Employee Future Prediction

Saraswati Devi
https://github.com/Saraswatinehra/UCDPA_Saraswati_Devi.git
Gurgaon, India

Abstract: - The biggest asset that a company has is not his business or the place or any other thing, but it's the employee. The employees pave the path for the success of the company and all the administration that is needed to run the company. But due to many reasons employee are not happy with the work or the company they work in and this tends or results for them to leave the company or finding a new company. It is one of the most important difficulties that company owners confront when their organizations lose their most brilliant personnel. A good employee is always a valuable asset to the company, and their departure can result in a variety of issues, including financial losses, poor overall performance, and the loss of accumulated expertise. Furthermore, compared to recruiting current employees, employing new personnel is significantly more expensive, stressful, and time-consuming. Hence, the authors propose a system to predict the future of an employee in a company taking in consideration various factors and using the algorithm to get the prediction with a great accuracy.

Keywords: - *Prediction Algorithm, Machine Learning, Employee Future Prediction, Unsupervised Learning*

In many of the company's employee, future has become a major. It's the responsibility of the HR (Human Resource) Manager to fulfil necessary needs of their employs. Employs generally leave the companies and to fulfil that particular position they have to assign another person for that particular position, which is not always possible for them. Since the worker's quality and skills represents the companies the growth and competitive advantage. [1] And there has been a method developed called Predicting Employee Attrition. The purpose of this is to support decisions are not based on personal aspects but on unbiased data analysis. The goal of this work is to analyse how purpose factors influence employee attrition, in order to identify the man causes that contribute to a worker's decision to leave a company, and to predict whether a particular employee will leave the company.

I. LITERATURE OVERVIEW

There few steps that are being used for performing ML processing, and they are Data Preprocessing, Data Analysis, Model Training, Model Validation, Model Prediction, Visualization of results [2] The consequences received from the data analysis constitute a starting point in the improvement of increasingly efficient worker attrition classifiers. the use of greater numerous datasets or certainly to replace it periodically, the software of feature [3] engineering to identify new sizable traits from the dataset and the supply of additional statistics on personnel would improve the overall information of the motives why personnel leave their agencies and, consequently, increase the time available to personnel departments to assess and plan the duties required to mitigate this risk (e.g., retention activities, employee substitution and/or venture redistribution).[4]. The report uses a combination of unsupervised learning algorithm with two data sets (Employee and Country data)

II. METHODOLOGY

A. Data Preparation

The information was taken from the well-known employee future prediction dataset on the internet. Due to the scarcity of data, the only method to run a model and get a forecast was to gather data from a reliable source. The dataset contains a variety of properties like education, gender, age, city, payment tier, joining year, leave or not, ever benched etc which can be seen in figure 1. The dataset is quite big and contains ~4653 entries which is taken from the employees in a company and 9 features along with it.

```
Education
JoiningYear
City
PaymentTier
Age
Gender
EverBenched
ExperienceInCurrentDomain
LeaveOrNot
```

Fig 1: Features/Attributes Overview

B. Data Preprocessing

Preprocessing is a great approach to improve data because both the actual world and our data have flaws. The speed of the procedure is determined on whether the data has been preprocessed. The better the model that is utilized, the more preprocessing that is done. Before eliminating the ID column, we check for all null values, which will have no effect on the results. After removing the null values, we performed a describe function to get to know the dataset more, which would also be beneficial in doing the visualization as it would give the right relation to perform. The figure 2 shows the describe table of the dataset.

	JoiningYear	PaymentTier	Age	ExperienceInCurrentDomain	LeaveOrNot
count	4653.000000	4653.000000	4653.000000	4653.000000	4653.000000
mean	2015.062970	2.698259	29.393295	2.905652	0.343864
std	1.863377	0.561435	4.826087	1.558240	0.475047
min	2012.000000	1.000000	22.000000	0.000000	0.000000
25%	2013.000000	3.000000	26.000000	2.000000	0.000000
50%	2015.000000	3.000000	28.000000	3.000000	0.000000
75%	2017.000000	3.000000	32.000000	4.000000	1.000000
max	2018.000000	3.000000	41.000000	7.000000	1.000000

Fig 2: Describing the dataset

C. Feature Visualization

The algorithm's features are crucial for producing reliable results. Visualization helps us to see all the different factors and how they influence the results. The figure 3 shows a heat map between the features/attributes and which signifies Our Target variable ('Leave or Not') has a very modest negative connection with Payment Tier and Age. The target variable ('Leave or Not') exhibits a very minor positive connection with 'Joining Year.' Figure 4 shows that Most of the Employees are having payment Tier 3 By observing the trend it seems that 'Payment Tier' Category is an Ordinal Variable Where, Tier 3 > Tier 2 > Tier 1. Figure 5 shows that most of the employees have a bachelor's degree but the employees with master's degree have left more. And in figure 6 it shows that young people have left early due to more job opportunities out in the market.

