Employee Churn Prediction Model using C4.5 Classification Algorithm

Nisrina Salma[1], Andry Alamsyah[2]

[1,2] Telkom University, Faculty of Economic and Business,

Abstract: ***Churn phenomenon commonly happens in customer problem and become jeopardy issues that any industries can suffer. Churn problem also can appear in organization, it is called employee churn. Employee churn is relatable with customer churn yet slightly distinct. Churn create a numerous adversely effects in the organization such as loss of employee can lead to unfairly distribution of workload, customer dissatisfaction, also company costs money and time for finding a replacement. Hence, it is important to know who, where, and why the employee is churning. Classification and prediction in data mining is implemented to predict the employee churn. Therefore, this research aims to present a case study that we present a study of C4.5 classifier algorithm for employee churn prediction model. In the prediction proposed model, the splitting of training and testing data distinguish into 2 different types of ratios. For dataset 1 the training dataset is 70% and testing dataset is 30%, while for dataset 2 training dataset is 80% and testing dataset is 20%. The classifier accuracy for dataset 1 and dataset 2 gains 94.8% and 95% respectively. Based on the accuracy level, C4.5 classifier is the proper method to predict employee churn.***

Keywords: employee churn, data mining, prediction model, classification**,** C4.5 algorithm

1. Introduction

The occurrence of churn frequently appears in unsteady consumer service markets such as subscription services, mobile phones, banking, and insurances [1]. Churn is defined as the tendency of customers to discontinue their subscription towards products or services and afterwards shift to other company within a given time period [2]. Customer churn is a notorious issue for most industries and become a serious concern for organizations [1, 2].

Employee churn has a similarity but slightly different with customer churn. Employee churn is the overall turnover, which refers to people leaving their jobs in an organization [1]. Churn implies to a cycle of continuous turnover. If organization has high turnover, they experience churn. Employee churn also can be called attrition [3]. Annually employee churn rates can be high as 12-15% and it is possible for the next year it might have a significant increase [1].

Churn leads to various adversely effects to organization [4]. Hence, it is crucial to know the number of future employee churn in the organization. More importantly, it is essential to know why, who, when, where churning is possibly happened and what are the possible reason for employee leaving the organization. Hence, it is crucial to know the future prediction of employee churn in an organization.

To estimate future churn, predictive technology is frequently applied. In present day, data mining techniques has given a great deal of attention in information industry and it has been recognized as a newly emerging analysis tool with great potential to help the company focus on the most important information in their data warehouse. Data mining technique is applicable to extract a large amount of data to discover the valuable and meaningful knowledge from it. Data mining extensively used in many areas, such as business and Human Resources, to find trend analysis and future planning [5, 6]. Thus, predictive technology is applied for employee churn modelling. Among the major techniques in data mining, classification and prediction technique are among the popular tasks in data mining. From all classification techniques, decision tree is the most popular techniques because its flexibility and easy to understand [7, 8]. The decision tree C4.5 is popular and widely used in supervised machine learning. The C4.5 classifier that was designed by Quinlan is extension of the basic ID3 algorithm to address the issues not dealt adequately by ID3 [6]. Behind its popularity decision tree C4.5 self-explanatory algorithm as the derived rules have a very straightforward interpretation [8, 9]. Those algorithms can predict future trends and behavior, enabling the company to be proactive, and encourage decision knowledge [5, 6, 7]. Due to these reasons, this study is aimed to seek the capability of C4.5 classifier for predicting employee churn.

In this study, recommendation for employee churn (transfer/not transferring) is considered as the target class in classification process. For employee dataset, we used employee's personal data from one of Indonesia Telecommunication Company as training and testing dataset. The first phase is data mining preprocessing task using the training dataset. In the second phase, the C4.5 classifier is used to generate employee churn prediction from 2 years period of time started from 2015 until 2017. The third step is the evolution measurement for C4.5 classifier.

2. Related Works and Theoretical Background

In this study, researcher presents based on two underlying theories which are employee churn and data mining. Therefore, in this chapter, researcher provides a linkage between how data mining can give an insight from employee data or Human Resource Information System (HRIS).

2.1 Human Resources Area

Jantan et al. [4] in their study presents a talent management prediction by using employee past performance score based on the performance appraisal standard in Malaysia. They suggested a classification and prediction techniques, which is C4.5 classifier. The experiment is one show the highest accuracy 95% of the total of test set samples that are correctly classified using full attributes. The second phase of experiment is to compare the accuracy of classifier for attribute reduction.

