A PREDICTIVE MODEL FOR IMPROVING EMPLOYEE ATTRITION RATE WITH K-NEAREST NEIGHBOR CLASSIFIER

Tsehay Admassu Assegie
Aksum University, Aksum, Ethiopia

ABSTRACT

The strategic success of an organization vitally depends on attracting and retaining top talented employees within organization for a longer time. However, retaining employee is a challenging task. This is because organization should determine the factors causing employee loss in order to avoid the problems causing employee loss in organization. This research focuses on investigation of the major factors that cause employee loss and develop a predictive model that support the decision making process of human resource management executive to understand and improve the determinant factor for employee loss and reduce the attrition rate. The research proposes K-nearest neighbors based model for employee attrition prediction that supports data driven decision making of organization to explore the factors causing employee lose and reduce the employee loss within organization. Finally, we have analyzed the performance of the proposed model with experimental test and result appears to prove that the model is effective on predicting the employee attrition. Overall, the model has performed with a predictive accuracy of 86.7%.

Keywords: *KNN, employee attrition, employee attrition prediction, predictive modeling, data analytics*.

1. INTRODUCTION

Employee attrition refers to the number of workers who leave an organization and employees replaced by new employees [1]. A high rate of employee attrition in an organization leads to an increased recruitment, hiring of employees to replace the vacant position due to employee loss. However, recruitment and hiring of new employees is challenging. Because qualified and competent replacements are hard to find. In recent years, employee attrition have become the major problem in many organizations [2-3]. Hence, human resource (HR) managers are required to identify the root causes for employee loss and take corrective actions or appropriate modifications for ensuring that the attrition rate decreases. This research focuses on exploring the answers to the following questions:

1) What is the major cause for employee loss at organization?
2) What is the relationship between age and employee attrition?
3) How can we decrease employee attrition rate of organization?
4) What is the accuracy of K-nearest neighbor for employee attrition prediction?

2. LITERATURE REVIEW

There has been many research works undertaken on reducing employee loss at organization by lowering the attrition rate. However, researchersrarely applied machine learning for decision support to deal with employee loss by reducing the attrition rate. This research focuses on reviewing the existing researches that applied machine learning for employee attrition prediction and formulate new hypothesis for testing. Moreover, this work compares the existing methods and machine learning algorithms for employee attrition prediction. In addition to that, this work validates the existing works, by analyzing whether the conclusions made by the researchers are sound and evaluates the results of prior works. Some of the existing researches [4-15] applied machine learning for employee attrition prediction. In [4], the authors applied XGBoost algorithm for employee attrition prediction. The proposed predictive model predicts whether an employee will leave or continues to work in organization with predictive accuracy of 90%. The result is promising however; there is still a room for improvement for effectively employee attrition prediction. In other study [5], the authors reviewed the application of supervised learning algorithms such as decision tree, random forest. Logistic regression, support vector machine and K-nearest neighbor for employee attrition prediction. Experimental result shows that logistic regression performed better as compared to decision tree, random forest, support vector machine and K-nearest neighbor. In [6], the authors compared the performance of K-nearest neighbor and support vector machine for employee attrition prediction and comparative result shows that the K-nearest neighbor performed better as compared to the support vector machine. The K-nearest neighbor has predictive performance of 83.74% for employee attrition prediction. A comparative study [7] on the performance of random forest and support vector machine for employee attrition prediction shows that the support vector machine has better performance compared to the random forest. Another comparison study [8], on random forest, support vector machine and K-nearest neighbor show that the K-nearest neighbor performed better compared to random forest and support vector machine in terms of

predictive accuracy. In [9], extreme gradient boosting (XGboost) based employee turnover prediction model is proposed for improving the employee retention by making better decisions. The model is tested on predicting employee attrition and result shows a promising result although there is a scope for improving the accuracy of the model.

3. RESEARCH METHODOLOGY

This research employeed K-nearest neighbor (KNN) for developing a predictive model to estimate employee attrition. The data used for predictive modeling is acquired from online kaggle employee attrition data repository. The Pearson correlation for conducting the investigation of factors for employee attrition through exploratory factor analysis investigation of correlation among employee dataset features.

