International Journal of Computer and Communication Technology

Volume 9
Issue 1 *Research on Computing and Communication Sciences.*

Article 8

July 2023

Employee Attrition System Prediction using Random Forest Classifier

Soumen Nayak
Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India, soumen.nayak@gmail.com

Pranati Palai
RNIASE, Cuttack, pranatipalai@gmail.com

Recommended Citation

Nayak, Soumen and Palai, Pranati (2023) "Employee Attrition System Prediction using Random Forest Classifier," *International Journal of Computer and Communication Technology*: Vol. 9: Iss. 1, Article 8.
DOI: 10.47893/IJCCT.2023.1445
Available at: https://www.interscience.in/ijcct/vol9/iss1/8

Employee Attrition System Prediction using Random Forest Classifier

Soumen Nayak[1*], Pranati Palai[2]
[1]Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India
[2]Department of Teacher Education, RNIASE, Cuttack
[*]Soumen.nayak@gmail.com, pranatipalai@gmail.com

Abstract
Despite rising unemployment, most job coverage of the COVID-19 outbreak has concentrated on layoffs. Employees have been fired for reasons related to the epidemic, which has been a less prominent issue. COVID-19 is still doing damage to the country's economy. Companies are in the midst of a recession, so they are beginning to fire off unproductive employees. Making critical decisions like laying off employees or cutting an employee's compensation is a challenging undertaking that must be done with extreme attention and accuracy. Adding negligence would harm the employee's career and the company's image in the industry. In this paper, we have predicted employee attrition using Logistic Regression, Random Forest, and Decision Tree techniques. Random Forest Classifier has outperformed other algorithms in this work. After using different machine learning techniques, we can say that Random Forest gives the best performance with a recall of 70%, and also, we have found Precision, Accuracy, and F1-Score.

Keywords: Logistic Regression, Decision Tree, Random Forest, Attrition, Layoffs, Recall, F1-Score.

1.Introduction

An organization must devote a lot of time and resources to each employee's training to meet the needs of the business. When an employee departs the firm, it loses not only one of its important workers but also the money it spent on hiring, selecting, and training that person for the position. On the other hand, the company needs to increase investments in hiring, training, and developing new employees to fill their open jobs. Due to these factors, every organization strives to reduce employee churn by creating work cultures and corporate policies that are more satisfying.

When companies face a recession, their HR cuts down their underperforming employees. Making such crucial decisions as cutting down employees or reducing an employee's salary is challenging, which is to be done with utmost care and precision. Adding negligence will attrite the employee's career and the company's reputation in the market. This also leads to a loss in the business section of the company, leading to business disruption. Not using an effective technique for deciding employees' faith will be more costly and time-consuming. At the point when a representative leaves his situation in an unforeseeable way, it influences the presentation of the gathering. Most importantly, assuming that concerns a worker in a critical position; It is a complicated and tedious undertaking to supplant him. Consequently, the HR administrator must screen and evaluate what is going on with workers on a few variables to expect his will for steady loss.

Deciding to remove an employee from a company is not easy; one must keep everyone from their jobs because it affects both the employee and the company. Attrition of an employee who has been a part of the company for a long time and is loyal and dedicated to his work will destroy that employee's morale, and his livelihood will affect a lot. At the same time, removing such and employee will hamper the image and reputation of the company. A good set of employees can provide a great working environment, increase the flow of work, help generate more revenue for the company.

In this project many factors are considered that helps to accurately calculate the worth of an employee so that not even a single worthy employee has to leave the company because of some tiny human or machine error. Some machine learning techniques such as Logistic Regression, Random Forest and Decision Tree are used to anticipate this accurately. Evaluation metrics like Precision, Accuracy, Recall and F1-Score are also calculated. This is crucial for attrition prediction since it allows us to accurately estimate future employees with a high chance of leaving the company and having attrition.

