Employee Attrition Prediction Using Machine Learning Algorithms

by Anbarasa Kumar A _

Submission date: 13-Jan-2023 11:35AM (UTC+0530)
Submission ID: 1992162077
File name: rition_Prediction_Using_Machine_Learning_Algorithms_Updated.docx (181.68K)
Word count: 3625
Character count: 19815

Employee Attrition Prediction Using Machine Learning Algorithms

Rajat Singh, Hrishikesh S G
Department of MCA, School of Information Technology and Engineering
Vellore Institute of Technology, Vellore – 632014, Tamil Nadu, India.
Anbarasa Kumar A.
School of Information Technology and Engineering
Vellore Institute of Technology, Vellore – 632014, Tamil Nadu, India
E-mail: rajat.singh2022@vitstudent.ac.in, gaikwad.hrishikesh2022@vitstudent.ac.in

Abstract

Machine Learning is a branch of AI which leverages data to imitate the pattern by which a human brain learns, in the process improving accuracy. Statistical principles are used to process data and learn from it efficiently. Machine Learning algorithms are extensively used to predict future outcomes based on past experiences that were recorded in a meticulous manner. These past experiences or data can help us to deduce preliminary insights about the data and what it represents. The current paper discusses certain Machine Learning algorithms where the prime objective is to classify the inputs into one of the two categories. The dataset in focus is the employee attrition dataset that gives various insights regarding the presumable reasons behind an employee leaving the job. The factors such as precision score, accuracy, f1_score and recall score for Random Forrest, XgBoost, Adaboost, Gradient boosting and Decision Tree Classifier have been ascertained and compared. Furthermore, Hyperparameter tuning, using the 'RandomSearchCV' python library is also implemented on the better performing algorithms, with the goal of achieving better performance.

Keywords: Machine Learning, Random Forrest, XgBoost, Adaboost, Decision Tree, Gradient Boost, Hyperparameters, RandomSearchCV.

Introduction

Employee attrition is when an employee of an organisation, leaves the company due to various reasons. It may be because his or her performance was not up to the mark. Or due to incompatibility with their colleagues and more so on. If the attrition happens with the will of the organisation then it is supposed to be fine. But if an organisation loses its valuable employee because the employee wanted it so, then it is a problem. The organisation might have gone through a long process of hiring and training of that employee. Both time and money of the organisation is invested on a particular employee, and when this employee leaves then it tends to be a concern for the employer. The organisation will have to do all that investments of time and resource again to hire another employee to replace the one that left. This problem can be analysed and to a great extent be solved by using machine learning algorithms. To do such a task, the most preliminary tool required to get started is the past data of the employee turnover along with ones who were retained. Machine learning models learn from the past data, or we might say, past experiences to predict the future, so if we feed this data consisting of employees that stayed and the employees that left, we might get a model that takes the employee data as input and tells if the employee is going to leave or not. Further analysis of this model will give us insights of what are the plausible factors leading to attrition of employees. This will help the management to take steps in the right direction

Previous Work and their short comings

Several researchers have investigated using ML algorithms to forecast employee behaviour. The issue of employee attrition has been the subject of research for several decades. Every organization experience staff attrition. Individuals either retire or resign. If this does not take place in a timely manner and if staff depart without notice, it may have serious repercussions for the continued existence of the organization.

Employee turnover might be seen as a theft of the company's intellectual property. The previous work focuses on the methods and strategies that various scholars have put forth for predicting employee attrition.

To forecast employee performance, the authors utilized a variety of techniques, including decision trees, the Naïve Bayes classifier, Random forests, Gradient Boost, SMV, and many more with various factors such as salary, job satisfaction, designation, age, gender, etc.

Previous studies used several machine learning algorithms and datasets to present various accuracy estimates. However, many researchers have focused on factors that are unrelated to employee attrition, and it has been noted that hyperparameter tuning was not done for datasets on employee attrition. The primary and most important objective of this research is to provide a thorough demonstration, description, and evaluation of machine learning algorithms for identifying attrition by using numerous significant factors and hyperparameter tuning.

