Employee Resignation prediction using Machine Learning Techniques

Mubassir Habib
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
Mubassir.Habib@northsouth.edu

Hosnara Happy
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
Hosnara.happy@northsouth.edu

Sifat Momen
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
Sifat.momen@northsouth.edu

***Abstract*—In this paper, machine learning approaches have been used to predict the resignation of employees. This shall help companies to be able to pinpoint the key reasons, why employees usually explore other opportunities. If the company can predict the issue of someone leaving then they can save time by investing in the right employees in a company who are willing to stay in the company**

Keywords—artificial intelligence, machine learning, employee resignation, classification, resignation prediction

I. INTRODUCTION

Employees are unquestionably a company's most valuable asset. Their labor is the reason behind the successful firms run smoothly in a well-organized manner.

In the era of the Large Resignation, employee retention schemes have been a significant topic of concern for many firms. Workers desiring more flexible working conditions, greater compensation and benefits, and professional progression have characterized the workforce.

According to our latest departure process study, roughly 85 percent of Asian employers think they can detect when someone on their team is going to quit. They think that if an employee is distracted, disengaged, less productive, and frequently absent, they may predict when he or she is considering quitting. [1]

In order to help the Companies to prevent employee unhappiness while also increasing consumer pleasure, various machine learning algorithms were used to make the best of best models to predict employee resignation rates. The paper discusses the models that are able to predict employee resignation rate with accuracies above 80%.

The rest of the paper is organized as follows: In section II, we briefly discuss the literature review. Section III outlines the methodology of the research work followed by results in section IV. Finally, in section V, the paper is concluded with remarks on future work.

II. LITERATURE REVIEW

Preecha Tangsukeesiri and collegues [2] researched the Factors Affecting Resignation of Young Employee in Hotel Business, Thailand The contrast facts and figures of 'Retention Policies' and 'Young employees Turnover' rate lead researchers to evaluate factors affecting resignation of young employees of hotel business in Thailand with the aim to reduce the problem of turnover as well as the costs of investment caused by the high rate of turnover. A quantitative approach using 500 questionnaires was distributed to the young employee age range between 18-36 years old who are working in Hotel Business in 5 regions of Thailand. Data were analyzed by using descriptive analysis. The findings contribute the organizations to realize the factors of resignation and existence of the young employee leading to improve the retention strategies in the Hotel Business in Thailand

Linos, Elizabeth, Ruffini, Krista, Wilcoxen, Stephanie [3]experimented on employee burnout. It is a chronic problem in government organizations around the world, since high job demands and limited job resources contribute to high rates of employee burnout. Despite the fact that four decades of study has identified the determinants and potential costs of frontline worker burnout, there is no causal information on how to prevent it. A multi-city field experiment (n=536) focused at enhancing perceived social support and affirming membership among 911 dispatchers was reported in this publication. Burnout is reduced by 8 points (0.4 SD) and resignations are reduced by more than half after a six-week intervention that encourages dispatchers to offer advise anonymously and asynchronously with their peers in other cities.

III. METHODOLOGY

The most important purpose behind this report is to introduce machine learning models that will help to predict employee resignation with the greatest accuracy.

A. Data collection

The first task was to precisely analyze and mark the important insights of the dataset. The dataset that was used in the employee resignation project is the IBM's hr analytics attrition dataset. There are 1,470 rows and 35 columns in the dataset, with no null values in the data.

B. Data preprocessing

The raw data were pre-processed before applying in different machine learning techniques. This way all the irrelevant and unnecessary data were dropped. The categorical values were converted into numerical values Visualization techniques used to plot features against attrition and checked the most significant impacts on the rate of employee resignation. Min-Max Scaler used to choose the min-max values between zeros and ones.

C. *Feature selection*

Feature selection by variance threshold is then used to select relevant features and remove unwanted features (Over18, StandardHours, EmployeeCount). The correlation was checked between the features and dropped the columns with correlation over 90% (columns [38, 10, 30]).

D. *Classification*

The raw data was split into 85% training and 15% testing sets by hold-out validation technique. The train set is then used to form a model by using various machine learning algorithms, including Decision Tree, K-Nearest Neighbor, Logistic Regression and Random Forest.

