Predicting Employee Attrition using various techniques of Machine Learning

Gargi Dongre [1]

[2]*Department of Computer Science and Engineering, MIT School of Engineering*
MIT Arts Design and Technology University, Pune, India

---***---

Abstract - *Each and every employee is the most precious asset of a company. It is only because of the employees that an organisation is able to run smoothly and hence Employee attrition is one of the key metrics that the comapnies are focusing on these days. Attrition may sometimes occur due to unavoidable circumsatnces such as transfer to a different city, retirement etc. But when the attrition starts causing a hole in the pockets of a business it needs to be monitored.A business spends huge amounts of its resources while hiring employees.To overcome the process of rehiring and to maintain a strong workforce the analysis of systematic machine learning models need to be adapted from which a suitable model can be chosen that measures the risk of attrition. This not only helps in saving resources of a business but also helps to maintain an equilibrium in the workforce.*

Key Words: **Employee, Attrition, Machine Learning, Analysis.**

1. INTRODUCTION

An employee is a boon to any company. Every employee who joins a business is bound to leave it at some point of time due to various reasons. Attrition can be thus defined as the exit of any employee due to avoidable or unavoidable circumstances including retirement, death, transfer, better opportunities, etc. The organization spends lots and lots of time and resources when hiring an employee. When employee's departure starts to affect the business in a negative way, it becomes a topic of concern for everyone in the business but especially for the HR. Due to the exit of skilled employee's the business not only loses its skilled professionals but also needs to rehire and train the new person. This makes its workforce weaker thus affecting the business as a whole. Due to increased globalization, especially post pandemic era, there has been a vast number of opportunities in every field. Due to better opportunities and for further growth an employee decides to depart from one business and joins another. This attrition influences a business in a negative way for a brief period of time. To maintain the manpower and to reduce costs Artificial Intelligence can be incorporated to predict the attrition.

This paper discusses about the various methods that can be used to predict employee attrition and also analyses the best possible solution with the help of model comparison.

Fig 1 shows the various reasons due to which an employee may decide to leave an organization.

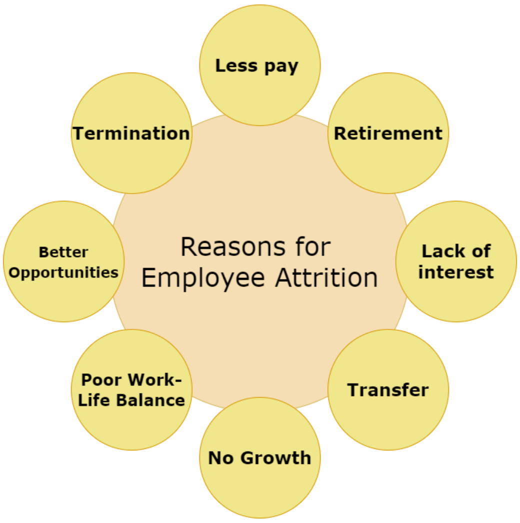

Figure 1.1: Reasons for Employee Attrition

2. Literature Survey

Many researchers have studied the causes and effects of employee attrition. One such paper states that [1] the maintenance of skilled and deserving employee's is a significant aspect that the HR needs to pay attention to. The study pointed out the most appropriate metrics which could help in the prediction of attrition. It highlighted that the number of job opportunities is directly proportional to employee's education and experience level. It also stated that some of the most agreeable factors that help in maintaining a workforce include good work-life balance, healthy workplace relationships, better policies, etc. Another such paper [2] states that in order for an organization to maximize its profits, it should give utmost importance and value to its employee's. This can be achieved by focusing on the development of opportunities and by bringing in new technologies that helps in maintaining the interest of an employee towards an organization. The study also highlights that it is necessary for an organization to conduct training programs, cultural events, etc. on a regular basis. These types of activities help in lowering the barrier of communication and also facilitates interaction and growth. The main idea of this study was to explain that it is necessary for an organization to have transparent work culture so that every person is well informed about their job and its outcome.

