International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 763-767 www.ijaem.net ISSN: 2395-5252

Prediction of Employee Attrition

Nikita Maurya[1], Rahul Kumawat[2], Ganesh Dash[3] and Prof. Sangeetha Selvan[4]

[1,2,3] *Student, Pillai College of Engineering, New Panvel*
[4] *Professor, Pillai College of Engineering ,New Panvel*

Submitted: 05-05-2021 Revised: 17-05-2021 Accepted: 20-05-2021

ABSTRACT-Employee is the most important aspect for any company or organization. Employee Attrition happens when an employee decides to leave the organization due to any reason, which will have impact on the particular organization's progress, as they lost their trained and experienced employee. Hiring and training the new employee will take time and money , which eventually slows the company growth. If we are able to predict which employee can leave in the future it will help the company to plan accordingly to avoid the upcoming loss. Random forest is used to extract feature importance and Artificial Neural Network (ANN) is used to predict the employee which will leave in the future.

KEYWORDS: Employee Attrition, Prediction, Neural Networks, Random Forest, Artificial Neural Network, ANN, Feature Importance, classifier

I. INTRODUCTION

Employee leaving a company because of any reason leads to attrition. Employees are valuable asset for any company or organization, losing employees will affect their progress. Prediction is when with the help of history data one can say what can happen in future. Predicting employee attrition means to predict which employee might quit from their job in the future.For an employee to leave, there are many factors to be considered personal, professional, etc. In our project model we have used random forest and Artificial Neural Network (ANN) to get our end result. The data set used is provided by kaggle website for IBM company employees.We have obtained feature importance considering all the attributes in the data set and it will help us to know which factor is more responsible for attrition. With the help of classifiers we can predict which employee can leave. By applying ANN layers we were able to get the yes or no value for prediction.The accuracy during training and testing were evaluated. Further a dashboard is prepared using flask which will make it easily accessible for the HR team to use the prepared model and access prediction values.

II. RELATED WORK

In [1] this paper the data set is obtained by kaggle website with 15000 observations and 10 attributes. They have explored the data and converted data with character values to numeric values. Data set to was applied to different conditions that helps in obtaining the prediction using gradient boosting, K means and random forest algorithms, and confusion matrix was used to view model performance. In [2] this paper many algorithms were used to obtain feature importance, and the performance of all the algorithms was evaluated with the help of AUC score. [3] Here, the dataset used was generated with the help of python script with 15,000 observations, they compared the performances of many algorithms on training data by using ROC-AUC where XGBoost gave better results than all others. [4] They have used data set of of 1470 observations and35 attributes and not considered all the attributes for prediction. Preprocessing was done by obtaining different graphs considering different relations between attributes. For prediction used different algorithms and random forest showed overall best accuracy. In [5] For preprocessing used heat map to find co relations between attributes and bar graphs with respect to attrition. Prediction model consist of KNN, SVM, Decision tree, Random forest where random forest performed better.

III. PROPOSED WORK

We proposed a prediction system for employee attrition by using Random forest and ANN. Random Forest is used to obtain the feature importance considering all the attributes from the data set. ANN gives the output as 0 or 1, 0 means no attrition and 1 means yes. The obtained classifier model is saved and applied to build a frontend. We build a system which is able to

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 763-767 www.ijaem.net ISSN: 2395-5252

analyse the important factors responsible for employee attrition and also displays the lists of employees which are about to leave in a dashboard.
StandardScaler is used for standardization where it finds mean values and replace it with unit variance. It is done while model trainingwhere first we find the mean and std to be used for feature scaling, then while fitting the model standardization is done by centering and scaling.**Random Forest**is an ensemble learning method also called as random decision forest as it uses forests of decision trees for classification, regression, etc. For prediction purpose the decision is made by voting that means every decision tree from the forest spits out a prediction value i.e 0 or 1, and then final vote is selected by calculating the votes the one with most votes is the final prediction for random forest.
Artificial Neural Networks (ANN)also called as Multi layer Perceptron (MLP) since it consist of many layers. It is an computer model algorithm that is supposed to produce output the way human brain analyzes or processes information. It is a self learning algorithm so while training it maps the input to output from the given example data and produces a function which will be used with new data. The basic structure consist of three layers .1. Input layer: to provide input that is a structured data.2. Hidden Layer: to perform mathematical calculations by the provided input. Amodel can have more than one hidden layer. 3. Output Layer: Gives the output prediction based on the given activation function.