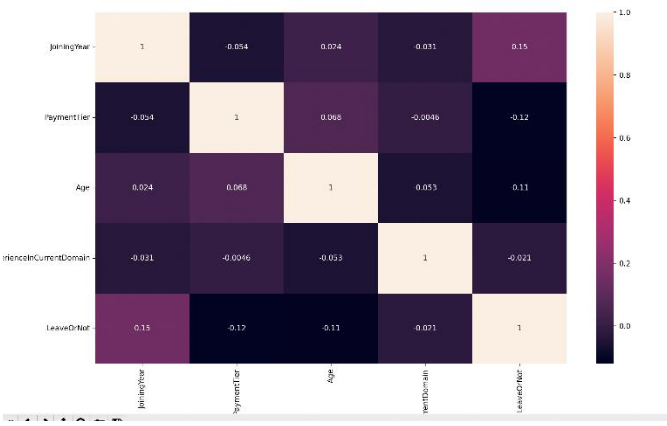

Fig 3: Heat map of Attributes

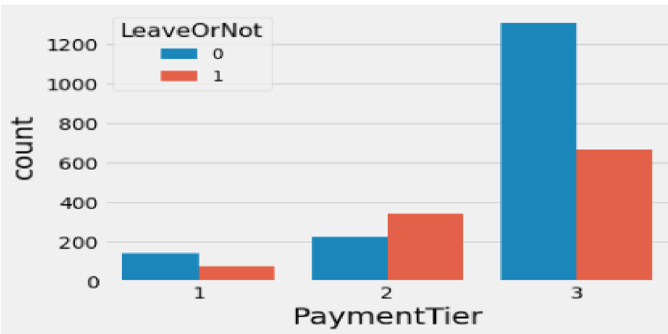

Fig 4: Payment Tier Visualization

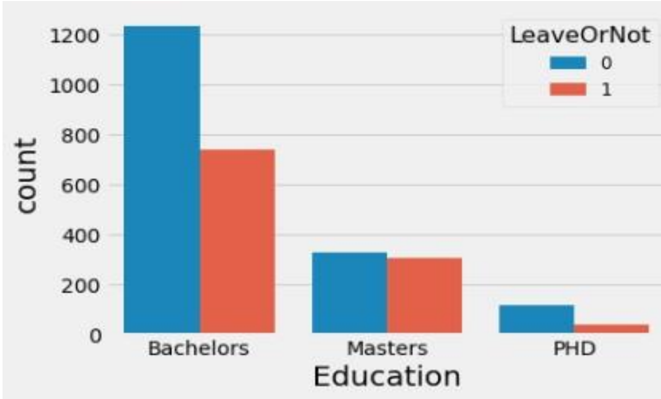

Fig 5: Education Visualization

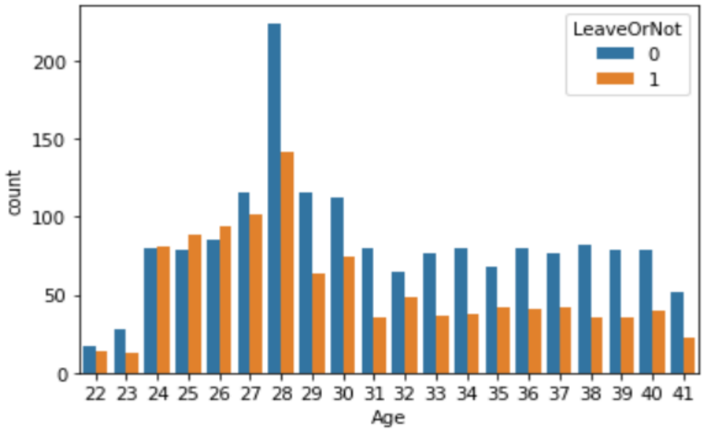

Fig 6: Age Group Visualization

D. Model Architecture

Gradient boosting is similar to random forest in that it groups weak learners together to build a stronger one. However, gradient boosting differs from random forest in that it adds predictors in a sequential manner, preventing additional improvement. The gradient booster uses the gradient of descent to help discover and fix inaccuracies in learners' predictions. In compared to decision trees and linear regression, it is memory efficient and fast, but it fails to visualize perfectly. Gradient boosting framework is not relevant to my model hence not used.

III. EXPERIMENTAL RESULTS

The steps followed from data preparation to preprocessing has helped the dataset to get better for the model and has played a vital role in the accuracy, and the algorithm used is unsupervised learning on country data where We examined the Life Expectancy (WHO) data set with the basic models in Machine Learning and got the results.

IV. CONCLUSION

Employees are the main asset to the company and the company can't run without them and they play an important role in shaping the company and sending the company to next heights i.e making it successful hence the data analysis used in the project help the HR managers to know what should be improved if an employee is thinking of quitting or finding a new job and by this, they would not loose a credible employee to any other rival company, So the results given by the analysis would make an impact on the present systems and increase the predictions by a very high percentage,

REFERENCES

[1]. “Employee-future prediction” Using Kaggle
[2].”employee-attrition-prediction”Using Medium.com
[3] UC Libraries (universityofcalifornia.edu)