Minireview Article

EMPLOYEE ATTRITION PREDICTING USING MACHINE LEARNING

Abstract---Almost every company nowadays is concerned about keeping their staff. They are nevertheless, unable to recognize the true reasons for their job resignations. This could be due to a variety of circumstances. Each business has its own approach to treating employees and ensuring their pleasure. As a result, many employees abruptly terminate their employment for no apparent reason. Machine learning approaches have grown in popularity among researchers in recent decades. It is capable of proposing answers to a wide range of issues. Then, using machine learning, you may generate predictions about staff attrition. In our Machine learning model we will be using IBM employee attrition a real time dataset to train our model. The goal of this study is to present a comparison of various machine learning algorithms for predicting which employees are likely to leave their firm. We will use two approaches to divide the dataset into train and test data: the 65 percent train 35 percent test split and the K Fold approaches. Three methods are three methods that we employ to train our model for accuracy comparison, and we will compare the accuracy of the models generated using these three Gradient Boosting Algorithms.

Keywords: Machine learning, Gradient Boosting Algorithms, K-Fold approaches, Light GBM Boost, and XGBoost.

1. INTRODUCTION

Employees are valued by companies that invest in them by offering thorough training and a nice working environment. They, too, are subjected to intentional attrition as well as the impacts of the environment. Skilled employees are being lost. Hiring is another issue; replacements cost the company a lot of money, including the expense of hiring, training, and interviewing applicants. Management will be able to act more swiftly by changing internal rules and methods if they can predict staff turnover. Where skilled workers on the verge of leaving are provided a range of incentives, such as a wage boost or further training, to reduce their odds of leaving. Machine learning techniques can be used to predict employee turnover. Using historical data from HR departments, analysts may construct and train a machine learning model that can predict which employees will quit the company.

In our project, we will leverage employee data provided by IBM's HR department, which is available on Kaggle, and we will train our Machine Learning models using the kfold validation methodology, using 70 percent 30 percent dataset splits. CatBoost, XGBoost, and LightGBM Boost were the Machine Learning algorithms we employed in our research so that we could pick the most accurate model out of all of them and compare 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. These results in the inability to replace a skilled worker. The increased interest in machine learning among company leaders and call centres necessitates that researchers investigate its application within businesses. 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 Neighbours, ANN, Random Forest etc. Machine learning has been used to predict employee behaviour 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 job title was the most important factor, while age had no discernible effect. The authors used a dataset of 1575 records and 25 features to test multiple data mining methods for predicting staff churn (or attrition). They employed naive Bayes, support vector machines, logistic regression, decision trees, and random forests as machine learning techniques. According to the findings, a support vector machine (SVM) with an accuracy of 84.12 percent should be used.

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 attrition may be a huge issue for firms, particularly when highly trained, technically skilled, and important staff leave for a better opportunity elsewhere. As a result, a skilled worker cannot be replaced so fast. So as the technology is evolving with the use of latest technological advancements in the IT industry we can make use of

Machine Learning and build ML models so that we can predict employee attrition.

In recent years we have seen remarkable progress in the Gradient Boosting algorithms like XGBoost, Cat Boost, and LightGBM Boost. So here in order to solve our problem above finding whether the employee will be retained 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 compare the accuracy of the corresponding algorithms and select the most accurate trained model so that we can make our predictions more accurately based on the input data given to the trained model based on the prediction the company act accordingly without incurring any loss to their company or businesses.

The process of establishing the architecture, components, modules, interfaces, and data for a system in order to meet specific criteria is known as system design. It's the application of systems theory to product development, in a nutshell. Object-oriented design and analysis methodologies are quickly becoming the most popular techniques for creating computer systems.