3.1. Correlation Model

The relationship among employee attrition dataset features is explored with Pearson correlation coefficient determined by a formula given in equation (1).

$$r = N\sum xy - (\sum x)(\sum y)/\sqrt{\sum x^2 - (\sum x)^2}(N\sum y^2 - (\sum y)^2) \quad \text{.....(1)}$$

Where: N= The number of features in emoployee attrtion dataset.

$\sum XY$= The sum of the products of of the features in emoployee attrtion dataset.

$\sum X$= Sum of emoployee attrtion dataset feature X.

$\sum Y$= Sum of emoployee attrtion dataset feature Y.

$\sum X^2$= Sum of squared emoployee attrtion dataset feature X.

$\sum Y^2$= Sum of sqaured emoployee attrtion dataset feature Y.

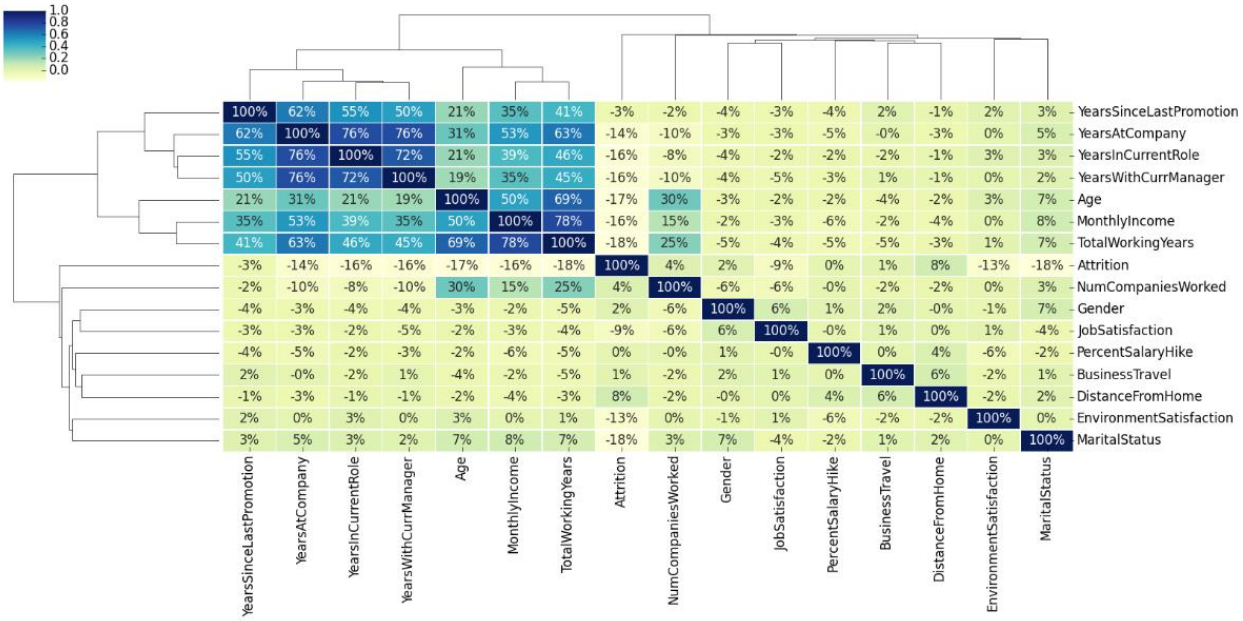

Figure 1. Employee attrition dataset features correlation

The correlation analysis shows that the employee attrition is strong correlated to the distance from home. This implies that the distance of an organization from the employee home has strong correlation. The number of organizations the employee has worked and gender and number of business travel has strong relationship with employee attrition.