Jantan et al. [5] in their work presents a significance of the study using data mining for talent management especially for classification and prediction. They use 5 classification algorithms including Decision Tree C4.5, Random Forest, Neural Network Multi-Layer Perceptron (MLP) and Radial Basic Function Network, and Nearest Neighborhood K*Star. They divided the dataset into 3 set of datasets. The results for full attribute present the highest accuracy of model is C4.5 (95.14%, 99.90%, and 90.54%). In conclusion C4.5 classifier algorithm is the potential classifier in the study. This technique can be used for building an employee churn model in the next prediction phase e.g. classification rules construction.

2.2 Churn Problem

Khodabandehlou et al. [6] presents a comparison of supervised machine learning techniques for customer churn problem. They compare three different including Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Decision Tre. ANN method used, Multi-Layered Perception (MLP) Radial Basis Function (RBF) algorithm, to their reliability and practicality in churning prediction problems. In the SVM method, the polynomial and RBF cores are selected due to their high-quality results in related works. In total they use 5 different algorithms; ANN-MLP, ANN-RBF, SVM-Poly, SVM-RBF and Decision Tree C5.0. Those 5 algorithms considered as the best method for customer churn prediction.

Saradhi et al. [1] presents a study of an employee churn prediction model using SVMs, random forests, and Naïve Bayes method for predicting employee churn. They divided the dataset into 3 types of datasets. The accuracy results for 3 different method shows that random forest and Naïve Bayes. However, SVMs success over random forest and Naïve Bayes in true positives accuracy compared to other methods.

2.3 Decision Tree

Decision tree is a diagram of decision-making rules used to categorize subjects into a few groups or to make predictions [10]. Decision tree classification approach is a flow chart like a tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node (Han et al, 2006:291). The advantages of decision tree are easy interpretation and understanding for decision makers to compare with their domain knowledge for validation and justify their decision [11]. Decision trees can easily be converted to classification rules. In the figure 2.6 below, shows an example of decision tree.

2.4 C4.5 Algorithm

The evolution of decision tree algorithm has growth from time to time. A lot of improved algorithms have emerged. Before Quinlan invented C4.5 algorithm in 1986, the predecessor of this algorithm is ID3 algorithm that also invited by Quinlan in 1993 [12]. Information gain ratio and entropy are the selection creation of C4.5 algorithm in order to overcome the defect of the predecessor algorithm which is ID3 [6, 12]. C4.5 also introduce a pruning technology for the creation of the tree [12]. The advantage of C4.5 includes error pruning reduction, avoidance of overfitting data, the capability of handling continues data, and handling data with missing attribute values. The aim of C4.5 algorithm is to obtain the rule set. This rule set is obtained in the testing phase and applied to the whole pre-processed data [6].

3. Methodology

The aim of this study is to provide a descriptive model and prediction model of employee churn. The C4.5 classifier is used to generate the prediction model. Because its ability to breakdown the complex decision-making process become much simpler [13]. Thus, the study presents the capability of C4.5 classifier towards the prediction of employee churn. The flow of this research consists of 3 phases.

3.1 Pre-Processing

Data which is used in this research is an employee database from XYZ company during a two years period starting from 2015 until 2017 with the initial dataset contains 16.649 instances of employee personal data.

Preprocessing data which involved data cleaning, data integration, data selection, and data transformation. The activities also include removing non-significant data, data labelling, merge the data, and replacing date with discrete number. This is the first phase of the research.

3.2 Data Processing

The second phase is data processing. The aim of data processing is to construct descriptive model to seek the pattern of the dataset using 16.649 records. Afterwards, generate the classification model using C4.5 classifier for training dataset. The selected attributes for training dataset is shown in table I. In building the prediction model, the dataset is divided into two sets of data, training dataset and testing dataset. In selecting the ratio of the training data and the test data, there is no exact formulation or resources [8]. In this study, we use two ratios; (1) 70% for training data and 30% testing data, (2) 80% for training data and 20% testing data.

3.3 Evaluation Measurement

The last phase of the research is evaluation and interpretation of the model. To ensure the prediction model able to generalized is to evaluate its classifier performance [7]. For evaluating the accuracy or the performance of predictor model, there are notable parameters including accuracy, precision, recall, and the F-measure [8]. In order to evaluate the classifier performance, those parameters are used based on the confusion matrix shown in table 1 [5, 6]. In the confusion matrix job transfer is represent churners, while not transferring represent non-churners.