Decision Tree: Visualizing and understanding a decision tree is easy, and we found that an entropy-based decision tree with seven nodes deep is optimum for us. Decision tree works with information gain and entropy to split the features such that the data is distributed in the tree as heterogeneously as possible. Eq. (1) shows the formula for entropy, which is calculated for every feature and the highest is selected as the root node:

$$E(s) = \sum_{i=1}^{c} -pi * log2 * pi \quad (1)$$

K-Nearest Neighbor: It is one of the most intuitive algorithms to understand, and a value of k = 9 with the distance metric as Manhattan distance (1-norm) was used for this dataset. KNN works by identifying k nearest data points and choosing the majority class from those. There are several distance metrics in use, which is used to calculate the nearest neighbors. Manhattan distance metric (see equation 2) is used here.

$$distance = \sum^{n} |Xi - Yi| \quad (2)$$

Logistic Regression: Logistic regression is a linear regression algorithm transformed using a sigmoid function to make it a classifier capable of predicting probabilities of each class. Linear regression uses a hypothesis function described in Eq. (3). A logistic regression passes the linear regressor into a sigmoid function, making it a probability- calculating classifier, as shown in Eq. (4)

$$h\theta(X) = \theta 0 + \theta 1X \quad (3)$$

$$\emptyset(z) = \frac{1}{1} + e^{-z} \quad (4)$$

Random Forest: A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max samples. parameter if bootstrap= True (default), otherwise the whole dataset is used to build each tree.

Hyperparameter tuning: Hyperparameter tuning was applied in the Decision tree classifier and Random Forest Classifier for better accuracy.

IV. Result

For the evaluation of the model, there are different metrics of machine learning that can be used to measure the performance such as accuracy, precision and f1 scores. In this model, accuracy has been used to evaluate the measure of the performance

A. *Accuracy*

Accuracy is one of the most commonly used measures of performance. The calculation of the performance is done by the total correct prediction made over the total number of predictions.

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

Where,

FP = False Positive
FN = False Negative
TP = True Positive
TN = True Negative

Classifiers	Training accuracy	Testing accuracy
Decision Tree	85%	82%
K Nearest Neighbor	87%	84%
Random Forest	100%	86%
Logistic Regression	90%	82%

Table 1: Accuracy results for classification model

Having applied gridsearch in the model, the testing accuracy of the decision tree algorithm increased slightly from 82% to 85%.

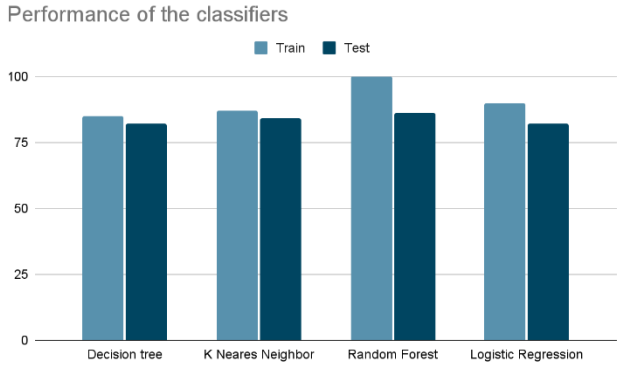

Table 2: Performance of the classifiers

V. Conclusion and future work

In this research, we employed machine learning techniques to forecast the likelihood of employee resignation. The dataset was collected from kaggle and in terms of accuracy our model on the dataset shows a decent prediction of 86% using random forest algorithm. In future, we plan to work on recommending changes that would improve the prediction rate of the resignation.

VI. References

[1] R. Walters, "5 reasons why employees resign - and how to stop them," [Online]. Available: https://www.robertwalters.com.ph/hiring/hiring-advice/5-reasons-why-employees-resign-and-how-to-stop-them.html.

[2] P. Tangsukeesiri, "Mendeley," 1 January 2018. [Online]. Available: https://doi.nrct.go.th//ListDoi/listDetail?Resolve_DOI=10.14457/NU.res.2018.48.

[3] E. R. K. W. S. Linos, "Mendeley," 1 january 2021. [Online]. Available: https://www.openicpsr.org/openicpsr/project/148502/version/V1/view?path=/openicpsr/148502/fcr:versions/V1/replication/replicationdata&type=folder.