Asian Journal of Research in Computer Science

Volume 15, Issue 3, Page 34-39, 2023; Article no.AJRCOS.98295
ISSN: 2581-8260

Machine Learning based Employee Attrition Predicting

Subhani Shaik [a*], P. Santhosh Kumar [b++], S. Vikram Reddy [b++], K. Sai Srinivas Reddy [b++] and Sunil Bhutada [b#]

[a] Jawaharlal Nehru Technological University Hyderabad, Hyderabad, Telangana, India.
[b] Department of Information Technology, Sreenidhi Institute of Science and Technology (Autonomous), Hyderabad, Telangana, India.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2023/v15i3323

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/98295

Original Research Article

Received: 28/01/2023
Accepted: 30/03/2023
Published: 04/04/2023

ABSTRACT

Now a day's variety of reasons for job resignations due to this, we have to take different types of measurements for prediction of job seekers. They have different reasons for not doing jobs well and fell like pressure. Many employees suddenly come to an end of their service without any reason. Techniques of machine learning have full-grown in fame in the middle of researchers in current years. It is accomplished of propose answer to a broad range of problems. Help of machine learning, you may produce prediction concerning staff abrasion. So machine learning model we will be using TCS employee attrition a genuine time dataset to train our model. The aim of this study is to at hand a comparison of different machine learning algorithms for predict which employees are probable to go away their society. We propose two methods to crack the dataset into train and test

---

[++] B. Tech Student;
[#] Professor and Head;
**Corresponding author: E-mail: subhanicse@Yahoo.com, subhanicse@gmail.com;*

data: the 75 percent train 25 percent test split and the K Fold methods. Three techniques are three methods that we employ to train our model for correctness comparison, and we will compare the exactness of the models generate using these three Boosting Algorithms.

Keywords: Machine learning; gradient boosting algorithms; K-Fold methods; light GBM boost; XG boost.

1. INTRODUCTION

Asset of companies are employees and valued by organizations that spend by offering methodical training and a pleasant operational surroundings. Capable employees are being lost due to different reasons. Employee hiring is another problem for company running successfully [1-3]. Replacements rate of the company a lot of capital, as well as the expenses of hire, preparation, and interviewing applicant. Due to this reason management can change their rules and regulations of employee hiring, this is lot of burden [2-7]. So many problems are facing by management, if the employee hiring, it impact boost the wages and incentives. Further training, to decrease their chances of leaving. Machine learning approaches can be used to predict employee earnings [8-11]. Using past data from HR department, analyst may build and train a machine learning model that can forecast the employees will give up the company [12,13].

In our paper, we will influence employee data provided by TCS's HR department, which is obtainable on Kaggle, and will train our models using the k-fold validation technique, using 75 percent 25 percent dataset split. CatBoost, XGBoost, and Light GBM Boost were the machine learning algorithms are employed in our research. So we pick the majority of exact representation out of all of them and evaluate their accuracies.

2. LITERATURE SURVEY

Attrition among employees can be a major problem for businesses, especially when highly trained, technical, and critical employees leave for a better opportunity elsewhere. The increased interest in machine learning among company leaders and call centers necessitates that researchers investigate its application within businesses [14,15]. One of the most serious issues confronting business owners is the loss of talented employees. In lot of research papers there are machine learning models developed using the various algorithms like Support Vector Machine, Decision trees, XGBoost, K Nearest Neighbors, ANN, Random Forest etc. Machine learning has been used to predict employee behavior in several researches. To predict employee performance, the authors employed decision trees (ID3 C4.5) and the Naive Bayes classifier in their research. They discovered that task title was the most important factor, while age had no discernible effect. The authors used a dataset of 1585 records and 27 characteristics to test manifold data mining techniques for predicting staff churn. They employed Bayes algorithms, SVM, logistic regression, decision trees, and random forests as machine learning methods. All of these findings, a support vector machine with an accuracy of 85.12 percent should be considered.

3. IMPLEMENTATION AND PROPOSED PROCEDURE

“As we know in a lot of company’s employees resign their job in the IT sector that becomes a major issue for the businesses. Employee abrasion may be an enormous problem for companies, predominantly when extremely taught, theoretically skilled, and significant employees go away for a better opening elsewhere. As a outcome, a accomplished worker cannot be replace so quick. So as the knowledge is developing with the use of most recent technical advancements in the IT industry, we can make use of machine learning and build machine learning models, so that we can forecast employee slow destruction” [16].