Machine Learning

The procedure of using computational methods to learn information directly from the data. This learning should be done without relying too much on a predetermined equation. This whole process is called Machine learning. The data samples used to train the model is very important. When the data models increase, the accuracy of this model might also increase. Since the process of Machine learning largely involves a lot of statistical computations, the data used to train such a model needs to be clean and relevant to the intended goals of the project. The data used should be consistent, that is, it should not have too many null values and understandingly the tabulated data should be correct. The relevance of the data used is important because if the data is not relevant than the model will predict wrongly for real world cases. For example, if we need a model to predict a disease in an individual, we need the medical data of that person rather than the data about his TV watching habits. The TV habits data can be useful for the model where subscription to a TV plan is predicted. There are many machine learning algorithms that use different statistical computation techniques to train a model. We have used some of these models during this research.

Hyperparameter Tuning

Machine Learning models are mathematical models associated with several parameters. These parameters are to be learned by the same machine learning model. These parameters are learnt by the model by training the model with the existing data. This process is called fitting the data to the model. There is another kind of parameters that cannot be directly learned by a model through the regular training process, known as Hyperparameters. These parameters are fixed even before the actual training process begins. Significant properties of a model is expressed by such parameters. The complexity or how fast a model must learn is also ascertained by Hyperparameters. In practice, there are two main strategies that are used to implement Hyperparameter tuning, Randomized Search and

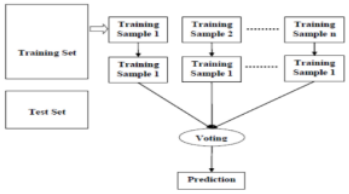
Grid search. For grid search technique, all the possible hyperparameters including the intermediate combinations of the hyperparameters are tried, that is, for each hyperparameter a new model is created. This is the reason why grid search is more computationally expensive. The Randomized search Tuning strategy is better than the "GridSearch" strategy because it tries only a fixed number of hyperparameters for the model. For this research, the "RandomizedSearchCV" python module is used.

Decision Tree

In each decision tree, the internal nodes stand for the features of the data set, leaf nodes for outcomes and branches for decision rules. All in all, this classifier is an ML algorithm which generates a model that can classify using a tree data structure. It is easy to implement and comprehend because tree structures are easy to understand. A tree structure also mirrors a human being's decision-making process, that is, the process of choosing one of the options at each step until the goal is reached or final decision is made. A decision tree starts with a question and based on the possible outcomes; it splits the tree at every level. The main issue in the decision tree algorithm is to choose the best attributes for the root node and the sub nodes. To do this task, we use the technique called ASM. With ASM the best possible attribute can easily be selected for the nodes of the tree.

Random Forest

In some cases, the problem is such that a single decision tree might not be enough to solve the problem. Random Forest consists of multiple decision trees associated with logical parts of the dataset. The Average is taken at the end and then checked if we are getting better accuracy. In most cases we get better accuracy. Random Forest is best suited for problems related to regression and classification. Random forest comes under the umbrella of ensemble learning. The process of using different classifiers together to tackle a complex problem and to improve the all over accuracy of the solution is called Ensemble learning. This is obvious that a greater number of trees in the forest will lead to better accuracy but also take a toll on performance. Random Forest is useful because it takes less time to train, predicts the output with high accuracy. It performs well even when a part of the data is missing.



XG-Boost

XgBoost stands for Extreme Gradient Boosting. Its library is written in C++, so it optimizes the Gradient boost training process. As the name suggests, XgBoost attempts to boost the Gradient Boosting model. Boosting is an ensemble modelling technique where several weak classification models are combined to form a stronger classification model. This processes simply adds weak models one after the other. At the beginning model is created on the data and then a second model is added in series. The second model tries to correct the errors made by the model previous to the current one. This cycle repeats itself until the maximum number of models is added or the complete data set is predicted. For

XgBoost, the multiple decision tree creation follows a sequence. Weights are important part of this process. The independent variables that are fed to the decision tree which predicts results are assigned with weights first. If an independent variable is predicted wrongly, the weight of that variable is increased and then fed to the next decision tree. These individual trees or classifiers are then ensembled together to obtain a more precise and stronger model.