To make the HRs of the company work more accessible and to not insult the company's reputation by kicking off efficient employees. Our objective is to use Machine Learning as a platform to predict whether an employee is eligible to remain in the company during this recession. The motivation for executing this project is not to make the wrong HR decision that might remove an employee by not considering all the necessary factors, which will damage the employee's career and the company's reputation if it is not so conscious of kicking off someone from the organization.

In this paper, an informative examination of the reasons is given, why that happens, and what pushes a worker to leave the organization. Then, we show how the models worked with a total assessment utilizing old style techniques.

This paper is coordinated as following. In Section 2, Literature Survey is presented that describes some of the prior works related to the fields or techniques used. Section 3 is committed to the methodology. Section 4 is dedicated to the outcomes and embraces nitty-gritty procedures. Last, in Section 5, conclusions are drawn.

2. Literature survey

Many researchers and academicians have proposed several reasons and techniques to assess employee attrition. Some of them are as follows. Ozolina [1] and

Slavianska [2] have discussed the effect of workers whittling down on human assets the executives and introduced a few viable practices to decrease this peculiarity. Mamun and Hasan [3] comprehensively examined worker turnover's primary drivers and elements. They additionally introduced a few doable techniques to guarantee that representatives go on in associations to build their viability and efficiency. At the same time, Berry and Michael [4] have shown how a model organizes the associations between Employee Engagement Factors and Intent to Leave. In this study [17], several retention strategies for the targeted personnel were improved using a novel technique centered on machine learning. This report also gives insight into factors impacting worker attrition rates and their potential remedies.

In a comparative setting, Babajide [5] shows the meaning of individual factors in the decision to leave the association made by delegates. Here creators moreover showed a backslide and assessment of the impact of the distinct components in specialist turnover objective. The larger pieces of the work have applied artificial mental ability methods to make models to predict whether agents will find business elsewhere. Dube etal.[6] have applied coordinated learning techniques for a matched decision to assume specialist wearing out. They have shown that using the KNN computation (48% in F-score) appeared differently about Naive Bayes (Gaussian), Logistic Regression, and MLP Classifier (ANN). Yedida et al.[7] in like manner, we have acquainted a twofold model with predicting which laborer will leave the association and which won't, using the determined to backslide procedure. The pipeline shown in this research [18] uses machine learning to forecast employee turnover while offering the business a low-cost strategy to keep the person on board.

Lao and Adeyemo [8] used a regulated game plan to portray the delegates as predefined objects for expecting the decision of hardheaded wearing out. The results show that the estimation C3.5 is more powerful (22.6% to the extent of the F score measure) and appeared differently concerning REP tree and Cart computations.

Fallucchi et al. [9] analyzed the factors and the reasons that make employees leave their jobs using IBM's enlightening assortment for data assessment. They present the complete evaluation of the AI models they work for predicting delegate debilitating. Yang et al. [10] utilized a connection lattice to eliminate the extra highlights from the dataset. The significant elements from the dataset were chosen using the Random Forest calculation. The qualities like month-to-month pay, age, and the number of organizations worked affected representative weakening. Eduvie et al. [11] used the decision tree model for the assumption of laborers wearing out.

In [12], a better predictor model—often called ensemble learning—is created by first combining the accuracy of five basic models. Compared to other methods like Adaboost and Random Forest, as well as third-placed SVM and gradient boosting, the accuracy obtained by the paper's combination of decision trees and linear regression is 86.39 percent. The article [13] uses the same IBM dataset to test various models, including SVM, Decision Trees, Random Forest, and KNN. By creating visualization graphs of gender, travel, and overtime vs. attrition, it assesses the following. Based on the provided variables, it forecasts employee turnover, which is the highest in Random Forest compared to other classifiers. Yadav et al. [14] employ the Brute Force technique, One Hot Encoding, and the Feature Selection strategy to extract accuracy from its 12 primary characteristics. The Feature Selection Approach using the random forest algorithm has the most incredible accuracy attained, which is a respectable level of accuracy. As a result, the article's authors decided to use a feature selection strategy to investigate the outcomes of ensemble trees, such as Random Forest and Gradient Boosting.