Artificial intelligence has led to exponential growth in all the fields. It has helped in finding solutions to many complex problems. Employee attrition is one such problem which is in talks these days. Artificial Intelligence has the ability to give a robust solution for this problem to various organizations. The incorporation of machine learning to predict attrition is helping companies worldwide. Similar research has been done where various models such as Support Vector Machines, Random Forest, KNN classifier, XG Boost are tried and tested. Table 1.0 depicts the information about some such researches.

Sr no.	Author	Object of Study	Recommend Technique
1.	Rahul Yedida, Rahul Reddy, Rakshit Vahi, Rahul J,Abhilash and Deepti Kulkarni[3]	Employee Attrition Prediction	KNN classifier
2.	B. Sri Harsha, A. Jithendra Varaprasad, L.V N Pavan Sai Sujith[4]	Early Attrition Prediction	Random Forest
3.	Yue Zhao, Maciej K. Hryniewicki, Francesca Cheng, Boyang Fu and Xiaoyu Zhu [5]	Prediction of Employee Turnover using Machine Learning	XG Boost
4.	Adarsh Patel, Nidhi Pardeshi, Shreya Patil, Sayali Sutar, Rajashri Sadafule and Suhasini Bhat[6]	Predictive model for Employee Turnover using Machine Learning	Random Forest
5.	Ozdemir, Coskun, Gezer and Gungor [7]	Using data mining techniques to predict attrition	SVM

Table 1.0: Survey Table

3 Design, Architecture and Dataset

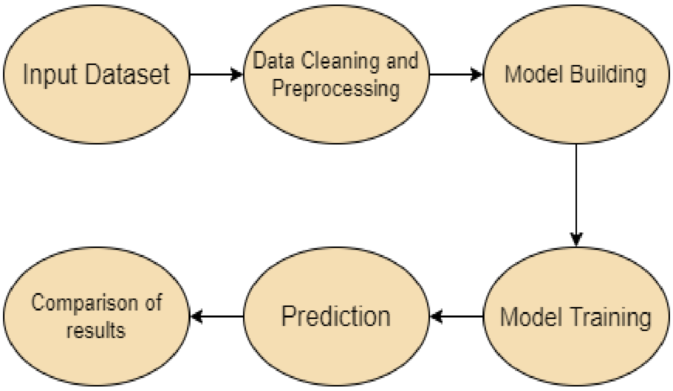

Fig 3.1: Design and Architecture

The above figure3.1 depicts the architecture of the system.The proposed system works on different models of machine learning.Each model attempts to predict attrition using the same dataset. The dataset consists of various employee records (both past and present).The input dataset is first cleaned and preprocessed by managing all the missing, Nan values , etc and removing unwanted columns.Then comes the model building phase where various models are taken into consideration for prediction.The dataset is then spilt into train and test sub dataset and the train set is used for training of each model used.All the predictions are compared on the basis of evaluation metrics and the best model is suggested.

The open source dataset consists of employee related information.All non – numerical values were given a designation (A1 ,A2 ,A3),etc. and all the unwanted parameters were discarded. Table 3.1 shows some parameters that are considered in the prediction.

