Employee Churn Forecast Study Based on Clustering Analysis and Machine Learning Models

Thu Zar Htet
Faculty of Information Science
University of Computer Studies (Meiktila),
Myanmar
dawthuzarhtetisdept@gmail.com

Soe Kalayar Naing
Faculty of Computer Science
University of Computer Studies (Yangon),
Myanmar
soekalayarnaing@ucsy.edu.mm

***Abstract*— In recent years, company is necessary to forecast churn customers in order to retain customers. Customer attrition is a crucial problem that every business must make the utmost effort to prevent. Employee churn or loss of workers would be similar to customer attrition, but the effect of losing a significant client for the company would probably be more traumatic while the effects of finding good employees instead of lost employees, as well as the expense of in-service training that could be offered to new employees. This paper finds out the groups of employees who left by using k-mean clustering and then predict the employee churn by using machine learning models. Finally, this paper assesses the performance of employee churn prediction and then compares the result with high accuracy. The purpose of this study is to predict employee churn, which allows the company to take the required steps.**

***Keywords*— Employee churn, k-mean clustering, machine learning models**

I. INTRODUCTION

Churn is described as the tendency of customers to discontinue their subscription to one party's goods or services and then turn to another party within a given period of time. For most businesses, customer churn is a notorious problem and is of significant concern to organizations [1] [2]. Within the organization called employee churn, the churn issue can arise. In theory, employee churn and customer churn have a common concept. The total turnover is employee churn, which refers to individuals leaving their positions in a company. Attrition can also be called employee churn. In consumer loyalty to brand goods or services, the churn phenomenon generally occurs. In certain ways, it is tougher to subscribe an employee with niche expertise. It affects current employees' continuing work and their productivity. The cost of recruiting new workers as a substitute includes the cost of jobs and the cost of training.

In this paper, we use the personal data of the employees of the Indonesia's renowned telecommunications firm. This paper determines the classes of the workers who left by using k-mean clustering [3] and then uses the machine learning models such as support vector machine, random forest, logistic regression and gradient boosting tree to estimate the employee churn [4, 5, 6, 7]. Finally, this paper compares the performance of the models and then it is found the method has a highest accuracy.

This paper is illustrated as follows. Section 2 describes the related research about different techniques or approaches implemented in relevant fields. Section 3 explores the employees churn prediction using machine learning models. Section 4 provides the model evaluation and the conclusion is in Section 5.

II. RELATED WORK

Most of the researchers' are attracted with enormous attention to predict churn such as the customer churn prediction. Verbeke et.al. proposes a benefit centered performance measure [10].To optimize the churn prediction problem with a decision support system [11] was studied by Coussement and Van den Poel. They found that it resulted in a substantial increase in predictive output by cooperating unstructured, textual information into a traditional churn prediction model. Wei and Chiu suggest the telecommunications customers churn forest in a related study by examining customer call details [12]. Coussement and Van den Poel apply the support vector machine method [13] for customer churn prediction.

The churn prediction research is widely discussed in the literature. In addition, there are few literature studies of the employee churn estimation in research area. By applying the Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest Method [14], Saradhi and Palshikar research workers expect churn. In another report, Khare et al. proposed the formula to predict the worker churn by using logistic regression [15]. Kane-Sellers' study on the Fortune 500 North American industrial automation manufacturer's sales force dataset is the last sample [16]. Kane-Sellers is used as the logistic regression as the main method. The literature calls for further focus on the part of researchers in this field, provided that the employee churn and the customer churn are not equal when comparing the results, and then they finds out the cost of the employee churn are even higher than the customer churns in some organizations. Therefore, as distinct from the above studies, we are trying not only to find out the groups of employees who left by using clustering analysis, but also to compare a number of classification models for the employee churn forest.

III. EMPLOYEE CHURN PREDICTION

Employee churn is the net turnover of the workers of a company as former employees depart and new ones are recruited. The churn rate is generally measured as the percentage over a given specified period of time of employees leaving the company. While some employee turnover is unavoidable, it is expensive to have a high degree of churn.