In this research [15], we offer a very reliable XGBoost model for predicting Employee Attrition using Machine Learning. This helps businesses anticipate employee loss and helps them develop economically by lowering the cost of their human resources. Using the dataset "IBM HR Analytics Employee Attrition Performance" and the tree-based Ensemble Machine Learning Model, Mehta and Modi [19] thoroughly analyze employee attrition. An employee's choice to quit the company is connected to several statistically essential factors. The study assesses the tree-based ensemble to acquire the best outcomes from the currently available tree approaches. To illustrate factors that influence the target candidate's decision and forecast the likelihood of candidate retention before training, Marvin et al. [16] effectively develop a variety of machine learning classifiers.

3.Materials and Methods

In this section, different steps that are followed to process the data have been discussed.

3.1 Data assortment

IBM Analytics disseminates the dataset that we utilized in this paper from Kaggle.com. This dataset contains 35 qualities connecting with 1471 perceptions. We have chosen solid expert and individual highlights by disposing of those given using human investigation and understanding. For instance: 'Job Satisfaction,' 'Relationship Satisfaction, and so on. Like this, the model we can develop is simpler to adjust to the different HR information. This informational collection comprises different subtleties and elements like age and orientation of the worker, training and month-to-month pay, and different variables that might affect the representative wearing down the choice. Table 1 depicts the rundown of the highlights that we utilized in this work. The data contains a target include, distinguished by the Attrition Variable (1 or 0), that addresses the twofold choice for worker steady loss.

TABLE 1
The elements used to foresee the attrition of employee

Years Since Last Promotion	Years With Current Manager
Years At Company	Years In Current Role
Training Times Last Year	Work Life Balance
Percent Salary Hike	Total Working Years
Monthly Rate	Numbers of Companies Worked
Marital Status	Over Time
Job Level	Monthly Income
Education	Hourly Rate
Department	Education Field
Daily Rate	Distance From Home
Age	Business Travel
Attrition(Target variable)	

3.2 Data handling

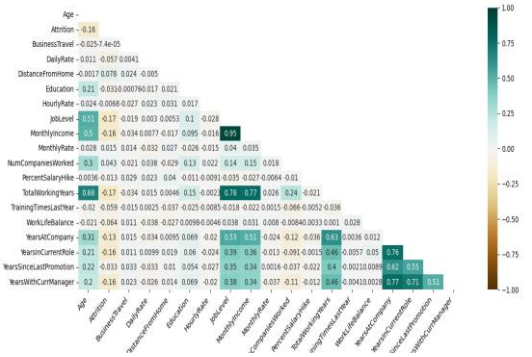
Information planning is one of the fundamental parts of ML that would influence the built model's nature. The principal task that we did was information determination by eliminating highlights that were not proficient. For instance, the moderate representative numbers of workers count, and so on. We killed the 'invalid' values into more elements since they would affect the right arrangement of the model and, consequently, produce misguided assumptions. We change features 'Orientation' and 'married or not' characteristics into Boolean ones instead of the message. Finally, we convert out and out factors: 'Division,' 'Occupation Role,' and 'Schooling' into farce factors as the AI model wellsprings of data.

3.3 EDA and Visualization

This stage's most vital phase was introducing the connection grid heat grid. Fig.1 shows the connection among the pre-owned factors. The dim variety addresses no relationship, and the overall force of brown and green separately addresses the positive and not positive relationship. We conclude from dissecting the connection framework that the accompanying element has a positive relationship:

- Monthly Income and Job Level (0.95)
- Total Working Years and Job Level (0.78)
- Years At Company and Years With Current Manager (0.77)