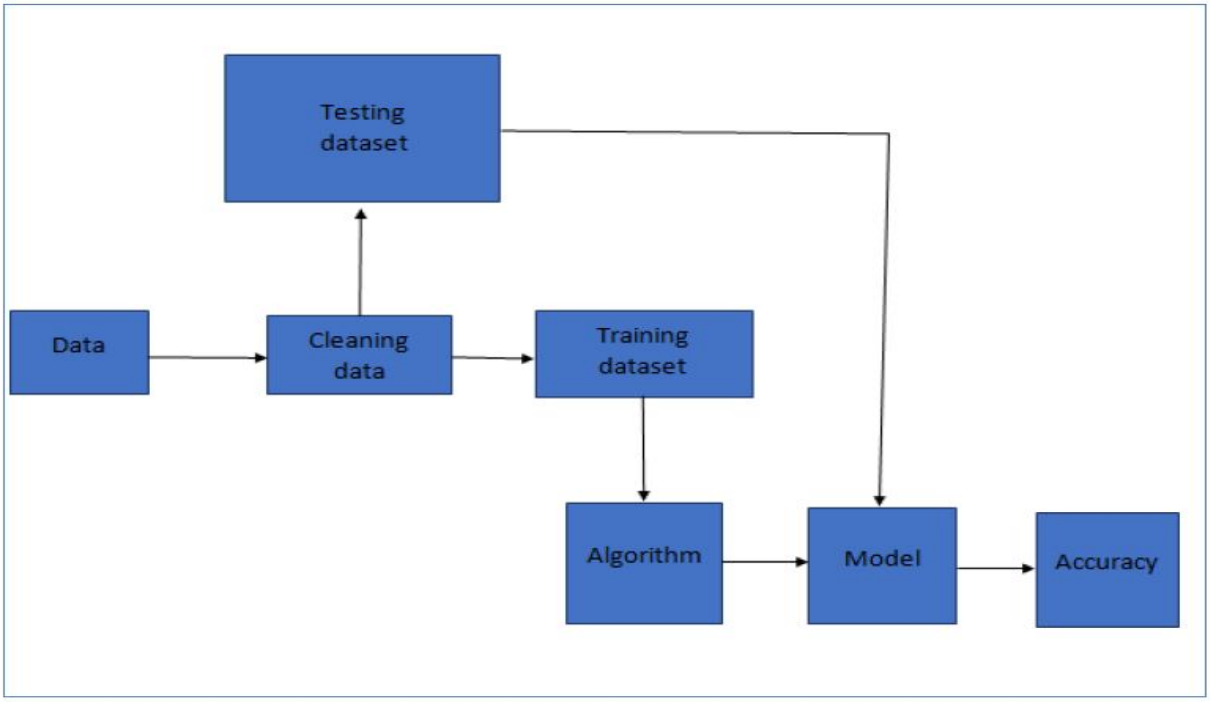

Figure 1: Architecture of Proposed System

Here in this module we will clean the dataset given by IBM HR department. The practice of correcting or deleting incorrect, corrupted, improperly formatted, duplicate, or incomplete data from a dataset is known as data cleaning. In general the steps for cleaning the dataset are removing any observations that are duplicated or irrelevant, Handling the missing data, Handling null values etc.. Here 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(integer) data by using the Label Encoder from scikit learn.

Here we will use two types of test, train data splits firstly we will use 65 percent train data and remaining 35 percent as test data, the second that we use is K Fold method to divide dataset into train and test dataset so that we can compare the accuracy of models build after training and make best out of the most accurate model trained.

After dividing the dataset into train and test data using the two methods described in above module now we will construct a model and train the model using the train dataset Here in our paper we will use three different types of Gradient boosting algorithms implementations they are namely CatBoost, LightGBM, XGBoost to train our model. We will train our algorithm for K Folds using folds values as 3,5,10 so that we can compare accuracy and pick a best model which is more accurate for our predictions.

After training the model we will test the trained model using the test data so that we can find the Accuracy of the model. As we have built models using different algorithms and different strategies of splitting test and train data we can pick the most accurate model out of all models tested depending on the accuracy so that we can predict the employee attrition more accurately. Using the most accurate model from the above module for every given input details of employee We will predict about the attrition of the employee.

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. In [19], 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 65 percent 35 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.

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.

