Employee Attrition System Using Tree Based Ensemble Method

Vimoli Mehta
Electronics and Communication Engineering
Institute Of Technology,Nirma University
Ahmedabad, India
vimolimehta@gmail.com

Shrey Modi
Information Technology
Charotar University of Science and Technology
Changa, India
modishrey007@gmail.com

2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4) | 978-1-6654-2013-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/C2I454156.2021.9689398

Abstract—**Employee attrition has become a vital problem across the world. It is one of the crucial issues faced by business leaders within companies where they lose the most talented employees. A good employee is always an asset to the organization and their resignation can lead to various problems like financial losses, overall performance, and loss of acquired knowledge. Furthermore, hiring new employees is far exorbitant, taxing, and time-consuming in comparison to recruiting the existing one. It is very time-consuming to recruit a new employee as it takes him months for training, adjusting to the culture, rules, and environment. Therefore, upcoming trends and technology using Machine Learning Algorithms must be exploited for the benefit of business organizations. Knowing the reason beforehand for the employee attrition, companies can mitigate this loss. This paper provides a conclusive review of employee attrition using the tree-based Ensemble Machine Learning Model from the dataset 'IBM HR Analytics Employee Attrition Performance'.A collection of statistically significant factors which connect to an employee's decision to leave are identified. The paper evaluates the tree-based ensemble to get the best results from the existing tree methods.**

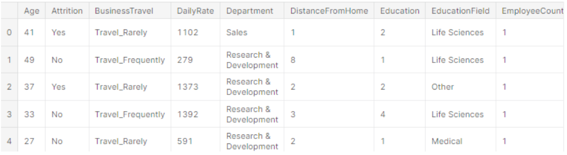
Index Terms—**Employee Attrition prediction, random forest, gradient boost, machine learning, ensemble model**

I. Introduction

Employee attrition can be expounded as the resignation of employees. Every organization needs to have a certain percentage of attrition to ensure the growth of the organization.[1] However, the success rate of the company is directly affected by losing highly-skilled employees. This occurrence can be of two types: Involuntary where the higher authorities fire the workers because of their unsatisfactory performance, efficiency, or disrespectful behavior. The latter is known to be Voluntary, where employees leave by their own volition. The common reasons are low remuneration, lack of opportunity in terms of growth, platitudinous life with the same work routine, lack of employee engagement with the superiors with management, or increased incentives offered by other firms. This negative attrition can lead to a decrease in a customer-client base, which diminishes the selling factor and it takes time and money in locating a qualified applicant and providing sufficient work training. Predicting employee attrition at a company can be beneficial in determining the factors which motivate the employees to leave. This prediction can be performed by analyzing the basic data(information) of the employee, which is possessed by the Human Resource (HR) department. The model helps act faster and prepare contingency plans based on the acumen collected. To make an effective retention plan, we need ample drivers such as pay, promotion, commute distance, relationship with colleagues, and satisfaction ratio. By diagnosing the following, we can formulate retention strategies and take preventive measures to provide employees a robust incentive to stay. The prediction of employee turnover also helps to indicate financial loss and production during attrition and analyses the trends of a candidate's performance from past data which can influence the future trends using data analytics and machine learning. Tree-based ensemble techniques, including the random forest(individual model) and gradient boost(ensemble), helps to improve the accuracy of the employee attrition prediction.

II. Literature Overview

Many types of research have been made on the given topic using different algorithms, which are explored below. The given algorithms get updated over time with the change in parameters contributing to the decision of employees to resign. For instance,[3] uses various models such as SVM, Decision Trees, random forest, and KNN on the same IBM dataset. It evaluates the following by plotting visualization graphs of gender, travel, overtime vs Attrition. Based on the given values, it predicts the employee turnover which turns out to be the greatest in Random Forest with an accuracy of 88.43 percent in comparison to other classifiers. Therefore, we have included the given method in our paper. In [2], the paper predicts the accuracy of five base models and later combines them to have a stronger predictor model which is often known as ensemble learning. The paper combines decision trees with linear regression to get an accuracy of 86.39 percent among others such as Adaboost and Random Forest and third as SVM and gradient boosting. In [4], the paper works on a dataset available on Kaggle and works on well-known classification techniques, namely, Logistic Regression, Support Vector Machine (SVM), Random Forest, Decision Tree, AdaBoost, and Neural Network. It focuses on 12 main attributes uses three methods to derive accuracy that is Brute Force method, One Hot Encoding, and the Feature selection approach. The accuracy achieved is highest in the Feature



	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1

Fig. 1. Dataset Overview

Selection Approach with algorithm Random Forest which is a good accuracy achieved. Therefore, the authors of the paper decided to explore the results of ensemble trees like Random Forest and Gradient Boosting by using a feature selection approach.