Present years, we have seen extraordinary development in the gradient boosting algorithms like boosting algorithms [17,18]. “Due to in order to resolve our issues above verdict, whether the employee will be retain or not we will be using three gradient boosting algorithms, namely XGBoost, Cat Boost, LightGBM Boost. In order to train model accurately we will use the basic 65% train, 35% test data splits and then we will move to K Fold Validation method for splitting train and test data so that we can evaluate the correctness of the equivalent algorithms and choose the majority accurate trained model so that, we can make our predictions more

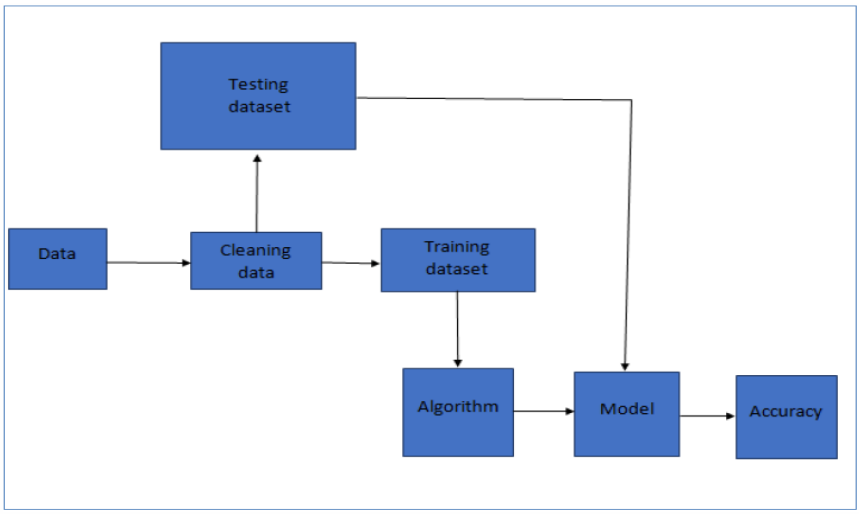

Fig. 1. Architecture of Proposed System

accurately based on the input data known to the trained replica based on the forecast the organization act for that reason without incurring any loss to their company or businesses" [16].

The procedure of establish the architecture, components data for a system in order to get together precise criteria is known as structure design. It's the application of systems theory to invention growth, in a nutshell. Object-oriented design and analysis methodologies are quickly becoming the majority of popular techniques for make computer systems [19,20].

TCS HR department, we will spotless the dataset given by this module. The put into put into practice of correcting or delete incorrect, dishonored, impolitely formatted, photocopy, or incomplete data from a dataset is known as data cleaning. In general the steps for onslaught the dataset are remove any clarification that are photocopy, conduct the absent data, management null values etc. In our dataset there are 4 irrelevant columns so we will remove them and there are some categorical data. So here we will change the categorical data into numerical data by using the Label Encoder from scikit learns.

We will use two types, test and train data split initially and we will use 75 percent train data and remaining 25 percent as test data, the second that we use is K Fold technique to split dataset into train and test dataset. So that, we can evaluate the accuracy of models build after training and make most excellent out of the accurate model trained.

After separating the dataset into train and test data using the two methods describe in above component, now we will build a model and train the copy using the train dataset, in our paper, we will use three dissimilar types of gradient boosting algorithms implementations are namely CatBoost, LightGBM, XGBoost to train our representation. We will train our algorithm for K Fold using folds values as 3,5,10. So that, we can evaluate accuracy and choose a most excellent model which is more correct for our prediction.

4. RESULTS AND ANALYSIS

"More research in the field of attrition may be found. Because the strategy to forecast employee attrition is quite similar to erosion, it enables us to predict alternative ways" [16] In [21], "combining various training previous observations per employee from Training Data improves the predicted performance of retention models compared to using simply the most relevant data. Another issue is that instead of obtaining several samples from the whole term of the individuals, they limit it to a small piece of data, implying that many jobs are once again eliminated".

"For implementation analysis, the data set is gathered from the Kaggle database, an open-access repository. Then trained data set Machine Learning models using the k-fold validation methodology, using 75 percent 25 percent dataset splits. CatBoost, XGBoost, and LightGBM Boost are the machine learning

algorithms employed in research to pick the most accurate model out of all of them and compare their accuracies" [16].