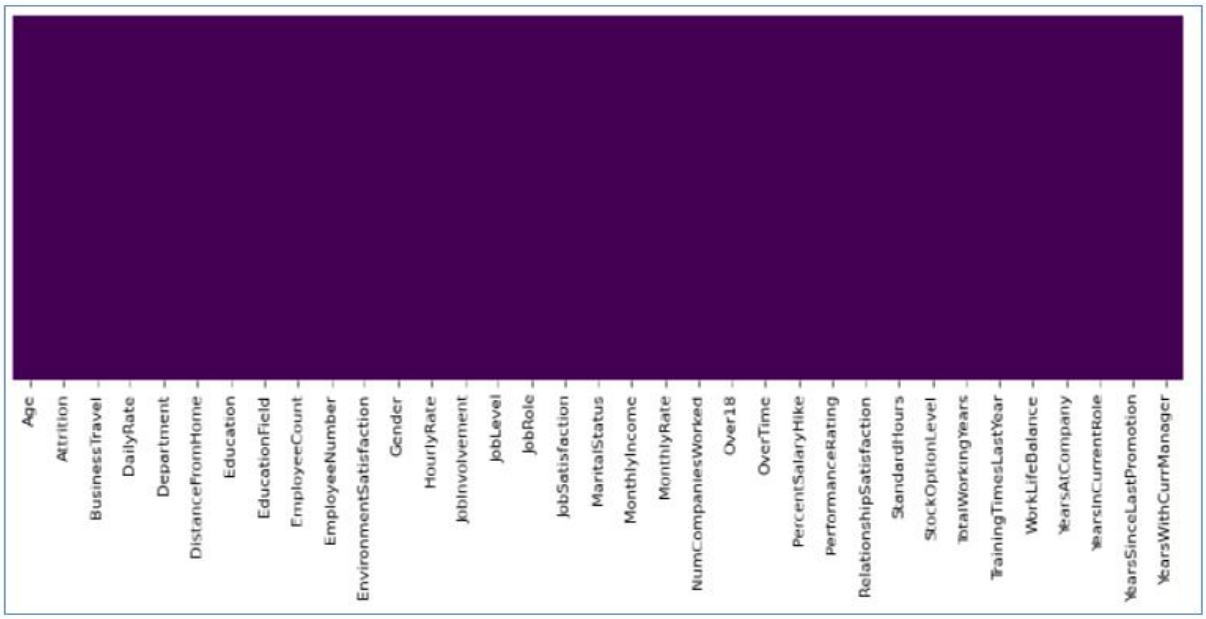

Figure 2: Heatmap

Firstly, the null values or the missing values are checked in the dataset with specified functions. Then the result is depicted in the form of heatmap that clearly showcases the null values or missing values in the dataset pictorially.