Table 1: Confusion Matrix

[redacted]		*Predicted*	
		Not transferring	*Transfer*
Actual	*Not Transferring*	True Positive (TP)	False Negative (FN)
	Transfer	False Positive (FP)	True Negative (TN)

1) TP (True Positives): the number of employee that should be in the not transferring category and the prediction algorithm determined the category correctly as not transferring.
2) TN (True Negatives): the number of employee that should be in the transfer category and the prediction algorithm has determined their category correctly as transfer.
3) FP (False Positives): the number of customers who are transfer but the algorithm incorrectly categorized them as not transferring.
4) FN (False Negatives): the number of customers who are not transferring but the algorithm incorrectly categorized them as transfer.

The C4.5 classifier calculates based on several parameters shown in table II. There are 5 parameters listed in table, those parameters are used to measure the capability of C4.5 classifier [9].

Table 2: Parameter Types

Parameter	*Description*	*Best Levels*
AUC	Area under the curve	≥0.85
CA	Accuracy classification score	1
F1	Weighted average of the precision and recall	1
Precession	How appropriate the model predicting the class	1
Recall	The number of true positive	1

4. Results and Analysis

The classification model generated from C4.5 classifier. This classifier can produce the decision tree and classification rules. The C4.5 classifier also constructs a tree for the purpose of improving the prediction accuracy [9]. The C4.5 classifier is among the powerful decision tree classifier. The C4.5 produce unequivocal result and interpretation [9]. Summary of analysis that derived from C4.5 classifier on the performance evaluation training dataset using several parameters.

As already stated in chapter 3, we use two ratios for training and data testing. Therefore, the dataset distinguishes into two datasets as shown in detail on table 3 The training dataset contains of data and testing dataset contains the remaining dataset or unseen data.

Table 3: Dataset Partitioning

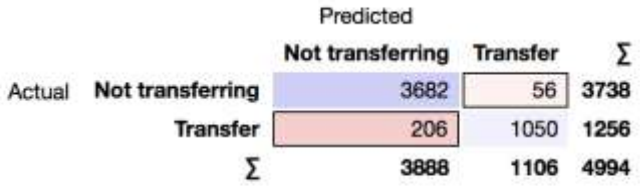
Dataset	*Training/Testing*	*Proportions*	*No. of Samples*
1	Training	70%	11.655
	Testing	30%	4.994
2	Training	80%	13.320
	Testing	20%	3.329

A classifier testing for both dataset is performs using parameter as shown in table II for training and testing dataset. Table 4 shows a comparison score of the parameters using both dataset 1 and dataset 2. The results generate 94.9% of accuracy on testing data for dataset 1, while dataset *2* generate 95% of accuracy on testing data.

Table 4: Comparison Results

Dataset 1			*Dataset 2*		
Parameter	*Trainin g Data*	*Test Data*	*Parameter*	*Trainin g Data*	*Test Data*
AUC	0.982	0.92 5	AUC	0.986	0.933
CA	0.974	0.94 8	CA	0.972	0.950
F1	0.945	0.88 5	F1	0.938	0.894
Precision	0.989	0.93 1	Precision	0.987	0.941
Recall	0.905	0.84 3	Recall	0.894	0.852

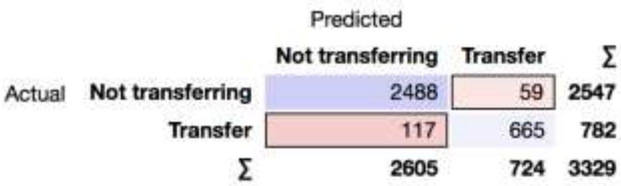
On top of that, confusion matrix for dataset 1 is presented in the form of table that shown in figure 1. From figure 1, it can be concluded the number of error generated by C4.5 classifier tested on testing dataset, when the actual class is transfer but predicted as not transferring is 206 instances, while for the actual class is not transferring but predicted as transfer is 56 instances.