III. Methodology

A. Data Extraction

Collecting data is a laborious task and it takes a significant amount of time to collect the data. Data Collection takes numerous features to be fulfilled which would help the algorithm to grasp the data and perform with better accuracy. It is simpler to collect data if one runs a firm with thousands of employees; but, the writers, who are students, are unable to acquire data since employee information is confidential. Therefore, we have selected a vast dataset created by IBM named 'IBM HR Analytics Employee Attrition Performance'. The dataset consists of 1470 different observations and 35 unique features which were asked to the employees of IBM as a questionnaire to get their responses. A glimpse of the dataset is shown in figure 1.

B. Data Preprocessing

Data preprocessing is an important aspect of the performance of the algorithm or the model. The speed of the algorithm depends on if the preprocessing of the data has been done or not. If done correctly, it can make the algorithm work quicker, and if applied to a large dataset, it will perform efficiently. To start with it, the first step would be to review the data and do a quality check on it. For this, the authors simply invoked a null call to determine whether any feature had a missing value or not, and the result was received as a boolean value (yes or no). If there is a null character, it must be filled with random data, ensuring that the dataset is full and free of missing values. The authors received all the features as false, so no data was missing. After that, the transformation of categorical data to dummies takes place. In a Categorical Variable, Dummy Variables serve as indications of the existence or absence of a category. The standard approach is that 0 denotes absence and 1 denotes presence. When categorical variables are converted to dummy variables, a two-dimensional binary matrix is created, with each column representing a different category. The dataset would then be made unique by removing duplicate characteristics and duplicate rows, ensuring that nothing was replicated. Finally, feature scaling would be used to normalize the independent characteristics included in the data in a given range; the dataset's age ranges from 18 to 60 years old, and the monthly income ranges from $1,000 to$10,000.from $2,094 to$26,999. If feature scaling isn't done, a machine learning algorithm will assume larger values to be higher and smaller values to be lower, regardless of the unit of measurement. Hence, normalization and standardization were performed on the dataset.

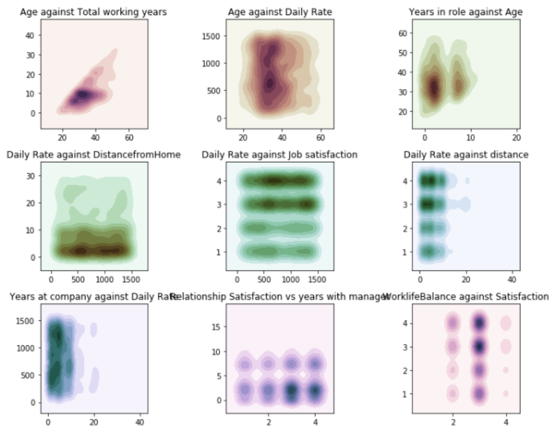


Fig. 2. KDE Plot

C. Features visualization

Feature selection is an important aspect of the model, the accuracy depends on the features infiltrated in the model. The more features, the more will be the accuracy of the model. To see the distribution of the data to the features of the dataset, the KDE(Kernel Density Estimate) plot is visualized to get the exact density of the dataset allocated in the features with each other. The KDE plot visualization can be seen in figure 2. The next tool in a data explorer's arsenal is that of a correlation matrix. We can get a good idea of how the characteristics are connected by creating a correlation matrix. A correlation matrix(Figure 3) is generally a table showing correlation coefficients between the variables. It is also used to summarize data as an input into more advanced analysis. So, from the correlation plot, we get to know that quite a lot of our columns seem to be poorly correlated with one another. Generally, when making a predictive model, it would be preferable to train a model with features that are not too correlated with one another so that we do not need to deal with redundant features. In the case that we have quite a lot of correlated features, one could perhaps apply a technique such as Principal Component Analysis (PCA) to reduce the feature space.PCA is a statistical method that transforms a collection of correlated variables into a set of uncorrelated variables via an orthogonal transformation. In exploratory data analysis and machine learning for predictive models, PCA is the most commonly used method. Hence, PCA tackles the situation of the correlation very well and helps us improve the model.

