HR Analytics

Problem Definition: Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees. But where HR Analytics fit in this? and is it just about improving the performance of employees?

HR Analytics

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.

Attrition in HR

Attrition in human resources refers to the gradual loss of employee's overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

How does Attrition affect companies? and how does HR Analytics help in analyzing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.

Attrition affecting Companies

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

Data Analysis:

Initially we need to import the hr. analytics csv file into the Jupiter notebook, we found than there are number of columns with categorical values and numerical values, then we are further checking the shape of the data frame present it, and it is found that the shape is (1470,35) and also, we check the data type present in the data frame

We found that the data type has integer and object and further we check the null values in the data frame by using is null command and we found that there are no null values present in it, then there is one more command to check null values it shows like true or false, if the null values is present it shows like true and if there are no null vales are present then it shows like false

Further we describe the data frame with count, mean, std, min, 25%, 50%, 75% and max. Finally, we need value count for the attrition with yes and no, we see that no has 1233 and yes has 237 counts in it

For analysis part we just see the comparison with yes and no divide by yes, we get the 80% of attrition guess

EDA Concluding Remarks:

Next, we came to the Exploratory data analysis part, in that we visualize the attrition with the count plot, it shows the graphical figure representation and then we use count plot with comparison on age and attrition

Pre-Processing Pipeline:

Then we see the value count of all the data frames and data types of the data frames, it shows the value count for all the columns in the data frame, then we need to remove unwanted columns that we want to drop and then we see the correlation of the data frame then we further se the correlation in the heatmap visualization, the color with high contrast is highly correlated and the color with low contrast is lowly correlated

We use label encoding technique to convert categorical variables into numerical values to fit the transform the columns and dropping the extra and unusable columns so than we can get accuracy in model prediction

Building Machine Learning Models:

Before building the model, we need to split the dependent variable and non-dependent variable then we need to call train test split function with test size 0.5 and random state is 0, further we are using random forest classifier and it is one of the best algorithms for this data frame

For the random forest it shows the accuracy of 97% and then we use confusion matrix algorithm the values for true positive are 309 and true negative are 1 and false positive are 49 and false negative is 9

Concluding Remarks:

Finally, by the above calculation with the confusion matrix we get the model testing accuracy is 86%, so by this random forest is the best algorithm that we Have chosen and final step is to save the model with the best algorithm that we have chosen.