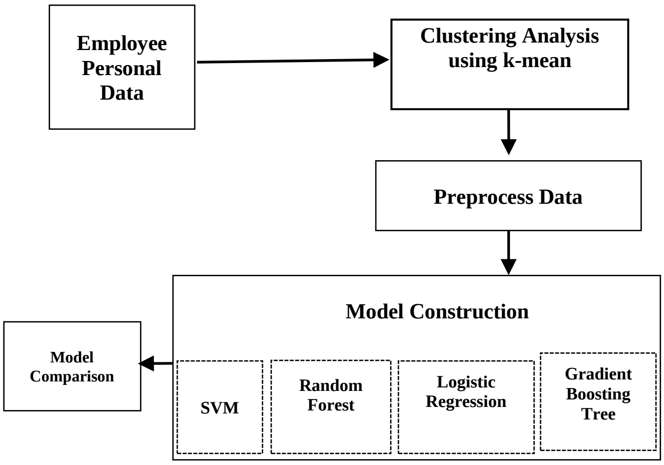

Figure 3.1: Research Stages for Employee Churn

The employee and customer churn are mentioned in the following points:

- Selects an employee from business to hire someone when who don't get to pick customers in marketing.
- Employees will be the face of business, and workers create all business does collectively.
- Revenues and brand value are impacted by losing a customer. Compared to maintaining the current customer, it is difficult and expensive to gain new customers. Employee churn is also agonizing for corporations and organizations. Finding and preparing a replacement takes time and effort.

Firstly, this paper accepts the employee personal data in Indonesia's renowned telecommunications firm and finds out the groups of employees who left by using k-mean clustering and then preprocess these data to improve data quality. After preprocessing these phase, this paper uses the machine learning models to predict the employee churn prediction that leave or not from organization in the future. And then calculates the performance of the models using confusion matrix [8]. Finally, this paper compares the performance of machine learning models and the stages are shown in Figure 3.1.

IV. Model Evaluation

4.1 Personal Record of Employee

This paper uses the employee data from Indonesia's renowned telecommunications firm [9]. Employee personal data has 10 parameters such as satisfaction level, last evaluation, number of projects, average monthly hours, time spend company, work accident, promotion last 5 years, department, salary, and left. The data include 14,999 samples of employee personal records.

4.2 Data Visualization

In the figure 4.1, we will see that about 3,571 were left out of 14,999, and 11,428 remained. The number of employee left is 23% of the total employment.

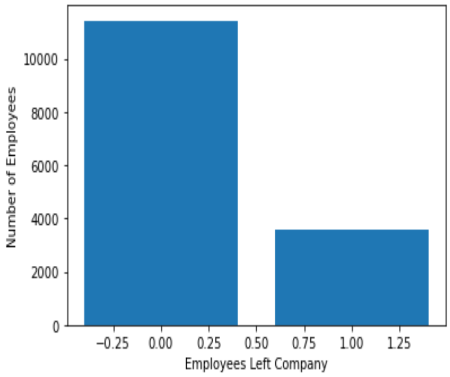

Figure 4.1: Employee Left

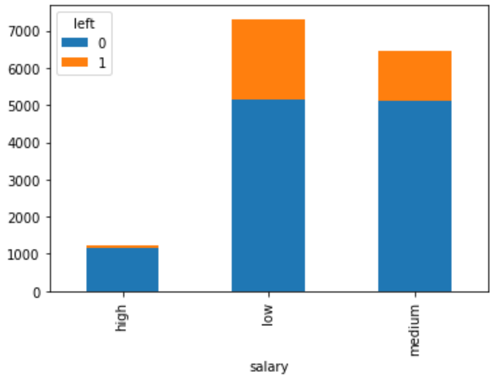

Figure 4.2: Employee Leaving from the Company with Salary

Figure 4.2 shows that whether the employee has left the company or not. i.e. 1 for left, 0 for another. From production in above figure 4.2, it appears that the highest ratio of the employee leaving the company is 30% of the low salary employee. Performance indicates that 20% of the medium salary employees are likely to leave the bank while in high salary employee is 7%.