Proposed approach

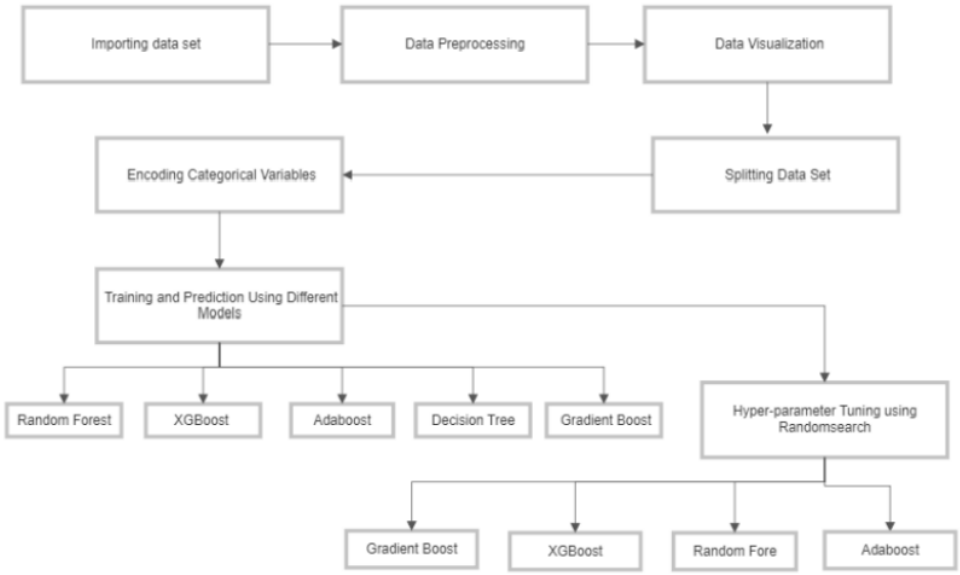
The dataset in use for this research has a total of 35 variables. These 35 variables are related to an employee working in a company. The dataset contains the data of employees who are not working in the company anymore with the ones who still work there. The percentage of employee in the dataset who left the company is 16 percent. Before using this data to train a model, the data must be put through some pre-processing stages. First the unwanted variables are removed from the dataset. As we want to train the model to predict employee attrition, we will make sure that the data does not have any variable that is directly related to employee attrition. The variables that were removed are: Employee count, standard hours, over18, employee number, and.

Now, the test and train split of the dataset is done. The training data makes of 70% and the test data makes of 30% of the entire dataset. Training data will be fed to the model so that it forms patterns for prediction. The testing data shall be used to evaluate the model.

The next step in pre-processing of data is encoding the categorical variables. For this, Ordinal and One-Hot encoder is used. Ordinal encoder is used when there is assumed ordering of categories. In other words, through the Ordinal encoder, we can show a hierarchy among the possible values of a variable for every tuple. So, for instance when encoding relationship satisfaction, the encoding will assume that low (0) is lesser than very high (3). On the other hand, the One-Hot encoder will create new columns indicating if the value of the categorical attribute is present or not, with binary values where 0 means absence and 1 means presence of such value.

Next step deals with balancing the target variable. It was noticed that the target variable is not balanced, that is, the employees who have left the company is much less than those who have not. To balance the variable, a python library called ‘SMOTE’ is used to synthetically create tuples that are similar but do not duplicate the already present tuples. This will balance the number of employees lost and retained equal.