Fig. 1 Correlation matrix heat map



For a negative association, the above network shows that Attrition is connected Conversely with Monthly Income, Total Working Years, Years in Current Role, and Job Level. As we said beforehand, we deal with an essential twofold component. We have 1234 (85%) discernments with a negative motivator for Attrition, and 236 (15%) insights are up-sides. Fig. 2 shows the dispersal of the objective variable inside the three individual components: 'Orientation,' Age, and 'Conjugal Status.' We reason that men have a more significant probability (17.02%) of leaving their posts than women (14.07%). For 'Conjugal status, single individuals are the ones who have a high probability of consistent misfortune decisions. The chart shows that a fourth of single delegates from the discernments leave their jobs and posts obstinately. Concerning', 'obviously, the likelihood of pursuing the steady loss choice declines as the worker age increments for people under 40. The chart demonstrates how choice whittling down could build relative maturity for people who spent 40 years of age. The month-to-month income is Among the expert variables that can influence the weakening choice. In Fig. 3, the month-to-month pay diminishes as the likelihood of whittling down selection builds: 34% of individuals with a month-to-month pay somewhere between 1000 and 2500 take the steady loss choice.

Notwithstanding, just 10% of individuals who have more than 5500 in month-to-month pay make this choice. 'Percent Salary Hike' presents the improvement compensation pointer (in rate). We can derive from the chart that this element could affect the whittling-down choice, assuming it is excessively low.

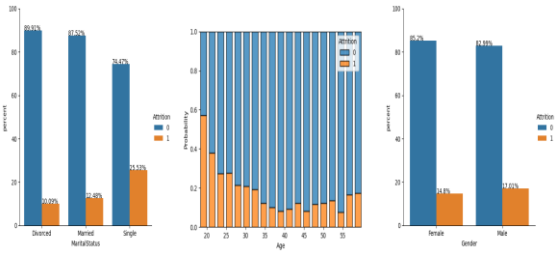


Fig. 2 Attrition distribution inside personal elements

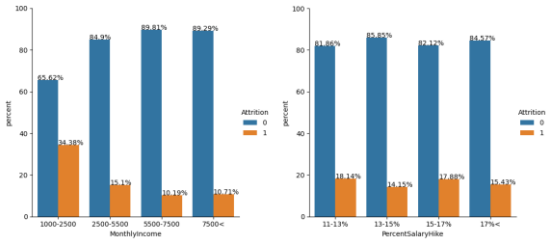


Fig.3Attrition distribution inside Monthly Income and Percent Salary Hike

Business Travel, Marital Status versus Attrition %

Since Lockdown measures in the current scenario, have restricted travelling, let us examine which type of employees frequently traveling or rarely traveling were affected the most. Also, we will check the marital status

of employees and its correlation with Business Travel and Attrition. This is represented in Fig. 4.

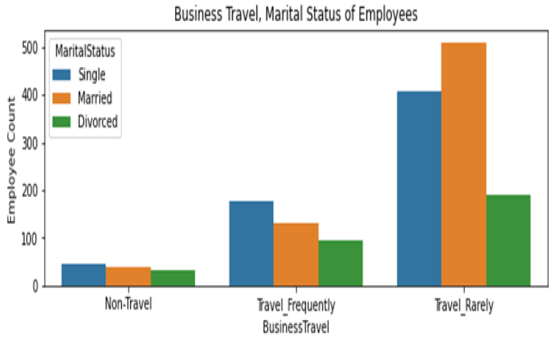


Fig. 4 Business travel vs. marital status

Here, we see that Married Employees rarely travel (about 500), which is highest in that category since they have to take care of their household, rarely making time for business travel. While in the Travel Frequently Category, Bachelors take up the highest count (about 180) since they don't have the burden of raising children, or household chores, unlike Married Employees. But interestingly, Divorced Employee Count is the lowest in all three categories. This can also be because of fewer records for Divorced Employees.