Fig -1 Proposed system architecture

Data preprocessing here, the provided datset has an attribute as "Attrition" as yes and no which will help the model in learning from the real data. Next the Data cleaning is done by finding that there is no null, unidentified values. Now we use One hot encoding method to convert the categorical attributes into binary vector for e.g. JobRole is an attribute for which different options are Representative, executive, Scientist, etc so after applying One Hot Encoding the attributes become JobRole_Representative, JobRole_executive, JobRole _Scientist and so on and they hold value 0 or 1 accordingly. We Drop some unnecessary attributes link EmployeeCount, Over18, etc.
Data Analysishere, the dataset is now separated into two x and y on the basis of attribute as "Attrition" where x has data with Attrition = 0 and y has data for Attrition = 1. These two data set will be used for fitting the classifier to get feature importance and further split into test set and train set for fitting the model.

IV. IMPLEMENTATION

By applying Random forest we get to know the feature importance in attrition and the top 15 attributes are as follows :

1) Age -- 68.78 %
2) DailyRate -- 58.08 %
3) DistanceFromHome -- 50.39 %
4) Education -- 47.89 %
5) EnvironmentSatisfaction -- 46.38 %
6) HourlyRate -- 44.06 %
7) JobInvolvement -- 43.06 %
8) JobLevel -- 41.47 %
9) JobSatisfaction -- 34.56 %
10) MonthlyIncome -- 32.82 %
11) MonthlyRate -- 32.55 %
12) NumCompaniesWorked -- 31.94 %
13) PercentSalaryHike -- 30.13 %
14) PerformanceRating -- 29.65 %
15) RelationshipSatisfaction -- 28.56 %

Now we started by initializing ANN as sequential and get a linear stack of layers to which we will add layers using dense. Add three layer input, hidden and ouput with activation as Sigmoid as we want the output between 0 and 1. After compiling with train data once the model is fitted we will test it on test data and the accuracy on both is as follows

Table -1: ANN results of prediction

	Training	Testing
ANN Accuracy	0.9966	0.8197

The overall performance of our model is given below:
Note, 0 = will not leave and 1 = will leave

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 763-767 www.ijaem.net ISSN: 2395-5252

Table -2: Model Performance

	precision	recall	f1-score	support
0	0.87	0.92	0.89	245
1	0.44	0.33	0.38	49
Accuracy			0.82	294
macro avg	0.66	0.62	0.64	294

Now we would require a dashboard that can be directly used by HR teams so that they can take precautions accordingly. At First we will provide an input to submit a dataset. Once dataset is submitted a page will be loaded that contains the values .i.e list of employees who can leave in future and analysis graph of the feature importance.

Fig -2 Results of prediction

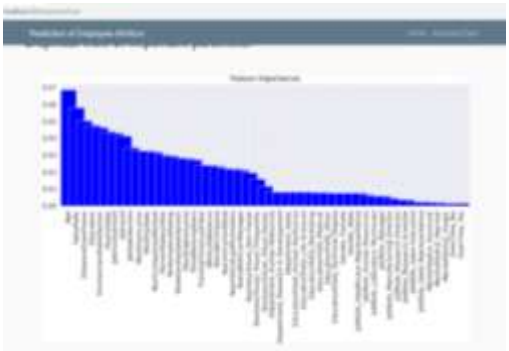

Fig -3 chart showing the important attributes.

V. REQUIREMENT ANALYSIS

The hardware and software requirements for implementation is given in this section.