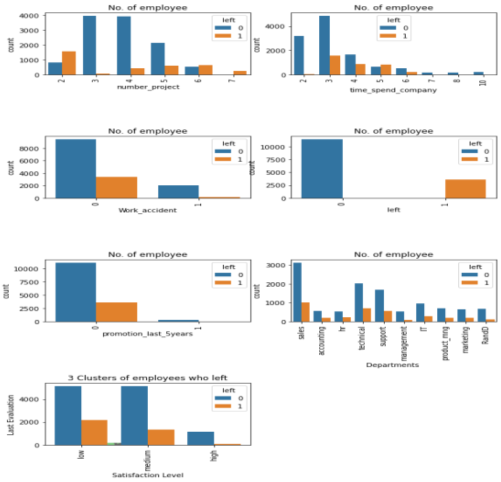

Figure 4.3: Subplots using Seaborn

In the figure 4.3, the following points can observe:

- Those employees who have more than 5 projects left the company.
- The employee who did 6 and 7 projects left the company seems to have been overwhelmed with jobs.
- Workers with 5 years of experience are leaving more because of no raises in the last 5 years and more than 6 years of experience are not leaving because of a company's love.

- All who have not left the promotion in the last 5 years, i.e. all those who have left, have not earned the promotion in the previous 5 years.

4.3 Cluster Analysis

This paper find out the groups of employees who left by using k-mean clustering. This analysis result is shown in below figure 4.4. It is possible to divide workers that have left the company into three categories of employees:

(1) **Winners** can also be called High Satisfication and High Evaluation (shaded in the graph by blue color).
(2) **Frustrated** can also be called Low Satisfication and High Evaluation (shaded in the graph by gray color).
(3) Moderate Satisfication and Moderate Evaluation can also be called '**bad match**' (shaded by green color in the graph).

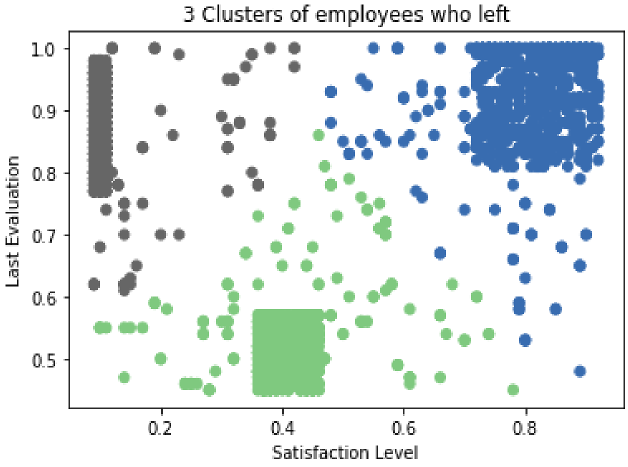

Figure 4.4: Group of Employee who left using k-mean

4.4 Data Preprocessing

Data preprocessing process performs data into consistent data. Numerical input data are required in machine learning algorithms. So, we transform categorical data to numerical data.

TABLE 4.1: DATASET PARTITIONING

Training/Testing	Proportion	umber of Samples
Training	70	10499
Testing	30	4500

In order to encode the personal data of the employees of Indonesia's renowned telecommunication company, that created labeled encoder object for salary and department columns. E.g. the meaning of the salary column may be depicted as low: 0, medium: 1 and high: 2. And then split the data into training and testing data that is shown in table 4.1. This implies that 70 % of the data will be used for model training and 30% for model testing. In data preprocessing step, all the models required the same format of data.

4.5 Model Performance

This paper predicts the employee churn by using machine learning models. Machine learning has recently become a common optional to solve classification and regression problems.

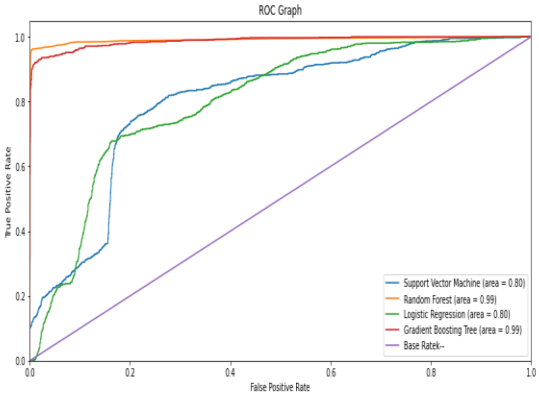

Figure 4.5: ROC Curves

The ROC curve is also a good indicator for selecting the best model. AUC stands for region under the curve, and the larger the model is, the better. We can visualize each ROC curve by adding the ROC curve node is provided in figure 4.5.