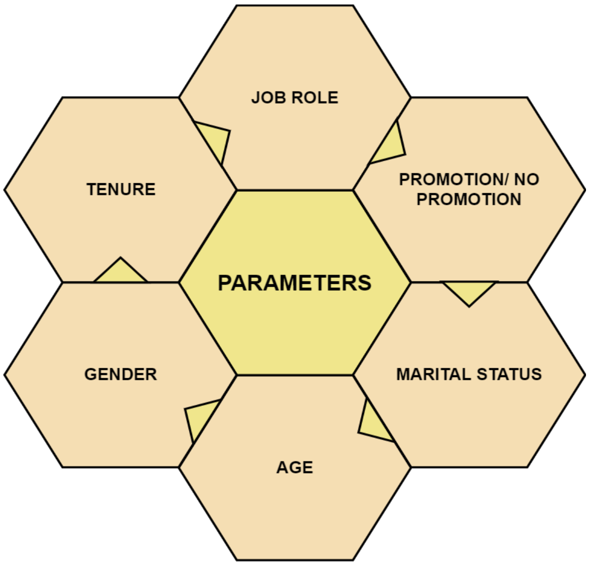

Fig 3.2: Parameters used in Prediction

4 Algorithms

A. Logistic Regression:

Logistic Regression is thought to be one of the most valuable statistical models. It is also a renowned data mining technique used by scientists and researchers for the analysis of proportional and binary kinds of datasets. One advantage that makes logistic regression special is that it has the ability to work for multi class problems as well [8].It is one of the most widely used algorithm for the purpose of classification.

Below is the equation that represents the equation for logistic regression:

$$log(y/(1-y)) = b_o + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_nx_n$$

where: y = dependent variable

x1, x2, x3…xn = independent variables

b0, b1, b2…bn = constants

B. Decision Tree:

When the word tree is used in computer jargon , the tree structure is visulaised. A decision tree consists of root, branches and leafs. The root node is considered to be the parent node. Every attribute is represented by nodes and the connection link between them are the branches. These branches are rules or decisions. The leaf is supposed to be the outupt or outcome.Some of the most commonly used decision tree algorithms include CHAID, ID3, CART [9].This algorithm is used for classification problems and can easily work with both continous and categorical values.

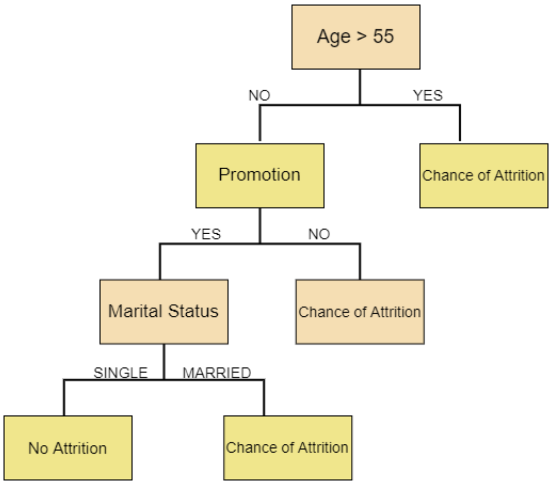

Fig 4.1: Decision Tree

C. K – Nearest Neighbours(KNN):

KNN is a supervised machine learning algorithm used for both classification and regression problems. The KNN algorithm uses the information about the input and predicts the output. The input is split into respective categories.The algorithm tends to search for the most optimal location for a new datapoint to lie in. The input data points are studied and the location for the new point is decided on the basis of it.Following is the algorithm used :

Step 1: Select number of neighbours (K)

Step 2: The Euclidean distance of K neighbours is calculated

Step 3: Identify K nearest neighbours by use of Step 2.

Step 4: Count number of points from each category.

Step 5: The new point is assigned to a category where the neighbours are more

Step 6: Finish

D. Support Vector Machines(SVM):

SVM is another type of widely used supervised machine learning model.It is mainly used for classification problems but can also be used for regression.The main idea of the algorithm is to create a line or a boundary which splits the space into n classes or categories. When a new data point is fed into this space it can easily search for its place in the created categories. The line that seperates these classes is also called as a hyperplane.When a straight line is enough for a problem of classification then the algorithm is linear.When a straight line is not sufficient and rather a crooked line is obtained then it is termed as non-linear SVM.

E. Random Forest:

Random forest is a machine learning algorithm used for regressions and classification type of problems.It is inherited from the concept of ensemble learning.It is similar to decision trees. This algorithm takes into consideration various trees by dividing the dataset to multiple subsets.Due to this method multiple results are obatined and the final result is the average of all the sub results.The more the number of trees and sub datasets are considered the more the accuracy of the algorithm.Due to this behaviour it is capable of managing huge amount of dataset [10].Following is the algorithm:

Step 1: Select K datapoints randomly from the train set

Step 2: Construct Decision trees of the subsets

Step 3: Select N which will be number of decisoon trees

Step 4: Repeat S1 and S2

Step5: Assign new datapoint a category according to the prediction of each tree.