Table -3: Hardware Requirements

Hardware	**Details**
RAM	512MB or above
Hard disk drive	500 MB or above (1GB or more recommended)
Processor	Pentium IV or above
Input Device	Standard Keyboard and Mouse

Software	**Details**
Operating System	Windows XP and above.
Frontend Application	Flask, HTML, CSS
Programming Language	Python 3.6 or above
Python Libraries	pandas, matplotlib, numpy seabom, sklearn, keras

Table -4: Software Requirements

The dataset used is taken from Kaggle Website provided by IBM analytics with 1470 observations and 35 features. The features consist of all kind of details of employees working life and personal life.

Data set Features are listed below :
Age
Attrition
BusinessTravel
DailyRate
Department
DistanceFromHome
Education
EducationField
EmployeeCount
EmployeeNumber
EnvironmentSatisfaction
Gender
HourlyRate
JobInvolvement
JobLevel
JobRole
JobSatisfaction
MaritalStatus
MonthlyIncome
MonthlyRate
NumCompaniesWorked

Over18
OverTime
PercentSalaryHike
PerformanceRating
RelationshipSatisfaction
StandardHours
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager

VI. CONCLUSION

In this paper we have analyzed ANN Classifier and used Random forest for feature extraction and prepared a model that can predict employee attrition. The accuracy of ANN on training set is 0.99 while the accuracy on test set is 0.82. According to random forest Age, DailyRate, Distance from Home, Education, Environment Satisfaction, Hourly Rate are the major factors responsible for attrition. For future purpose we can more hidden layers to our ANN to improve its performance.

ACKNOWLEDGEMENT

We owe our great regards to our supervisor Prof. Sangeetha Selvan for her guidance and patience throughout the process. We are very thankful to our Head of the Department Dr. Satishkumar Varma and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to presenting this work.

REFERENCES

[1]. "Employee Churn Rate Prediction and Performance Using Machine Learning" by Aniket Tambde, Dilip Motwani International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8,Issue-2S11, Sept 2019

[2]. "Employee Attrition Prediction using Data Mining Techniques" byJeel Sukhadiya, Harshal Kapadia, Prof. Mitchell D'silva-International Journal of Management, Technology And Engineering Volume 8, Issue X, OCTOBER/2018 ISSN NO : 2249-7455

[3]. "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms" A case for Extreme Gradient Boosting Rohit Punnoose,Pankaj Ajit-International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 9, 2016

[4]. "Foreseeing Employee Attrition Using Diverse Data Mining" by Jalpesh Vasa and Kanksha Masrani-International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019

[5]. "Employee Attrition Predictive Model Using Machine Learning" by Adarsh Patel, Nidhi Pardeshi, Shreya Patil, Sayali Sutar, Rajashri Sadafule, Suhasini Bhat -International Research Journal of Engineering and Technology e-ISSN: 2395-0056 Volume: 07 Issue: 05 | May 2020 p-ISSN: 2395-0072

[6]. "Data science vs Big Data vs Data analytics" [Online] Availablehttps://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article

[7]. "MachineLearning" [Online] Availablehttps://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

[8]. "Random Forest Algorithm" [Online] Availablehttps://www.javatpoint.com/machine-learning-random-forest-algorithm

[9]. "Artificial Neural Network [ANN]" [Online] Availablehttps://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_neural_networks.htm

[10]. S. Saranya, J. Sharmila Devi, "Predicting Employee Attrition Using Machine Learning Algorithms and Analyzing Reasons for Attrition," International Journal of Advanced Engineering Research and Technology (IJAERT), Volume 6, Issue 9, September 2018, pp. 475-478.

[11]. Kaggle, "HR-Employee-Attrition." [Online]Availablehttps://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

[12]. Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari, "Evaluation of machine learning models for employee churn prediction," International Conference on Inventive Computing and Informatics (ICICI 2017).

[13]. M.Sudheer Kumar, Obulesu Varikunta, K.Ramakrishna, "Employee Attrition and +Retention Strategies in Manufacturing: An Empirical Study in Amara Raja Batteries Limited," International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8, Issue-7, May, 2019, pp. 2962-2968.

[14]. "Confusion Matrix" [Online] Availablehttps://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
[15]. 'Neural Network' [Online] Availablehttps://data-flair.training/blogs/artificial-neural-networks-for-machine-learning/