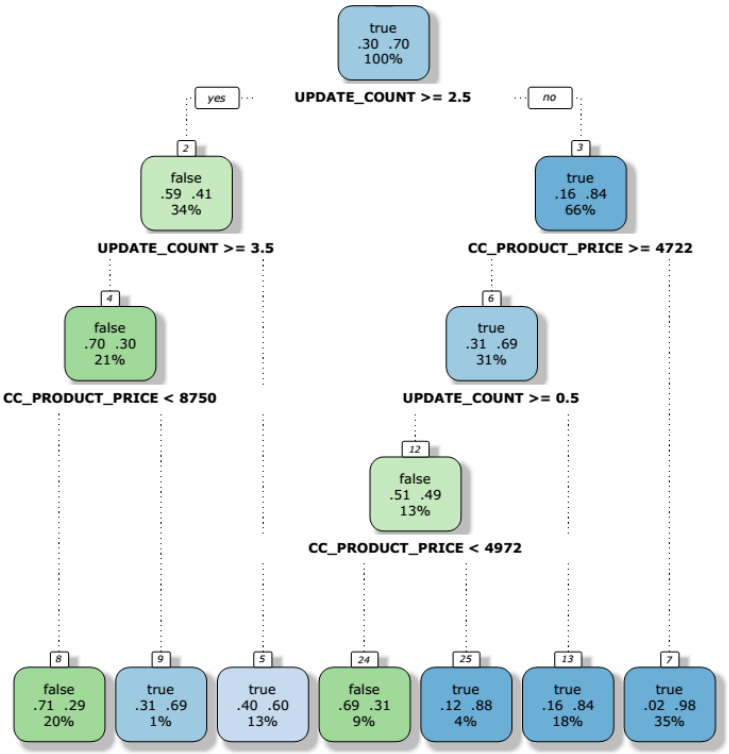

Fig. 3. Visualization of tree on Low Price customer segment based on purchase and auto-renewal data

Performance evaluation of the model on predicting defection using the C4.5 decision-tree algorithm is presented in Tables 4 and 5. The minimum object is set to 40, and the complexity of the tree is set to 0.005. The results indicate that using purchase and auto-renewal data, we can obtain a better prediction model of customer defection and conclude that the new features we acquired in data preparation are useful in predicting customer defection in the case company.

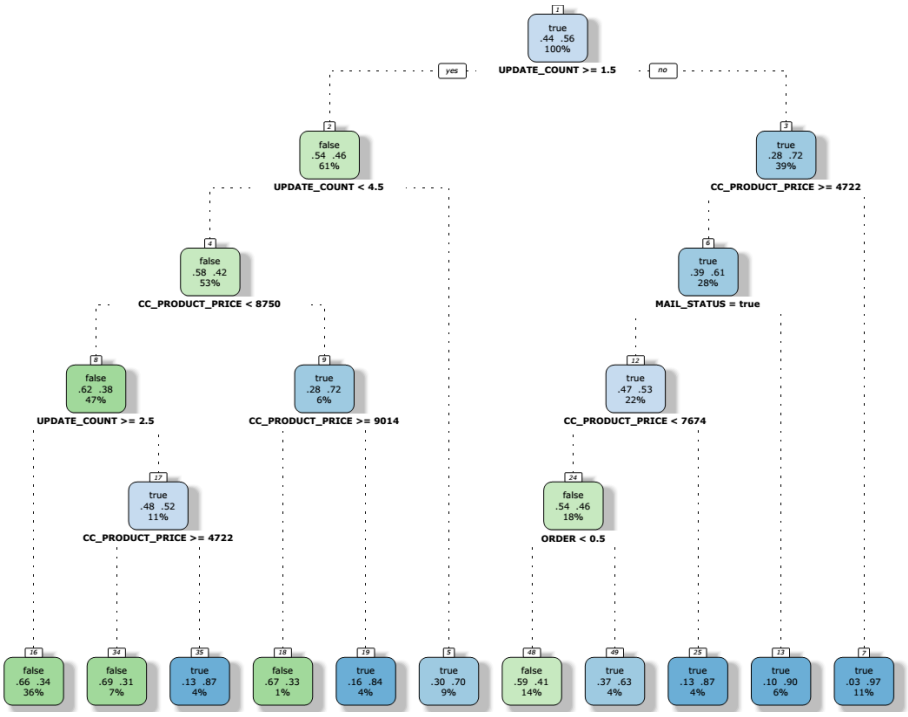

Fig. 4. Visualization of tree on Low Price customer segment based on web log data

Table 5. Predictive accuracy of C4.5 decision tree based on the purchase and auto-renewal data

