HR Analytics Project (Understanding the Attrition in HR)

This project is about HR analysis on the attrition of employees from the organisation.

The application of analytical techniques to an organization's human resource department with the goal of enhancing employee performance and, consequently, obtaining a higher return on investment is known as human resource analytics, or HR analytics. HR analytics include more than just collecting data on worker productivity. Instead, it collects data and uses it to make pertinent judgments on how to enhance each process in order to give insight into it.

Absolutely, attrition can indeed pose significant challenges for companies. High attrition rates can lead to increased recruitment and training costs, loss of institutional knowledge, decreased morale among remaining employees, and disruptions in productivity and workflow. Human resources (HR) professionals play a crucial role in addressing and mitigating attrition.

Data Pre-processing:

In this database there were 35 columns and 1470 rows present. There is no null value present in this dataset.

There are some columns which I thing would not contributing for predicting the target variable. Such as 'EmployeeCount','EmployeeCount','StandardHours','Over18'. So, I have remove these columns from the dataset. After removing this columns I got 31 columns remain with me.

In the description of the dataset I found no major difference between 50% value with mean value. It show there is no outliers present in this dataset.

There are some categorical columns present in this dataset which I had changed it to continues data. For that I used Lable Encoder technique to change the data format. I checked the outliers if present in this data by ploting Distplot and I found some outlier present in some columns i.e. YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion & YearsWithCurrManager. I again confirmed this by ploting boxplot graph. I used Z score outlier removable technique to remove outliers. After removing outliers, Now I have to check the Multicollinearity issue in the columns. For that I used Variance inflation factor. I found two columns i.e. JobLevel and MonthlyIncome are high VIF value so, I decided to remove these columns. After removing unwanted columns, now I have 29 columns. I used these columns for model building.

Handling Imbalanced Data:

In Attrition column, data was imbalanced. Out of 1470 employees total 229 employees were leaved the company. So, I used SMOTE technique to balance the data in this column.

Model Building:

After doing all these PDA activities, I started building classification model. I imported all the necessary libraries. After checking the Accuracy score of all the models. I found every models gave good accuracy score. So, I decided to check Cross Validation score of the models. After checking the cross validation score of all the models, ExtraTreesClassifier had given me highest CV score i.e. 90%. After that I decided to tune the parameters to check the more accuracy score by using GridSearchCV technique.

Final Results:

After tuning the parameters I got highest final score i.e. 93%.

In this way I handle the dataset and build the model. 93% accuracy score is best score.