"In k-fold cross-validation, we initially rearrange data to ensure that the sequence of the dependent and independent variables is fully random. This process is executed to ensure that none of inputs are skewed. Next, we divided the dataset into k sections. Thus eliminated the over fitting issue, when a classifier is developed utilizing all of the data in one brief and gives the greatest prediction performance" [16].

Initially, the null values are check in the dataset with particular functions. Then the outcome is depict in the shape of warmth map that obviously showcase the null values in the dataset pictorially.

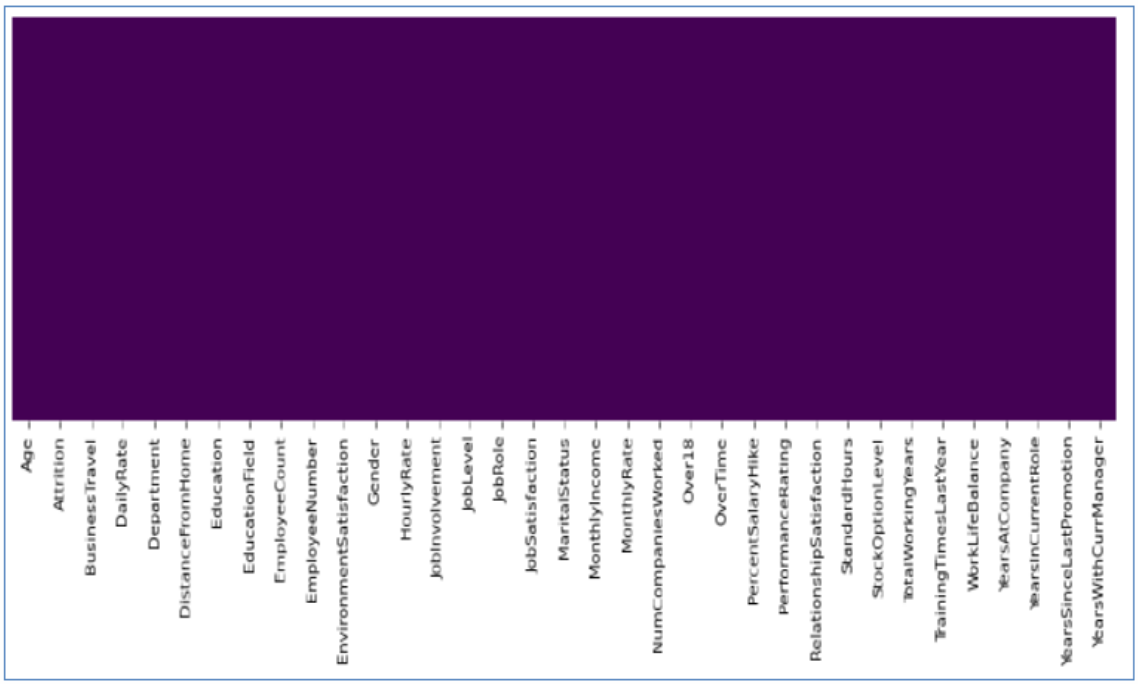

Fig. 2. Heat map

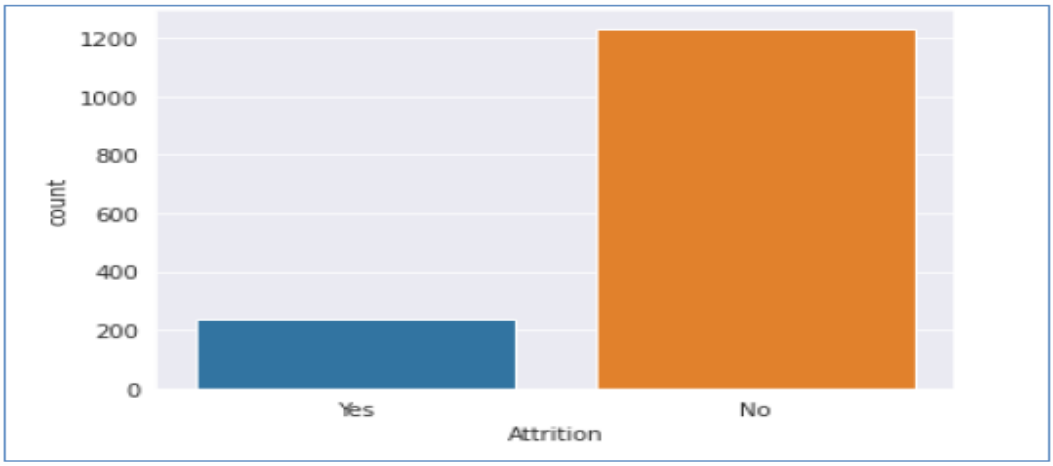

Fig. 3. The above graph represents count of attritions

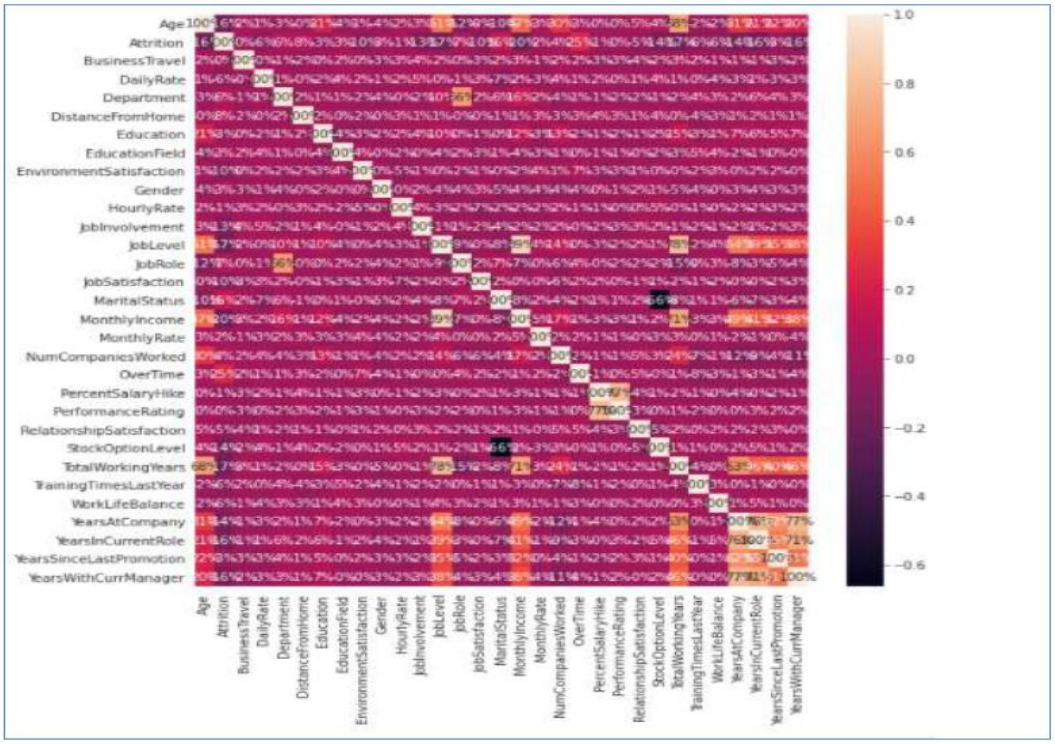

Fig. 4. Correlation matrix

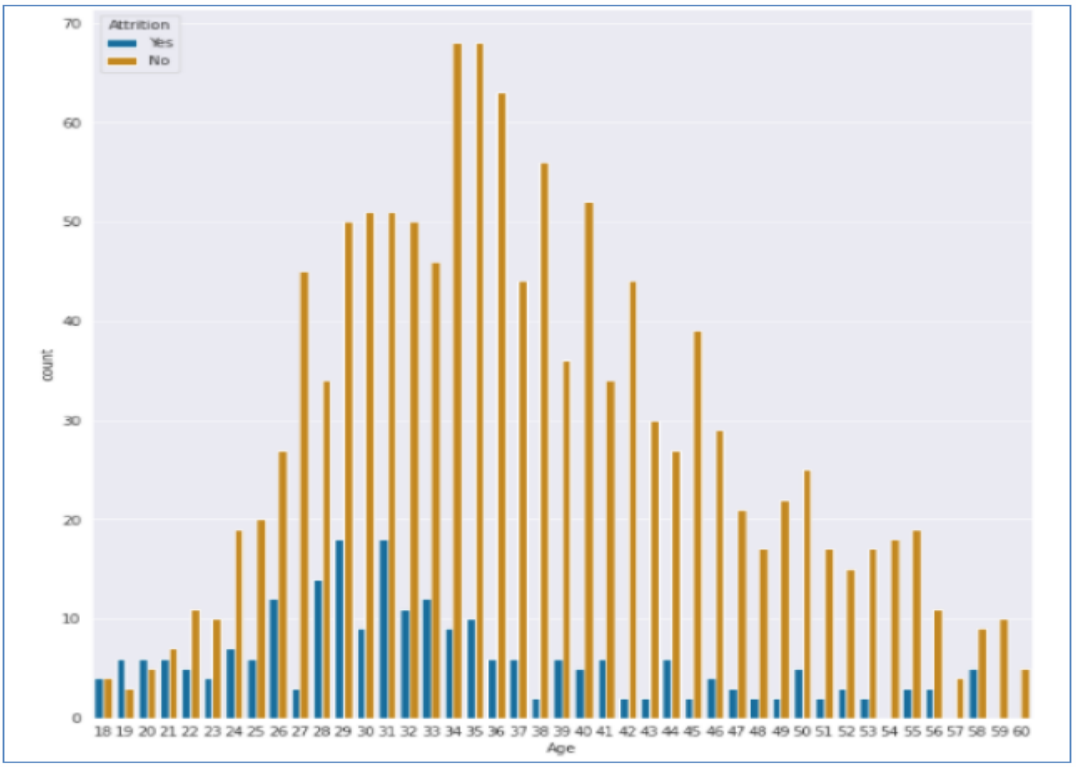

Fig. 5. Relation between Age and Attrition

5. CONCLUSION

Using these three boosting algorithms namely CatBoost, Light GBM, XGBoost with 75% train and 25% test dataset crack the LightGBM gave us the extra precise model than other two algorithms. When K Fold justification is used for algorithms namely CatBoost, LightGBM, XGBoost, then we got more correct models with 90.47% accuracy for two algorithms namely Cat Boot and XGBoost