The data is fed to the models: Random Forest, XgBoost, decision tree, gradient boosting and Adaboost. The training is done and the recall, F1_score, accuracy and precision is calculated by using the confusion matrix. After this, RandomSearchCV was used to tune the best performing models. After the hyperparameter tuning, again the models were evaluated using the confusion matrix.



Measure of Performance: Confusion Matrix

The table that lists the number of correct as well as incorrect guesses. The effectiveness of a classification model is required to judge its performance. The Confusion Matrix(CM) shows the recall, F1_score, accuracy, and precision to judge the performance of a classification model. **True positives are the values where the both the predicted and actual values are true. On the other hand, true negatives are the values where the actual value is false, and the predicted value is also false. False Positives are the values where the predicted value is false, and the actual value is true. False negatives are the value where the actual value is true and predicted value is false.**

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

CM for binary classification (2x2 matrix)

Accuracy is simply the frequency of correct predictions. It is the proportion between the number of accurate predictions and all predictions combined.

$$\text{Accuracy} = \frac{Correct\ Predictions}{Total\ Predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision indicates the level of accuracy attained in real predictions. Out of all the samples that really belong to the positive class, the proportion of samples that were accurately predicted.

$$\text{Precision} = \frac{Predictions\ Actually\ Positive}{Total\ Predicted\ Positive} = \frac{TP}{TP+FP}$$

Recall measures how well actual observations match predictions. It is also referred to as sensitivity.

$$\text{Recall} = \frac{Predictions\ Actually\ Positive}{Total\ Actual\ Positive} = \frac{TP}{TP+FN}$$

The harmonic mean(HM) of recall and precision is the F1 score. The F1 score is responsible to keep precision and recall the classifier in balance.

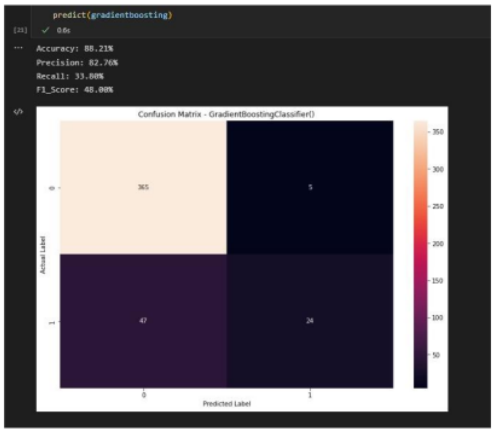
$$\text{F1-Score} = 2 * \frac{Recall * Precision}{Recall + Precision}$$