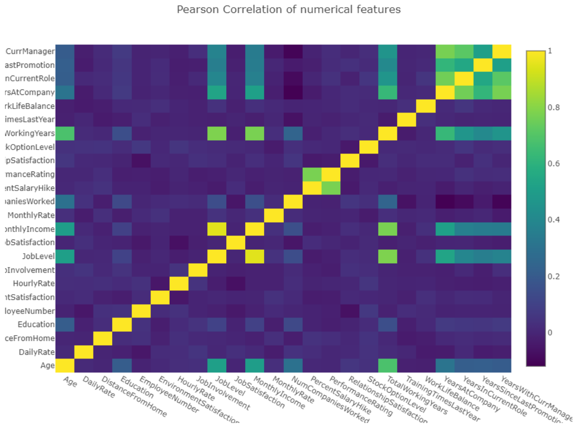
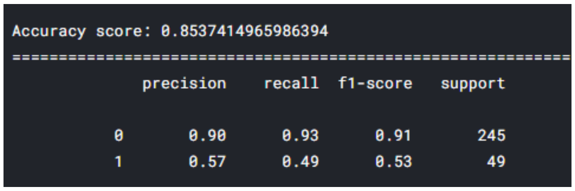


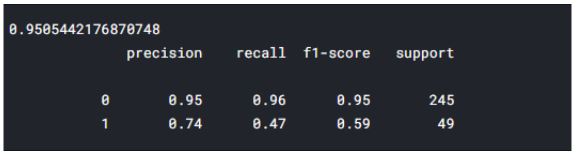
Fig. 3. correlation matrix



```
Accuracy score: 0.8537414965986394
=============================================================
              precision    recall  f1-score   support


           0       0.90      0.93      0.91       245
           1       0.57      0.49      0.53        49
```

Fig. 4. Random Forest Result



```
0.9505442176870748
              precision    recall  f1-score   support


           0       0.95      0.96      0.95       245
           1       0.74      0.47      0.59        49
```

Fig. 5. Gradient Boosting Result

D. Model Architecture

1) Random Forest: It is a method that comes under the group of ensemble models which exploits the bagging technique. It can be used for classification and regression to get a good predictive model. The base estimators used here are decision trees. Decision trees on their own act as poor predictors, but when combined with groups of decision trees, it becomes a stronger predictor. The decision trees vote on how to classify a given instance of input data and output the class that is the mode of the classes in case of classification tasks or the mean of predictions in the case of regression tasks [2]. By this, one can reduce overfitting without parameter tuning. The reason behind using random forests is it can also handle large datasets efficiently.

2) Gradient Boosting: Gradient boosting is similar to random forest where there is a grouping of weak learners to form a stronger one however, gradient boosting differs from random forest since it adds its predictors in a sequential order which means no further improvement is possible before. The gradient of descent aids the gradient booster in identifying and correcting errors in learners' predictions. It is memory efficient as well as good in speed but fails to visualize with perfection in comparison to decision trees and linear regression.

E. Building the model

The basic architecture of the model which we will employ is discussed above in the model architecture. To start building the models, we need to play with the dataset a bit to obtain the desired output. Train-test-split is a very common method used to estimate the performance of the algorithm. This requires partitioning the dataset into training and testing data, which is accomplished with the help of the sklearn library. After that, we would be taking the models individually and performing the algorithm used in it to determine the accuracy of the taken models. The authors used scikit learn for performing both the models, in random forest we would determine the set of parameters that we will feed into our Random Forest classifier. Having defined our parameters, we can initialize a Random Forest object by using scikit-learn's RandomForestClassifier and unpacking the parameters by adding the double asterisks symbols, after that the process of building a forest of trees using our training set and fitting it to our attrition target variable is done, and finally we get the scores which would determine how the algorithm has performed. Similarly, we follow the same process for the gradient boost method.

IV. Experimental Results

As it can be seen, our Random Forest predicts with an accuracy of around 86 percent, and at first look, this appears to be a fairly decent model. When we consider how skewed our target variable is, where yes and no responses are distributed at 84 percent and 26 percent, our model only predicts marginally better than random guessing. In the categorization report outputs, balancing the accuracy and recall scores would be more helpful. It comes down to business considerations whether one measure should be prioritized over the other, i.e. Precision versus Recall. While in the gradient boost method which is an ensemble method it performs fairly better than the random forest algorithm as it not only increases the accuracy, it also balances out the accuracy and recall scores. From this ensemble technique, we are getting an accuracy of approx 95 percent, which is far better than the random forest method. The accuracy of both the models has been shown in figure 4 and figure 5.