when K=10 in K Fold validation. So from our research, we have got more accurateness than the SVM classifier, decision trees, KNN, Random forest as mentioned in the some of the research papers as discusses in Literature Survey.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Available:https://ieeexplore.ieee.org/document/8605976
2. Available:https://ieeexplore.ieee.org/abstract/document/8746940
3. Available:https://ieeexplore.ieee.org/document/6216220
4. Available:https://www.researchgate.net/publication/326029536_Employee_Attrition_Prediction
5. Available:https://www.irjet.net/archives/V7/i5/IRJET-V7I5737.pdf
6. Available:http://www.jicrjournal.com/gallery/24-jicr-december-2247.pdf
7. Available: http://www.iosrjournals.org/iosr-jbm/papers/Vol20-issue2/Version-4/A2002040127.pdf
8. Available:https://ieeexplore.ieee.org/document/8541242
9. Available:https://ieeexplore.ieee.org/document/8541242
10. EMC Education Services, “Data Science and Big Data Analytics - Discovering, Analyzing, Visualizing and Presenting Data”, July 2015.
11. Pavan Subhash, IBM HR Analytics Employee Attrition & Performance; 2016.
12. Available:https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attritiondataset
13. Available:https://ieeexplore.ieee.org/document/9033784
14. Dr. Sunil Butada and Subhani Shaik. IPL Match Prediction using Machine Learning, IJAST. 2020;29(5).