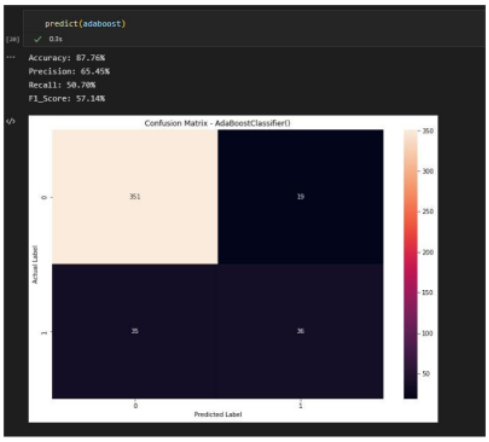
Results and discussions

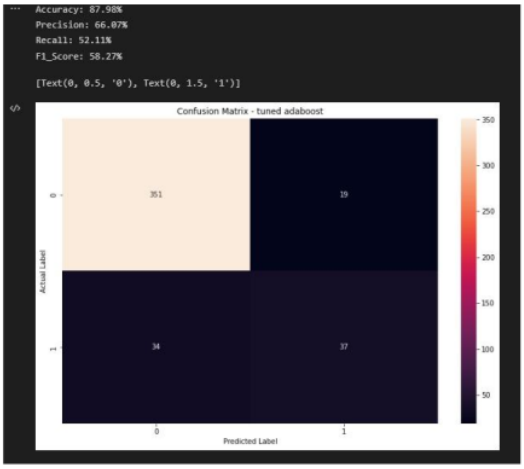
The exploratory data analysis conclusions for this research are twofold. When analysing the categorical variables, it is observed that most employees who left the company belonged to the Research and Development department, with most of them being laboratory technicians, sales executives, or research scientists. It was also observed that the employees who left the company scored excellent performance ratings. It is not good to lose such high-quality employees. Most of these employees had bachelor's degree and their education field was mostly life sciences, medical and marketing. Many employees showed high work involvement along with the dissatisfaction with the work environment. Looking at the attrition per age histogram, its noticeable that as the age of an employee increases, the lesser are the chances for such employees to leave. Most of the attrition is made in the ages ranging between 25 to 35. The data also indicates that if an employee invest more years in a company and at a same role, he/she is less likely to leave. Talking about incomes, most employees who have left belong to the category of smaller income employees. Also, those who have less percentage salary hike also tend to leave more than those with a higher percentual salary hike.

In this research there are 2 cycles of training and evaluation of the data. First cycle is regular, second cycle involves the tuned data. Data is tuned using Hyperparameter tuning explain above. For evaluation, confusion matrix is being used. Confusion matrix talks about the recall, F1_score, accuracy and precision of the model. Out of these performance measures Recall is important for this research. Considering the main goal to identify the employees that are more susceptible to voluntary attrition, the recall score is the one in focus.

For the first cycle, gradient boosting gave the best accuracy score (88.21%) while ada-boosting had the best recall score (50.70%). For second cycle, that is, with hyperparameters, ada-boosting gave the best accuracy (87.98%) and the best recall (52.11%).







Conclusion

High staff turnover rate is a big issue for any organisation. When a high-performance employee leaves the company, it is very difficult to find a replacement for that employee. If that employee was high performing, that means a lot of resources had been invested in his/her training. To replace such an employee, this cycle must be performed again which makes if very inefficient. As the possibility of successors is quiet low, it is imperative that companies should look at ways to make the work environment such that it is easy for an employee to work in a company/organisation for long. The main goal of this research is to train the different ML models and evaluate their performances. The comparison of performances of different models is also essential. The findings of this study demonstrate that algorithms for data extraction can be used to create precise and trustworthy models for employee attrition forecasting. According to the recorded results, Ada-boost is the best performing model as it has the best recall score at 52.11% after hyperparameter tuning. In the future, data to be trained to the model should be from a company that works in a totally different field. Furthermore, the possibility of practical utilization of these models should be explored extensively.

References

[1] Shankar, R.S., Rajanikanth, J., Sivaramaraju, V.V. and Murthy, K.V.S.S.R., 2018, July. Prediction of employee attrition using datamining. In 2018 ieee international conference on system, computation, automation, and networking (icscan) (pp. 1-8). IEEE.

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

[3] Alduayj, S.S. and Rajpoot, K., 2018, November. Predicting employee attrition using machine learning. In 2018 international conference on innovations in information technology (iit) (pp. 93-98). IEEE.

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

[5] Martin, L., 2020. How to retain motivated employees in their jobs?. Economic and Industrial Democracy, 41(4), pp.910-953.

[6] Jhaveri, S., Khedkar, I., Kantharia, Y. and Jaswal, S., 2019, March. Success prediction using random forest, catboost, xgboost, and adaboost for kickstarter campaigns. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1170-1173). IEEE.

[7] Kabiraj, S., Raihan, M., Alvi, N., Afrin, M., Akter, L., Sohagi, S.A. and Podder, E., 2020, July. Breast cancer risk prediction using XGBoost and random forest algorithm. In 2020 11th international conference on computing, communication, and networking technologies (ICCCNT) (pp. 1-4). IEEE.

[8] Bardenet, R., Brendel, M., Kégl, B. and Sebag, M., 2013, May. Collaborative hyperparameter tuning. In International conference on machine learning (pp. 199-207). PMLR.