Data set		Accuracy	Precision	Recall	F-score
Low Price	Baseline	69.7	69.6	49.1	50.5
	C4.5	**82.8**	82.6	82.9	82.7
Middle Price	Baseline	61.6	61.7	38.0	47.1
	C4.5	**72.4**	72.6	72.1	71.8
High Price	Baseline	58.5	58.2	33.9	42.8
	C4.5	**83.3**	82.8	73.8	82.9

Table 6. Predictive accuracy of C4.5 decision tree based on the web-log data

Data set		Accuracy	Precision	Recall	F-score
Low Price	Baseline	56.4	56.4	31.8	40.7
	C4.5	**72.7**	72.7	74.4	72.8
Middle Price	Baseline	50.7	50.7	25.8	34.2
	C4.5	**69.8**	69.8	70.7	69.6
High Price	Baseline	60.8	60.8	36.9	46.0
	C4.5	**73.1**	72.1	79.2	78.4

To summarize and clarify how our methods apply to the case company, we consider several questions and answers that are appropriate based on the experiment results.

1. *Which customers are defecting?*
   If "defecting" is defined clearly based on the original data sets, the answer to this question is straightforward to the number of "opt-outs" that appear in the database query and it is not factual. Thus, analysis using our method clearly indicates that customers who do not have renewal records are actually defecting from the company.
2. *What variables characterize the defecting customer?*
   Defecting customers are characterized mainly by their loyalty attributes: how long they stay with the company, how many purchases they make, and their auto-renewal activity. In addition, the number of total payments the customer has made represents the likelihood that the customer will defect.
3. *Can we determine what strategy may keep a customer from defecting?*
   Based on the previous questions and answers, we can determine what strategy in each customer segment may keep a customer from defecting. For example, since customer loyalty is the main characterizing variable, a company may direct more marketing campaigns toward customers with a low loyalty average.

7 Conclusion and Future Works

One key activity of customer defection management is predicting customer defection. This paper presents several procedures that contribute to resolving some novel problems in predicting defection. We provided an algorithm that is beneficial in determining which customer is truly defecting from the company and produced a new feature (UPDATE_COUNT) from the available data that can be used as a measurement of customer loyalty. Using machine learning, we then identified important variables for classifying defecting customers. Finally, we built a prediction model of customer defection using both purchase and auto-renewal data and web log data.

This study does not capture the dynamic of customer activity and characteristics. Thus, future work will seek to integrate machine learning with a more dynamic approach, such as agent-based modeling and simulation. Agent-based modeling will provide a computational model for simulating interactions between customers from the micro level to a macro level. In addition, machine learning can provide predictive accuracy regarding customer behavior that will be useful for validating the agent-based model.

References

1. Colombus, L.: Predicting Enterprise Cloud Computing Growth. Forbes (April 9, 2013), http://www.forbes.com/sites/louiscolumbus/2013/09/04/predicting-enterprise-cloud-computing-growth/ (accessed July 20, 2014)

2. Huang, B.Q., Kechadi, M.-T., Buckley, B.: Customer Churn Prediction for Broadband Internet Services. In: Pedersen, T.B., Mohania, M.K., Tjoa, A.M. (eds.) DaWaK 2009. LNCS, vol. 5691, pp. 229–243. Springer, Heidelberg (2009)
3. Wei, C., Chiu, I.: Turning telecommunications call detail to churn prediction: A data mining approach. Expert Systems with Applications 23, 103–112 (2002)
4. Yung, S., Yen, D., Wang, H.: Applying data mining to telecom churn management. Expert System with Applications 31, 515–524 (2006)
5. Tiwari, A., Roy, R., Hadden, J., Ruta, D.: Churn Prediction: Does Technology Matter. International Journal of Intelligent Systems and Technologies 1 (2006)
6. Prasasti, N., Ohwada, H.: Applicability of Machine-Learning Techniques in Predicting Customer Defection. In: International Symposium on Technology Management and Emerging Technologies (ISTMET 2014) (2014)
7. Prasasti, N., Okada, M., Kanamori, K., Ohwada, H.: Customer Lifetime Value and Defection Possibility Prediction Model using Machine Learning: An Application to a cloud-based Software Company. In: Nguyen, N.T., Attachoo, B., Trawiński, B., Somboonviwat, K. (eds.) ACIIDS 2014, Part II. LNCS (LNAI), vol. 8398, pp. 62–71. Springer, Heidelberg (2014)
8. Quinlan, J.: Induction of Decision Trees. Machine Learning 1, 81–106 (1986)
9. Xiong, Y., Syzmanski, D., Kihara, D.: Characterization and Prediction of Human Protein-Protein Interaction. In: Biological Data Mining and Its Applications in Healthcare, pp. 237–260 (2014)
10. Breiman, L.: Random Forests. Machine Learning 45, 25–32 (2001)
11. Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J.: Random Forest for Classification in Ecology. Ecology 88(11), 2783–2792 (2007)