V. Conclusion

Employee turnover is caused by a variety of factors, including compensation and other financial factors such as promotions. The study's main objective is to build reliable and accurate models that can help employers save money on acquiring and keeping talented people. This might be accomplished by employing proper data mining tools to determine the attrition status of the individual in question. The authors have constructed a very simple pipeline predicting employee attrition, from some basic Exploratory Data Analysis to feature engineering, as well as implementing two learning models in the form of a Random Forest and a Gradient Boosting classifier.The model achieved an accuracy of 86 percent for the random forest algorithm and 95 percent accuracy for the gradient boost algorithm, and hence, a tree-based ensemble has been performed with great accuracy which can be used by the HR industry for the employee.

VI. Future Work

The given model uses two approaches from which the Gradient Boost classifier provides more accuracy in comparison to Random Forest. The model can be further implemented on Adaboost classifier as it can reduce the loss function. Furthermore, the dataset used is a medium-sized dataset and one can increase the robustness of the model by including more important features that directly affect employee turnover. Another drawback that can be solved in future work is that the model is restricted to only supervised machine learning which takes a lot of computational time and new features have to be added every time as an input which takes a lot of human effort. The model can also further be extended by deploying more data visualization between parameters and attrition to get a better idea of the factors that influence the most.

References

[1] Joseph, Richard, et al. ”Employee Attrition Using Machine Learning And Depression Analysis.” 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2021.

[2] Qutub, Aseel, et al. ”Prediction of Employee Attrition Using Machine Learning and Ensemble Methods.” Int. J. Mach. Learn. Comput 11 (2021).

[3] Patel, Adarsh, et al. ”Employee attrition predictive model using machine learning.” International Research Journal of Engineering and Technology (IRJET) 7.5 (2020).

[4] Yadav, Sandeep, Aman Jain, and Deepti Singh. ”Early prediction of employee attrition using data mining techniques.” 2018 IEEE 8th International Advance Computing Conference (IACC). IEEE, 2018.

[5] Kakad, Shital, et al. ”Employee attrition prediction system.” Int. J. Innov. Sci., Eng. Technol. 7.9 (2020): 7.

[6] G. Marvin, M. Jackson and M. G. R. Alam, ”A Machine Learning Approach for Employee Retention Prediction,” 2021 IEEE Region 10 Symposium (TENSYMP), 2021, pp. 1-8, doi: 10.1109/TENSYMP52854.2021.9550921.

[7] Jain, P.K., Jain, M. Pamula, R. Explaining and predicting employees’ attrition: a machine learning approach. SN Appl. Sci. 2, 757 (2020). https://doi.org/10.1007/s42452-020-2519-4

[8] A. C. Patro, S. A. Zaidi, A. Dixit and M. Dixit, ”A Novel Approach to Improve Employee Retention Using Machine Learning,” 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), 2021, pp. 680-684, doi: 10.1109/CSNT51715.2021.9509601.

[9] A. Mhatre, A. Mahalingam, M. Narayanan, A. Nair and S. Jaju, ”Predicting Employee Attrition along with Identifying High Risk Employees using Big Data and Machine Learning,” 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 269-276, doi: 10.1109/ICACCCN51052.2020.9362933.

[10] Giovanni Seni; John Elder, Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions , Morgan Claypool, 2010.

[11] Alhashmi, Saadat M. ”Towards Understanding Employee Attrition using a Decision Tree Approach.” 2019 International Conference on Digitization (ICD). IEEE, 2019.

[12] G. Martínez-Muñoz, D. Hernández-Lobato and A. Suárez, ”An Analysis of Ensemble Pruning Techniques Based on Ordered Aggregation,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 245-259, Feb. 2009, doi: 10.1109/TPAMI.2008.78.

[13] Pratt, Madara, Mohcine Boudhane, and Sarma Cakula. ”Employee Attrition Estimation Using Random Forest Algorithm.” Baltic Journal of Modern Computing 9.1 (2021): 49-66.

[14] V. Vijay Anand, R. Saravanasudhan and R. Vijesh, ”Employee attrition - a pragmatic study with reference to BPO industry,” IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012), 2012, pp. 769-775.

[15] A. C. Patro, S. A. Zaidi, A. Dixit and M. Dixit, ”A Novel Approach to Improve Employee Retention Using Machine Learning,” 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), 2021, pp. 680-684, doi: 10.1109/CSNT51715.2021.9509601.

[16] Modi, Shrey, and Mohammed Husain Bohara. ”Facial Emotion Recognition using Convolution Neural Network.” 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2021

[17] R. S. Shankar, J. Rajanikanth, V. V. Sivaramaraju and K. V. S. S. R. Murthy, ”PREDICTION OF EMPLOYEE ATTRITION USING DATAMINING,” 2018 ieee international conference on system, computation, automation and networking (icscan), 2018, pp. 1-8, doi: 10.1109/ICSCAN.2018.8541242.