TABLE 4.2: MODEL EVALUATION

Model	Accuracy	Precision	Recall	F1	AUC
Support Vector Machine	0.785	1.000	0.096	0.175	0.799
Random Forest	0.988	0.991	0.956	0.973	0.991
Logistic Regression	0.760	0.494	0.238	0.321	0.803
Gradient Boosting Tree	0.972	0.958	0.921	0.939	0.988

In the above table 4.2, we have tested four models and the result has shown that the random forest performs best with an **accuracy** of 98.8%. **Precision**: Precision is to provide precise. In other words, the correct prediction results of the model. In prediction case, when the random forest model predicted the employee is going to leave that left 99.1% of the time. **Recall**: If there is an employee who left present in the test set and the random forest model can identify it 95.6% of the time. **F1 score:** this offers a balance between accuracy and recall, and is the test to be used if the sample is unbalanced. Random Forest has the best F1 score of 97.3%.

V. CONCLUSION

Employee churn result is important for financial, time and effort loss of organization. Through the use of the customer churn prediction model, potential churners can be detected in the organization and, as a result, the employee can take some measure to deter them from leaving. The purpose of this paper is twofold. First, we discussed the employee churn problem and cluster analysis that has been used to build employee churn model. Second, we presented a case study that the demonstration that machine learning techniques such as support vector machine, random forest, logistic regression, and gradient boosting tree can be used to predict models for the employee churn. Based on the experimental results obtained, it was emphasized that the best model for the churn forecast is random forest with the highest accuracy of 98.8%. The second best approach gradient boosting tree is 97.2%. The classification models such as support vector machine and logistic regression are the lowest accuracy. To improve predictive models and increase trust employees to these prediction projects, several features of employees should be examined. As the future direction, robust and universal models create as a potential path that the company can use for the benefit of the employee, cost efficiency and future prospects.

ACKNOWLEDGEMENT

I would like to express my special thanks to **all my teachers** who gave me their time and guidance, and all my friends who helped in the task of developing this paper. Finally, I would like especially to thank **my parents** for their continuous support and encouragement throughout my whole life.

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] Youguo Li, Haiyan Wu, “A Clustering Method Based on K-Means Algorithm,” International Conference on Solid State Devices and Materials Science, 2012.

[4] Weston. J., Watkins. C., “Support vector machines for multi-class pattern recognition,” In: Proc. Seventh European Symposium on Artificial Neural Networks, pp. 219-224, 1999.

[5] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[6]G.L.Nie,R.Wei,L.L.Zhang,et al, “Credit card churn forecasting by logistic regression and decision tree,” Expert Systems with Applications,vol.38,pp.15273- 15285,2011.

[7] Alexey Natekin and Alois Knoll, “Gradient boosting machines, a tutorial”, 2013.

[8] 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. ½, pp. 65-93, 2017.

[9] Employee data of Indonesia’s renowned telecommunications firm, https://www.kaggle.com/ liujiaqi/hr-comma-sepcsv

[10] W. Verbeke, K Dejaeger, D. Martens, J. Hur, and B. Baesens, “New insights into churn prediction in the telecommunication sector: A profit driven data mining approach”, European Journal of Operational Research, vol. 218, no. 1, pp. 211-229, 2012.

[11] K. Coussement and D.Van den Poel, “Integrating the voice of customers through call center emails into a decision support system for churn prediction”, Information & Management, vol. 45, no. 3, pp. 164-174, 2008.

[12] C. P. Wei and I. T Chiu, “Turning telecommunications call details to churn prediction: a data mining approach”, Expert systems with applications, vol. 23, no. 2, pp. 103-112, 2002.

[13] K. Coussement and D. Van den Poel, “Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques:, Expert systems with applications, vol. 34, no. 1, pp. 313-327, 2008.

[14] V. V. Saradhi and G. K. Palshikrar, “Employee churn prediction”, Expert Systems with Applications, vol.38, no. 3, pp. 1999-2006, 2011.

[15] R. Khare, D. Kaloya, C. K.Choudhary, and G. Gupta, “Employee attrition risk assessment using logistic regression analysis”

[16] M. L. Kane-Sellers, “Predictive models of employee voluntary turnover in a North American professional sales force using data-mining analysis.”, PhD thesis, Texas A&M University, 2007.