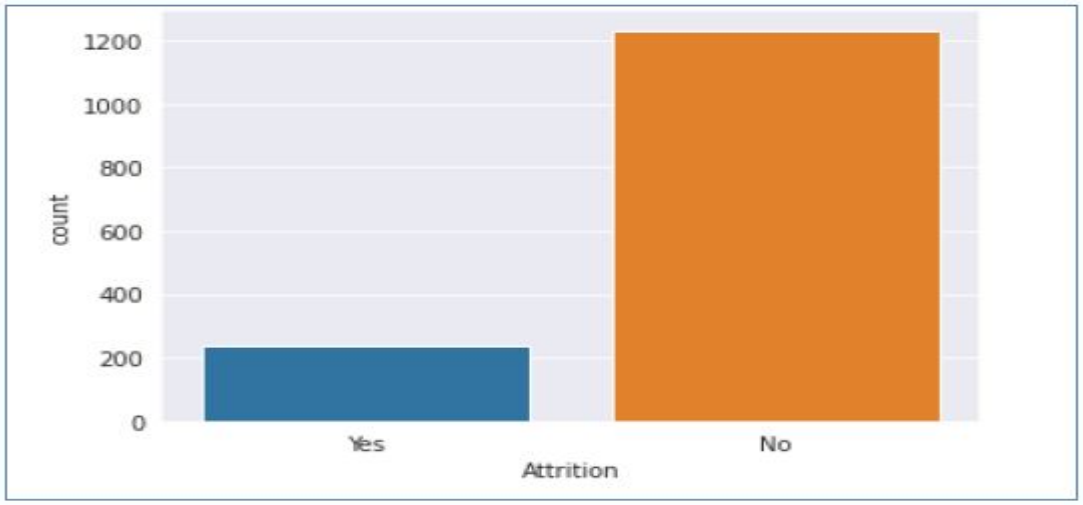

Figure 3: The above graph represents count of attritions

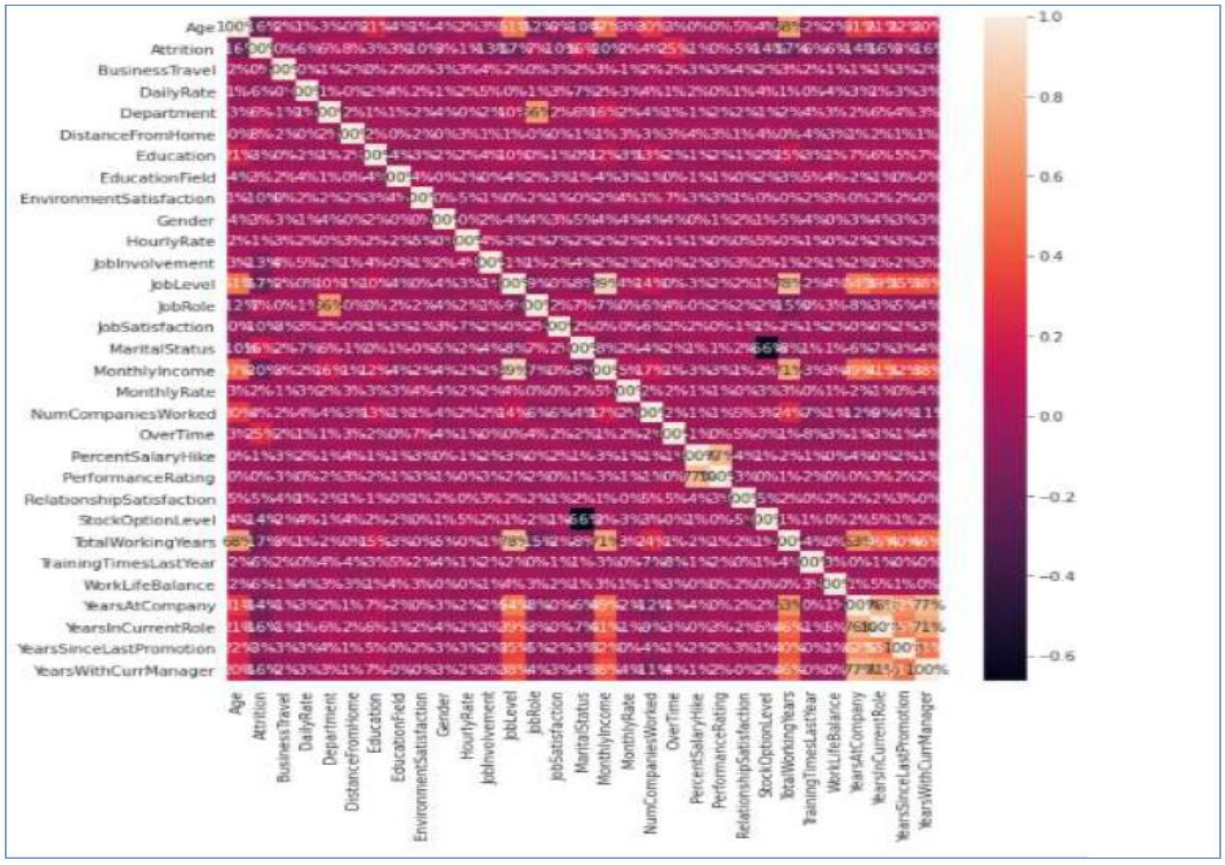

Figure 4: Correlation matrix

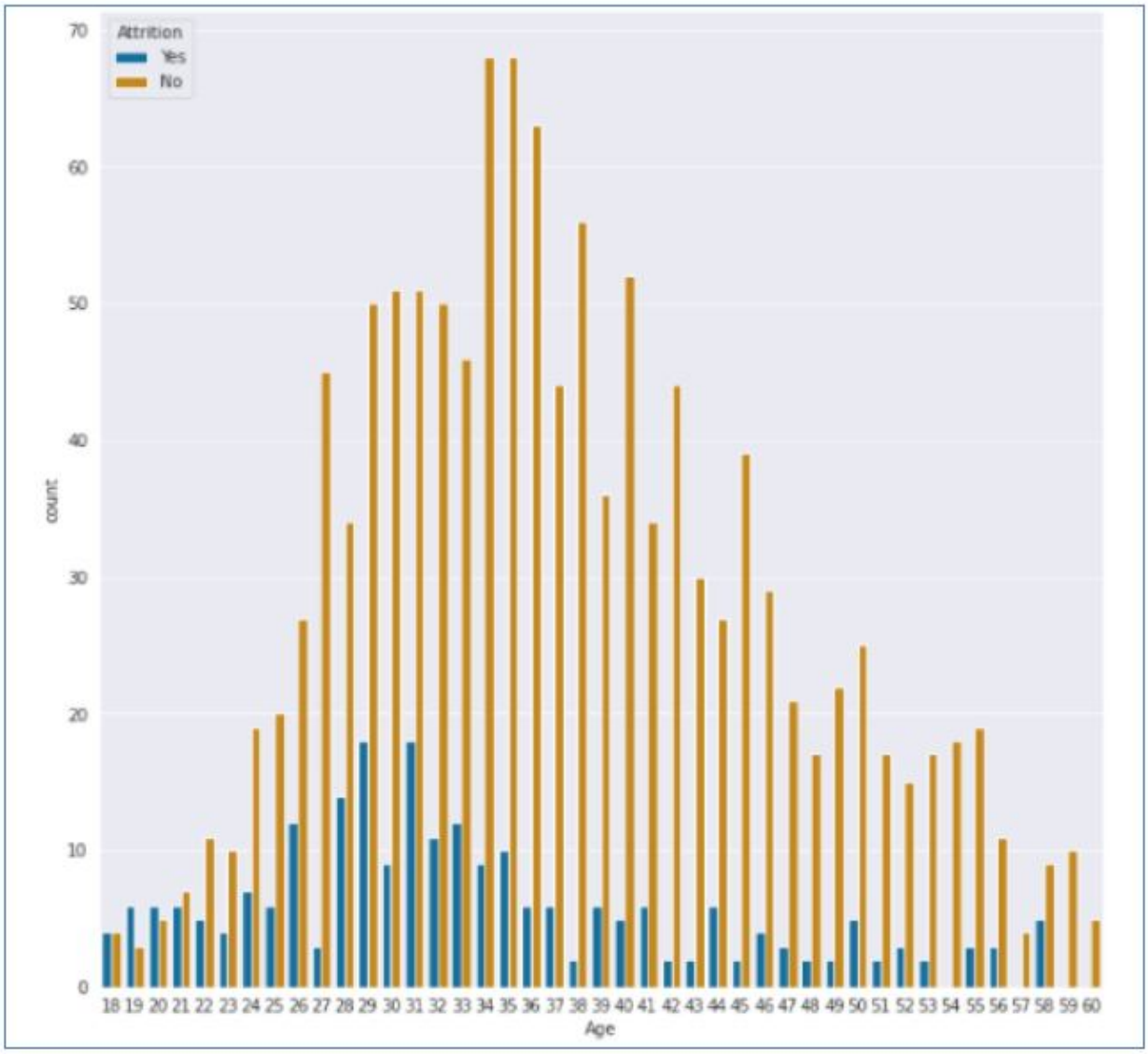

Figure 5: Relation between Age and Attrition

5. CONCLUSION

Using the three boosting algorithms namely CatBoost, Light GBM, XGBoost with 65% train and 35% test dataset split the LightGBM gave us the more accurate model than other two algorithms. When K Fold validation is used for algorithms namely CatBoost, LightGBM, XGBoost we have got more accurate model with 90.47% accuracy for both the algorithms namely Cat Boot and XGBoost when K=10 in K Fold validation. So from our study we have got more accuracy than the SVM classifier, Decision. Trees, KNN, Random forest as mentioned in the some of the research papers as discusses in Literature Survey.

REFERENCES

1. https://www.researchgate.net/publication/326029536_Employee_Attrition_Prediction
2. https://www.irjet.net/archives/V7/i5/IRJET-V7I5737.pdf
3. http://www.jicrjournal.com/gallery/24-jicr-december-2247.pdf
4. http://www.iosrjournals.org/iosr-jbm/papers/Vol20-issue2/Version-4/A2002040127.pdf
5. https://ieeexplore.ieee.org/document/8605976
6. https://ieeexplore.ieee.org/abstract/document/8746940
7. https://ieeexplore.ieee.org/document/6216220
8. https://ieeexplore.ieee.org/document/8541242
9. 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", www.kaggle.com, 2016.
12. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attritiondataset
13. https://ieeexplore.ieee.org/document/9033784