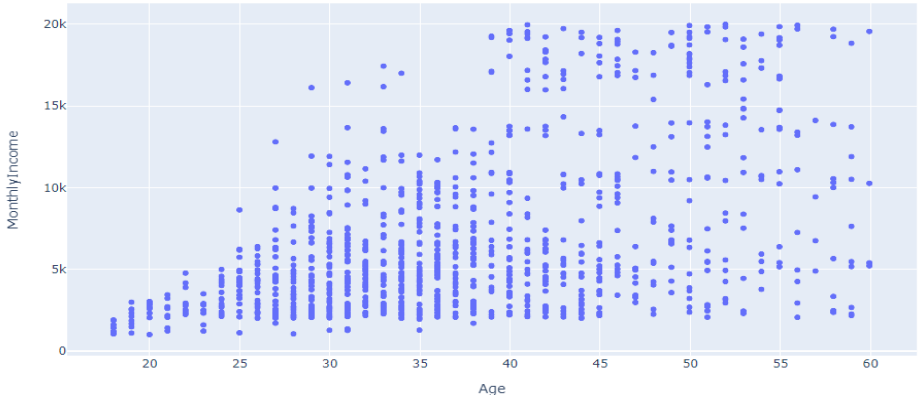

Figure 2. Moderate positive correlation between age and monthly income

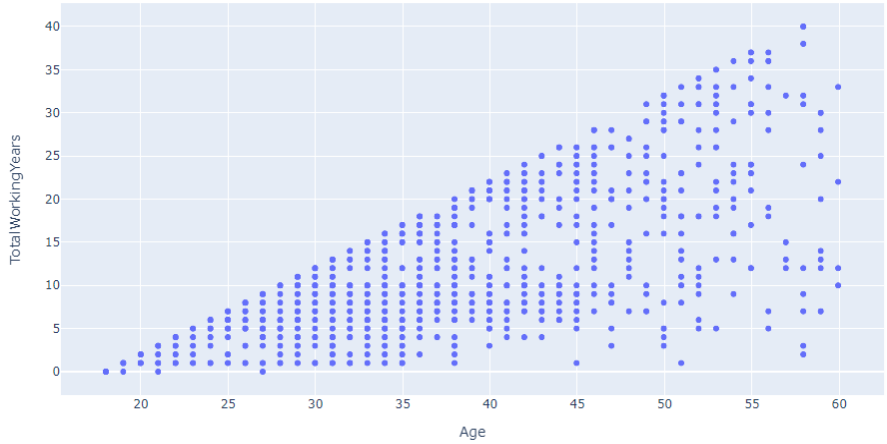

Figure 3. Strong positive correlation between age and total working years

3.2. Dataset Description

The employee attrition dataset has 1058 observations of employees leaving an organization and not leaving an organization. Each observation in the dataset has 16 features describing them as leaving or staying at an organization. Table 1 describers each of employee dataset feature employed in this study.

Table 1. Employee attrition dataset feature description

Observation No.	Feature	Description
1	Age	The age of employee
2	Attrition	(0=continues to work, 1=employee leaving organization)
3	Business travel	(1=No Travel, 2=Travel Frequently, 3=Travel Rarely)
4	Distance From Home	The distance from work to home
5	Environment Satisfaction	Satisfaction with the environment (1 'Low' 2 'Medium' 3 'High' 4 'Very High')
6	Gender	(0=Female, 1=Male)
7	Job Satisfaction	Satisfaction with the job (1 'Low' 2 'Medium' 3 'High' 4 'Very High')
8	Marital status	(2=Divorced, 1=Married, 0=Single)
9	Monthly Income	Monthly salary
10	Number companies Worked	No. of companies worked at
11	Percent Salary Hike	Percentage increase in salary
12	Total Working Years	Total years worked
13	Years at company	Total Number of years at the Company

14	Years In Current Role	Years in current role
15	Years Since Last Promotion	Number of years
16	Years With Current Manager	Years Spent with current manager