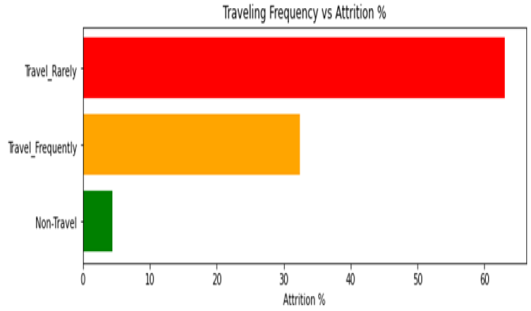


Fig. 5 Traveling frequency vs. attrition%

Jobs involving rare traveling have been affected the most, having more than a 60% Attrition Rate, as depicted in Fig. 5. At the same time, Jobs involving non-travel have the least Attrition Rate of less than 5%. It makes sense because, in recent times, most firings have taken place for working in IT Sectors as Software Engineers in companies like Capgemini, Cognizant, Uber, Ola Cabs, etc. This concludes that most jobs that were affected rarely involved traveling; restrictions didn't impact jobs as much as the loss of revenues due to the lockdown measures.

3.4 Model Construction

Different controlled AI computations are used to assemble models for expecting the worker to whittle down a choice. We want to check and analyze the desired outcomes for every model to choose the most capable among them. Our benefit is predicting the delegates who would decide to leave the association. In this manner, we will search for the calculation that could limit the false negatives. Individuals fled the organization, and the analysis wrongly arranged them. We utilized the accompanying measures to anticipate the representative steady loss choice:

- Decision tree
- Logistic Regression
- Random Forest

In the information preparation stage, we partition the information into a training set that contains 70% of the data perceptions and we utilized the other information (30%) for tests.

Fig. 6-The workflow of proposed model

4. Result and Discussions

This sectiowill shothe outcomes and nature of every taken-on modeBefore this, we utilized the confusion matrix to infer some significant measurements to have a communication of the proficiency of every calculation quantitatively:

- **Accuracy**: Accuracy is the proportion of number of right expectations to the complete number of information tests.
- **Recall**: Recall = $2 \times (TP \div (TP + FP))$

FN is the number of false negatives.

- **F1-score**: F1– Score = $2 \times ((precision \times recall) \div (precision + recall))$
- **Precision**: precision = $TP \div (TP + FP)$

TP isthe number of true positives, and FP is the number of false positives.

Table2 shows the assessment for every one of the calculations utilizing the actions that made sense above. We could express, as per the below table, that Random Forest is the best performer model as it achieves our objective. Figure 7 depicts the confusion matrix for Random Forest.

TABLE 2
The Algorithms Performance Evaluation

Models	Train Accuracy (%)	Test Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)

Decision Tree	**90**	**71**	**30**	**39**	**57**
Logistic Regression	**61**	**60**	**23**	**35**	**55**
Random Forest	**98**	**85**	**16**	**24**	**70**

The confusion matrix for the results of random forest is represented in Fig.7.

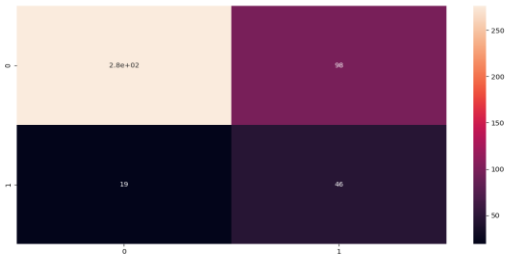


Fig. 7 The confusion matrix of Random Forest**.**

In this paper, the dataset of IBM Analytics from Kaggle.com is taken. The methodologies we have taken in this paper are Logistic Regression, Decision Tree, and Random Forest. Using different ML techniques, we obtained accuracy levels of 85%, and Random Forest outperformed other algorithms in this case.

5.Conclusion and future work

This paper, it is shown how employee attrition has effects on a company and what impact it generates. Also, it is observed that what factors can play an essential role in deciding if an employee leaves their post? Machine learning models are provided for solutions to estimate those who want to quit that organization. We split the data in the pre-processing part into training data, which has been used to train the models, and also test data to check how the models perform. Evaluation metrics like precision, accuracy, recall, and F-score are used to choose the best-performing model. Random Forest is the model that portrayed the foremost out of all, as it can minimize false positives (70% Recall). This work shows an essential part of human resources management's daily problems.