[9] Schratz, P., Muenchow, J., Iturritxa, E., Richter, J. and Brenning, A., 2019. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling, 406, pp.109-120.

[10] Shi, X., Wong, Y.D., Li, M.Z.F., Palanisamy, C. and Chai, C., 2019. A feature learning approach based on XGBoost for driving assessment and risk prediction. Accident Analysis & Prevention, 129, pp.170-179.

[11]Alhashmi, S.M., 2019, November. Towards Understanding Employee Attrition using a Decision Tree Approach. In 2019 International Conference on Digitization (ICD) (pp. 44-47). IEEE.

[12]Sisodia, D.S., Vishwakarma, S. and Pujahari, A., 2017, November. Evaluation of machine learning models for employee churn prediction. In 2017 international conference on inventive computing and informatics (icici) (pp. 1016-1020). IEEE.

[13]Hebbar, A.R., Patil, S.H., Rajeshwari, S.B. and Saqquaf, S.S.M., 2018, May. Comparison of machine learning techniques to predict the attrition rate of the employees. In 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 934-938). IEEE.

[14]Dubey, R. and Bisht, G., 2009, April. Key Result Employee (KRE) Retention:" Entrapping the Mammoth". In 2009 International Association of Computer Science and Information Technology-Spring Conference (pp. 272-275). IEEE.

[15]Brockett, N., Clarke, C., Berlingerio, M. and Dutta, S., 2019, December. A system for analysis and remediation of attrition. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 2016-2019). IEEE.

[16]Singh, M., Varshney, K.R., Wang, J., Mojsilovic, A., Gill, A.R., Faur, P.I. and Ezry, R., 2012, December. An analytics approach for proactively combating voluntary attrition of employees. In 2012 IEEE 12th International Conference on Data Mining Workshops (pp. 317-323). IEEE.

[17]Joseph, R., Udupa, S., Jangale, S., Kotkar, K. and Pawar, P., 2021, May. Employee Attrition Using Machine Learning And Depression Analysis. In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1000-1005). IEEE.

[18]Jain, R. and Nayyar, A., 2018, November. Predicting employee attrition using xgboost machine learning approach. In 2018 international conference on system modeling & advancement in research trends (smart) (pp. 113-120). IEEE.

[19]Mhatre, A., Mahalingam, A., Narayanan, M., Nair, A. and Jaju, S., 2020, December. Predicting employee attrition along with identifying high risk employees using big data and machine learning. In 2020 2nd international conference on advances in computing, communication control and networking (icacccn) (pp. 269-276). IEEE.

[20]Alduayj, S.S. and Rajpoot, K., 2018, November. Predicting employee attrition using machine learning. In 2018 international conference on innovations in information technology (iit) (pp. 93-98). IEEE.

[21]Zhou, N., Gifford, W.M., Yan, J. and Li, H., 2016, June. End-to-end solution with clustering method for attrition analysis. In 2016 IEEE International Conference on Services Computing (SCC) (pp. 363-370). IEEE.

[22]Ray, A.N. and Sanyal, J., 2019, October. Machine learning based attrition prediction. In 2019 Global Conference for Advancement in Technology (GCAT) (pp. 1-4). IEEE.

[23]Sadana, P. and Munnuru, D., 2022. Machine learning model to predict work force attrition. In Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications (pp. 361-376). Springer, Singapore.

[24]Bhartiya, N., Jannu, S., Shukla, P. and Chapaneri, R., 2019, March. Employee attrition prediction using classification models. In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT) (pp. 1-6). IEEE.

Employee Attrition Prediction Using Machine Learning Algorithms

ORIGINALITY REPORT

3%	2%	3%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.i-scholar.in Internet Source	1%
2	Sahil Sharma, Vijay Kumar. "Voxel-based 3D face reconstruction and its application to face recognition using sequential deep learning", Multimedia Tools and Applications, 2020 Publication	1%
3	"Advances in Computing and Data Sciences", Springer Science and Business Media LLC, 2022 Publication	1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		