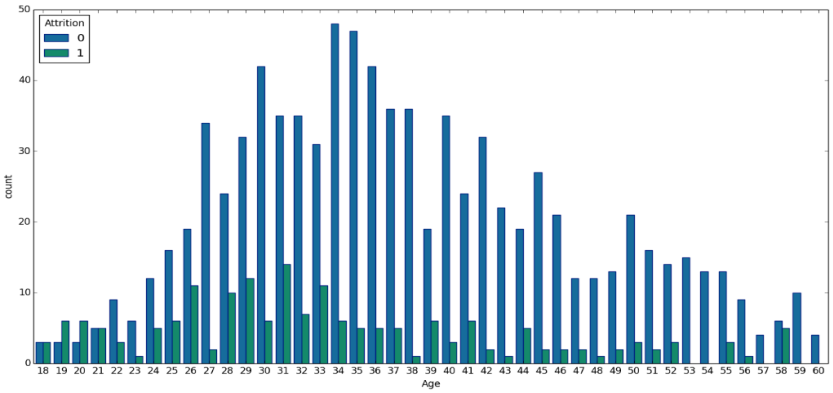

Figure 4. Age vs employee atrrition rate

As shown in figure 4, the attrition rate of employee is higher for age between 26 and 45 years. The employee attrition is lower for younger employees and older employee above the age of 45 years as shown in figure 4. The highest employee attrition rate is at the age of 34. Algorithm 1 shows the procedure for employee attrition prediction with K-nearest neighbor as follows:

Algorithm: 1

1. **Input**: Employee attrition features {f=X1, X2,....X16}
2. **Output**: Attrtion label {y=1 if employee leaves organization and y=0 if employee continues to work
3. Split employee dataset {X=traning set and y=test set}
4. Knn_model← initialize the KNN model
5. fit knn_model on X
6. test Knn_model on y

4. RESULT ANALYSIS AND DICUSSIONS

The experimental evaluation of the performance of the proposed model on employee attrition prediction shows a promising result. The predictive accuracy, receiver operating characteristics curve (ROC) are performance metric employed to evaluate the proposed model‘s effectiveness for employee attrition prediction for improving employee retention at organization by making better decision based on the results.

4.1. Accuracy vs K values of the proposed model

The performance of the proposed model for different values of K, is demsontrated in figure 5. as demonstrated in figure 5, the accuracy of the KNN model tends to decsrease with an increase in K values and the highest percoemance is achived with K=3 for the K valaues comapred in the range 1 to 15 as demosntrated in figure 5.

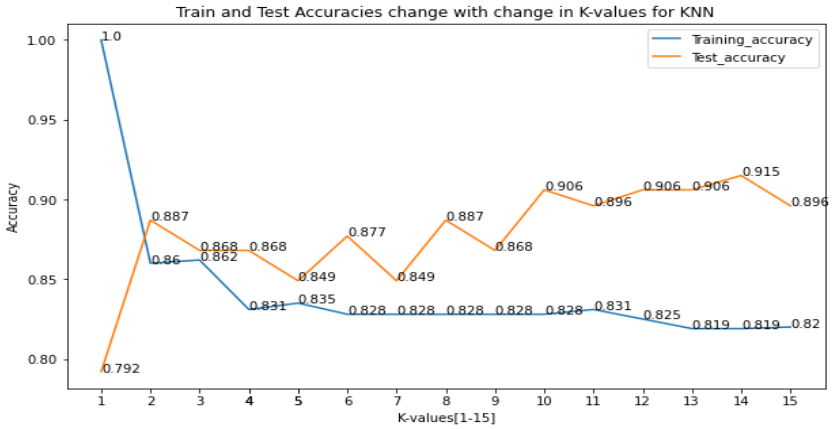

Figure 5. Accuracy vs K-value

Figure 5 shows the accuracies for different values of K ranging from 1 through 15. As shown in figure 3, the highest accuracy score for K-Nearest neighbor (KNN) is achieved when K-value of 2 is used for training the model.

4.2. Receiver Operating Characteristic Curve (ROC)

To analyze the behavior of the proposed model on predicting the employee attrition classes (employee leaving organization and employee continuing to work in organization) the receiver operating curve of the model is demonstrated in figure 6.