F. Naive Bayes:

Naiye Bayes algorithm is formulated using the Bayes Theorem and is a popular supervised machine learning method.It is probabilistic in nature and so the working of this algorithm is based on the probability of an object.It is usually used for problems where text classification is needed but it can be used for other classification problems as well [11].

Following is the formula for Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where Posterior probability is denoted by P(A|B), Likelihood probability is denoted by P(B|A). P(A) is Prior Probability and P(B) is marginal probability.

5. Evaluation and results:

5.1 Heatmap:

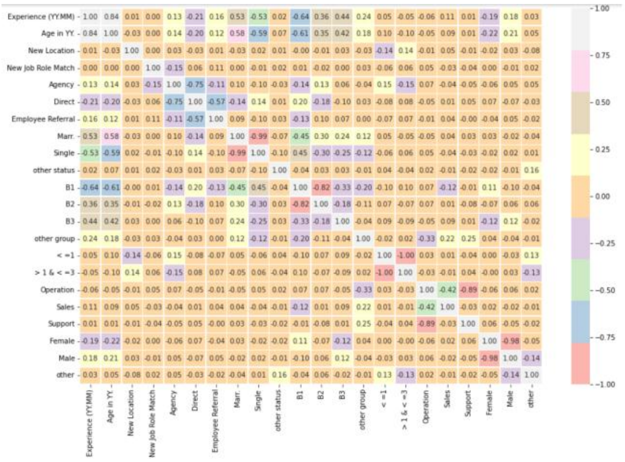

Figure 5.1.1: Heatmap

The above fig 5.1.1 is a heatmap that helps in identifying the strong and weak correlation between the attributes considered.

5.2 Graphs:

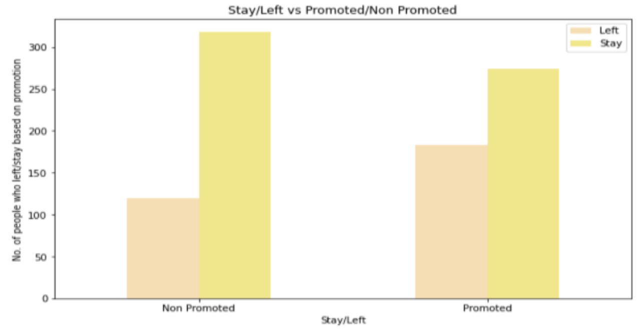

Figure 5.2.1: Promtion Vs Attrition Graph

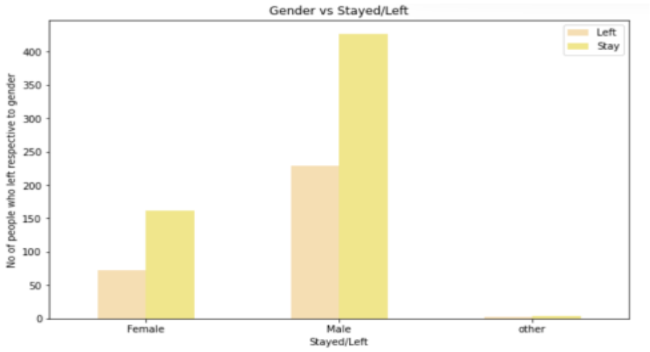

Figure 5.2.2: Gender Vs Attrition Graph

The above fig 5.2.1 and fig 5.2.2 represents the information about some of the attritubutes with respect to attrition in a graphical format.

The first graph is a relation between promtion and attrition. It is visible from the graph that an employee is more likely to stay in the organisation in the case where there has been a promotion.

The second graph is a representation of the effect on attrition based on gender.The graph shows how male candidates are more likely to stay in the organisation than female candidates.