		Predicted		
		Not transferring	Transfer	∑
Actual	Not transferring	3682	56	3738
	Transfer	206	1050	1256
	∑	3888	1106	4994

Figure 1: Confusion matrix (showing the proportion of missclassified) dataset 1

The confusion matrix for dataset 2 is shown in figure 2. It can be concluded that the number of error the number of error generated by C4.5 classifier tested on testing dataset, when the actual class is transfer but predicted as not transferring is 117 instances, while for the actual class is not transferring but predicted as transfer is 59 instances.



		Predicted		
		Not transferring	Transfer	Σ
Actual	Not transferring	2488	59	2547
	Transfer	117	665	782
	Σ	2605	724	3329

Figure 2: Confusion matrix (showing the proportion of missclassified) dataset 2

Table 5 shows the rules evaluation using both dataset. This result indicates the accuracy of the classification for the new dataset. The accuracy results indicate that using proportion ratio of dataset 2 have a slightly better accuracy than proportion ratio in dataset 1 although both of dataset have the accuracy >90%.

Table 5: C4.5 Classifier Evaluation

Dataset	*Status*	*Number of Data*	*Accuracy (%)*
1	Correctly classified	4.732	94.8%
	Incorrectly classified	262	
2	Correctly classified	2.509	95%
	Incorrectly classified	220	

5. Conclusions and Suggestions

Based on the experimental results obtained, emphasized that for prediction model C4.5 classifier algorithm successfully predicted future employee churn with highest number of accuracy 95%. The accuracy results indicate that using proportion ratio of dataset 2 have a slightly better accuracy than proportion ratio in dataset 1 although both of dataset have the accuracy >90%. Thus, based on the accuracy level we can conclude that C4.5 classifier is a one of proper method to predict employee churn.

For the future research, in order to get the best accuracy for employee churn a comparation of classification model is strongly recommend. Therefore, the researcher may know which model is the best model to predict employee churn.

References

[1] V. V. Saradhi and G. K. Palshikar, "Employee Churn Prediction," Expert System with Application, vol. 38, no. 3, pp. 1999-2006, 2011.

[2] V. Bhambri, "Data Mining as a Tool to Predict Churn Behavior of Customers," International Journal of Computers & Organization Trends, vol. 2, no. 3, pp. 85-89, 2012.

[3] P. S. Martins, "Working Churning and Firms' Wage Policies," International Journal of Manpower, vol. 29, no. 1, pp. 48-63, 2008.

[4] R. Punnoose and P. Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms (A case for Extreme Gradient Boosting)," International Journal of Advanced Research in Artificial Intelligence (IJARAI), vol. 5, no. 9, pp. 22-26, 2016.

[5] H. Jantan, A. R. Hamdan and Z. A. Othman, "Data Mining Classification Techniques for Human Talent Forecasting," P. K. Funatso, Ed., InTech, 2011.

[6] A. M. Florence.T and M. Savithri.R, "Talet Knowledge Management Acquisition using C4.5 Classification Algorithm," International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS), vol. 4, no. 406, p. 410, 2013.

[7] D. Eswaramurthy and S. Induja, "A Study on Customer Retention using Predictive Data Mining Tecniques," International Journal of Computer & Organization Trends, vol. 12, no. 1, pp. 48-63, 2014.

[8] S. Khodabandehlou and M. Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," Journal of Systems and Information Technology, vol. 19, no. 1/2, pp. 65-93, 2017.

[9] H. Jantan, A. R. Hamdan and Z. A. Othman, "Human Talent Prediction in HRM using C4.5 Classification Algorithm," International Journal on Computer Science and Engineering (IJCSE) , vol. 02, no. 08, pp. 2526-2534, 2010.

[10] E.-B. Lee, J. Kim and S.-G. Lee, "Predicting customer churn in mobile industry using data mining technology," Industrial Management & Data Systems, vol. 117, no. 1, pp. 90-109, 2017.

[11] Q. A. Al-Radaideh and E. Al Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performances," International Journal of Advanced Computer Science and Applications (IJACSA) , vol. 3, no. 2, pp. 144-151, 2012.

[12] S.L. Pang, J.Z. Gong ”C5.0 classification algorithm and application on individual credit evaluation of banks”, Systems Engineering—Theory & Practice, vol. 29, no. 12, pp. 94-104. 2009.

[13] H. M. Setiadi, C. Ariandika and A. Alamsyah, "Prediction Models Based on Flight Tickets and Hotel Rooms Data Sales for Recommendation System in Online Travel Agent Business," The 7 h Smart Collaboration for Business in Technology and Information Industry (SCBTII), vol. 6, 2016