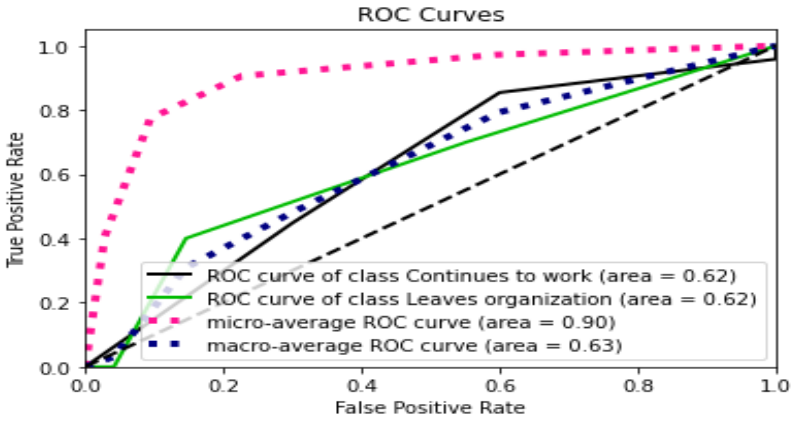

Figure 6. ROC curve for the proposed model

Figure 6, shows the receiver operating characteristic curve (ROC) for the proposed model. As shown in figure 6, the performance of the proposed model on employee attrition of both class (employee leaving organization and continues to work at organization) shows that the model has acceptable performance on predicting both classes.

```
True employee attrition observations: [0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0]
 Predicted employee attrtions       : [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
```

The predictive performance of the model is tested on employee attrition observation and the model predicted six observations incorrectly out of the total 30 observations. The result reveals that all of the incorrectly predicted observations are false negative (FN) employees actually leave the organization but the model predicted as employee that continues to work within organization.

4.3. Precision Recall Curve Analysis

The precision recall curve demonstrates the effectiveness of the K-nearest neighbor classifier on positive and negative class. The precision recall metric shows the effectiveness of the model on predicting employee leaving organization (positive class) and employee that continue to work in organization (negative class). As demonstrated in figure 4, the precision recall curve of the class continues to work is greater than the class leaves organization, which means the model classifies most of the observations as positive class (employee that continues to work in organization).

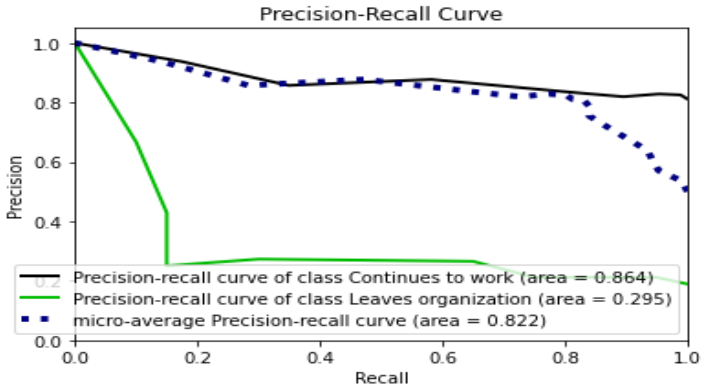

Figure 7. Precsion recall curve

4.4. Reliability Curve Analysis

The reliability calibration shows the probability associated with the predictive performance of the proposed model. The probability curve demonstrates how well the proposed model is calibrated and analyze the confidence on the prediction of employee attrition.

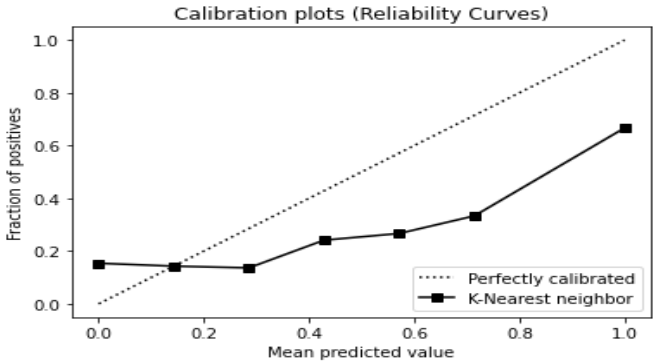

Figure 8. Relaiabiity curve

4.4. Cumulative Curve Analysis

The cumulative gain curve demonstrates the effectiveness of the proposed model on the employee attrition prediction. Figure 9 shows that the model's performance tends to increase with an increase in percentage of training sample.