5.3 Results:

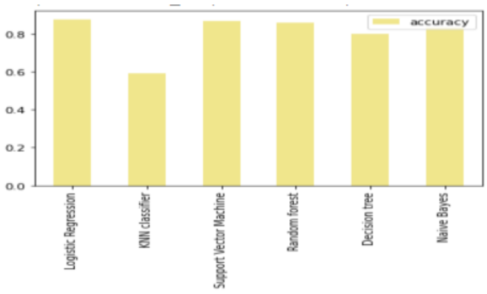

Figure 5.3.: Graphical result representation

The above fig 5.3.1 is a graphical representation of the result obtained when all the mentioned machine learning algorithms are applied on the dataset.

MODEL	ACCURACY
Logistic Regression	0.877095
KNN Classifier	0.592179
Support Vector Machines	0.865922
Naiye Bayes	0.832402
Decision Trees	0.804469
Random Forest	0.832402

Above table shows the test accuracy of each model seperately.

From the above table it is clear that Logistic Regression has performed best as it has the most amount of accuracy followed by Random forest.

The following table 5.3.2 and 5.3.3 gives an overview about the classification report of the two best models obtained from our dataset.

-	Precision	Recall	F1 score	Support
Stay	0.88	0.91	0.89	118
Leave	0.81	0.75	0.78	61

Table 5.3.2: Classification report of Random Forest Algorithm

-	Precision	Recall	F1 score	Support
Stay	0.91	0.90	0.91	118
Leave	0.81	0.84	0.82	61

Table 5.3.3: Classification report of Logistic Regression Algorithm

6. Conclusion

The moto of this paper is to determine which algoritm yields the best result for the chosen dataset to predict the attrition of employees.A total of six machine learning algorithms were applied on an opensource dataset and the output obtained was informed.It can be inferred from the output that logistic regression performs the best on the dataset followed by random forest algorithm.The attributes mentioned in the paper are some of the main causes of attrition and there can be many more parameters that can be added according to an organisations requirement.The aim of this paper is to compare some of the most widely used machine learning models so that it can help various kinds of organisations to maitain its workforce and lessen the rate of employee attrition.

REFERENCES

1. Journal of Interdisciplinary Cycle Research Volume XI, Issue XII, December/2019 ISSN NO: 0022-1945-A SURVEY PAPER ON EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING TECHNIQUES

2. VSRD International Journal of Business and Management Research , Vol VI Issue VII August 2016 - EMPLOYEE ATTRITION AND STRATEGIC RETENTION CHALLENGES IN INDIAN MANUFACTURING INDUSTRIES : A CASE STUDY

3. https://www.researchgate.net/publication/326029536_Employee_Attrition_Prediction-Employee Attrition Prediction

4. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 03, MARCH 2020 ISSN 2277-8616 3374 IJSTR©2020 www.ijstr.org EARLY PREDICTION OF EMPLOYEE ATTRITION

5. https://www.researchgate.net/publication/328772915_Employee_Turnover_Prediction_with_Machine_Learning_A_Reliable_Approach- Employee Turnover Prediction with Machine Learning

6. Employee Attrition Predictive Model Using Machine Learning - International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 05 | May 2020 e-ISSN: 2395-0056 p-ISSN: 2395-0072

7. F. Ozdemir, M. Coskun, C. Gezer and V.C Gungor, “Assessing Employee Attrition Using Classifications Algorithms,” In Proceedings of the 2020 the 4th International Conference on Information System and Data Mining, pp. 118-122, May 2020.

8. International Journal of Data Analysis Techniques and Strategies 3(3):281-299 July2011 – Logistic Regression in Data Analysis: An overview DOI 10.1504/IJDATS.2011.041335

9. JCSE International Journal of Computer Sciences and Engineering Vol.-6, Issue-10, Oct. 2018 E-ISSN: 2347-2693 - Study and Analysis of Decision Tree Based Classification Algorithms

10. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012 ISSN (Online): 1694-0814 - Random Forests and Decision Trees

11. International Journal of Advance Engineering and Research Volume 4, Issue 11, November -2017 - Short Survey on Naive Bayes Algorithm