15. Srivastava AK, Khan R. Fake news detection system using stance detection and machine learning approaches. International Journal of Forensic Software Engineering. 2022;1(4): 378-389.
16. Kakulapati V, Subhani S. Predictive Analytics of Employee Attrition using K-Fold Methodologies. I. J. Mathematical Sciences and Computing. 2023;1:23-36.
17. Ch. Shravya, Pravallika and Subhani Shaik, Prediction of Breast Cancer Using Supervised Machine Learning Techniques, International Journal of Innovative Technology and Exploring Engineering. 2019;8(6).
18. Santosh P, Subhani Shaik. Heart disease prediction with PCA and SRP, International Journal of Engineering and Advanced Technology. 2019;8(4).
19. Shiva Keertan J, Subhani Shaik. Machine Learning Algorithms for Oil Price Prediction, International Journal of Innovative Technology and Exploring Engineering. 2019;8(8).
20. Surya Teja KP, Vigneswar Reddy, Subhani Shaik, Flight Delay Prediction Using Machine Learning Algorithm XGBoost, Jour of Adv Research in Dynamical & Control Systems. 2019;11(5).
21. Kashyap GS, Malik K, Wazir S, Khan R. Using Machine Learning to Quantify the Multimedia Risk Due to Fuzzing. Multimedia Tools and Applications. 2022; 81(25):36685-36698.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/98295