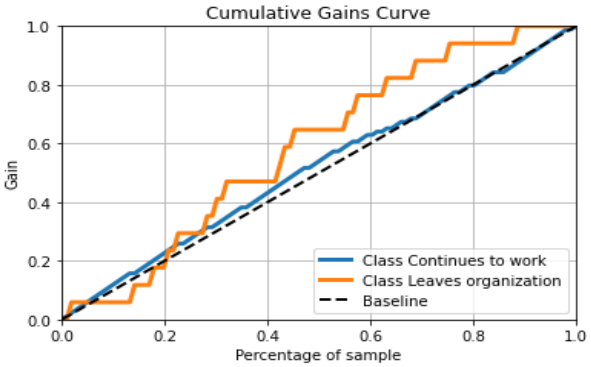

Figure 9. Cumulative gain curve

4.5. Confusion matrix

The employee attrition prediction by the model and real observations of employee attrition are compared in table 2.

Table 1. Confusion matrix for the proposed model

Observation	Predicted (that an employee will leave organization)	Predicted (that an employee will continue to work in organization)
Actually leaves organization	True Positive (TP):91	False negative (FN):1
Actually did not leave organization	False Positive (FP):13	True Negative (TN):1

The confusion matrix shown in table 1 illustrates, employee attrition problem with employee attrition record to predict which employee are likely to continue working in an organization. The percentage of correct prediction, the accuracy of the model is calculated with the formula shown in equation (2).

$$\text{Accuracy Score} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100 \dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

Hence, by substituting TP=91, TN=1, FP=13, FN=1 into equation (1), the accuracy of the proposed model can be determined as follows:

$$\text{Accuracy Score} = \frac{91 + 1}{91 + 1 + 13 + 1} * 100 = 86.7\%$$

5. CONCLUSIONS

Predictive modeling have become an important part of almost every data driven decision-making processes. In this research we have proposed K-Nearest Neighbor classifier based model for employee attrition prediction in organization. The effectiveness of the proposed model is tested and the result is analyzed. The experimental test on the model shows that the model performed with 86.7% accuracy. Overall, the experimental test result with performance measures such as receiver operating characteristic curve (ROC) and predictive accuracy appears to prove that the proposed model is effective and maximizes employee retention in organization. The proposed model is significant to support the decision making process as such can be effectively used for improving employee retention in organization.

6. ACKNOWLEDGEMENTS

I would like to express my gratitude for my lovely wife Alemtsehay Belay for her support in editing this manuscript.

7. REFERENCES

[1]. Abhiroop Nandi Ray, Judhajit Sanyal, Machine Learning Based Attrition Prediction, Global Conference for Advancement in Technology, IEEE, (2019).
[2]. Sandeep Yadav, Aman Jain, Deepti Singh, Early Prediction of Employee Attrition using Data Mining Techniques, IEEE, (2018).
[3]. Shawni Dutta, Samir Kumar Bandyopadhyay, Employee attrition prediction using neural network cross validation method, International Journal of Commerce and Management Research, (2020).
[4]. Rachna Jain, Anand Nayyar, Predicting Employee Attrition using XGBoost Machine Learning Approach, International Conference on System Modeling & Advancement in Research Trends, IEEE, (2018).
[5]. Yue Zhao, Maciej K. Hryniewicki, Francesca Cheng, Boyang Fu, Xiaoyu Zhu, Employee Turnover Prediction with Machine Learning: A Reliable Approach, Springer Nature Switzerland AG 2019.
[6]. Sarah S. Alduayj, Kashif Rajpoot, Predicting Employee Attrition using Machine Learning, 13th International Conference on Innovations in Information Technology (IIT), IEEE, (2018).
[7]. Rohit Hebbar A, Rajeshwari S.B, Sanath H Patil, S S M Saqquaf, Comparison of Machine Learning Techniques to Predict the Attrition Rate of the Employees, International Conference on Recent Trends in Electronics, Information & Communication Technology, IEEE, (2018).
[8]. Usha.P.M , N.V.Balaji, An Analysis of the Use of Machine Learning for Employee Attrition Prediction – A Literature Review, Journal of Information and Computational Science, (2020).
[9]. Rohit Punnoose, Pankaj Aji, Prediction of Employee Turnover in Organizations using Machine Learning Algorithms, International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 9, (2016)