Talking about future work, our main agenda will be to solve as many problems and challenges in the human resource department of an organization like predicting under-qualified candidates who appear for a job interview for that organization.

References

[1] Ozolina-Ozola, I. (2014). The impact of human resource management practices on employee turnover. *Procedia-Social and Behavioral Sciences*, *156*, 223-226.

[2] Slavianska, V. (2012). Measuring the impact of human resource management practices on employee turnover. *Problems of Management in the 21st Century*, *4*(1), 63-73.

[3] Al Mamun, C. A., & Hasan, M. N. (2017). Factors affecting employee turnover and sound retention strategies in business organization: A conceptual view. *Problems and Perspectives in Management*, (15, Iss. 1), 63-71.

[4] Berry, M. L., & Morris, M. L. (2008). The Impact of Employee Engagement Factors and Job Satisfaction on Turnover Intent. *Online Submission*.

[5] Babajide, E. O. (2010). The influence of personal factors on workers turnover intention in work organizations in South-West Nigeria. *Journal of Diversity Management (JDM)*, *5*(4).

[6] Dubey, A., Maheshwari, I., & Mishra, A. (2018). Predict Employee Retention Using Data Science. *Journal of Diversity Management (JDM)*, *5*.

[7] Yedida, R., Reddy, R., Vahi, R., Jana, R., GV, A., & Kulkarni, D. (2018). Employee attrition prediction. *arXiv preprint arXiv:1806.10480*.

[8] Alao, D. A. B. A., & Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms. *Computing, Information Systems, Development Informatics and Allied Research Journal*, *4*(1), 17-28.

[9] Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting employee attrition using machine learning techniques. *Computers*, *9*(4), 86.

[10] Yang, S., & Islam, M. T. (2020). IBM employee attrition analysis. *arXiv preprint arXiv:2012.01286*.

[11] Eduvie, R., Nwaukwa, J., Eric, T., & Uloko, F. (2021). Predicting employee attrition using decision tree algorithm. *Global Scientific Journal*, *9*(9).

[12] Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R., & Alghamdi, H. S. (2021). Prediction of employee attrition using machine learning and ensemble methods. *Int. J. Mach. Learn. Comput*, *11*(2), 110-114.

[13] Patel, A., Pardeshi, N., Patil, S., Sutar, S., Sadafule, R., & Bhat, S. (2020). Employee attrition predictive model using machine learning. *International Research Journal of Engineering and Technology (IRJET)*, *7*(5).

[14] Yadav, S., Jain, A., & Singh, D. (2018, December). Early prediction of employee attrition using data mining techniques. In *2018 IEEE 8th international advance computing conference (IACC)* (pp. 349-354). IEEE.

[15] Kakad, S., Kadam, R., Deshpande, P., Karde, S., & Lalwani, R. (2020). Employee attrition prediction system. *Int. J. Innov. Sci., Eng. Technol.*, *7*(9), 7.

[16] Marvin, G., Jackson, M., & Alam, M. G. R. (2021, August). A machine learning approach for employee retention prediction. In *2021 IEEE Region 10 Symposium (TENSYMP)* (pp. 1-8). IEEE.

[17] Jain, P. K., Jain, M., & Pamula, R. (2020). Explaining and predicting employees' attrition: a machine learning approach. *SN Applied Sciences*, *2*, 1-11.

[18] Patro, A. C., Zaidi, S. A., Dixit, A., & Dixit, M. (2021, June). A novel approach to improve employee

retention using Machine Learning. In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 680-684). IEEE.

[19] Mehta, V., & Modi, S. (2021). Employee Attrition System Using Tree Based Ensemble Method. In *2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4)* (pp. 1-4). IEEE.