HR ANALYTICS AND EMPLOYEE TURNOVER PREDICTION

DATA SYNOPSIS

The dataset used in the project contains the data of employees of a hypothetical company.It is taken from kaggle.The link to dataset is https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset The dataset contains 1470 observations and 35 features.Out of the 35 features,9 features are categorical and 26 features are numerical.The target variable (y) which is a categorical variable mapped to 0 and 1 is attrition which indicates whether an employee has left from the company or not.

GOALS

- Exploratory Data Analysis is performed on the dataset and visualisations are plotted so as to get information which will aid the HR Department in the optimization of the workforce.

- Building a Machine Learning model which will predict whether an employee will leave or not from the company(y variable).

- Finding the first 10 employees among various categories who are more probable to leave the company

SPECIFICATIONS

1.AGE

Numerical Value-Age of the employees

2.ATTRITION

Categorical Value-Employee leaving the company

3.BUSINESS TRAVEL

Categorical Value-tells whether the employee has to travel for business purposes

4.DAILY RATE

Numerical Value - daily rate of salary

5.DEPARTMENT

Categorical Value-tells the department in which the employee is working

6.DISTANCE FROM HOME

Numerical Value - the distance between workplace and residence of employees

7.EDUCATION

Numerical Value-The level of education

8.EDUCATION FIELD

Categorical Value-tells the educational background of the employees

9.EMPLOYEE COUNT

Numerical Value

10.EMPLOYEE NUMBER

Numerical Value - employee id number

11.ENVIRONMENT SATISFACTION

Numerical Value - tells the level of satisfaction of work environment of the employee

12.GENDER

Categorical Value-tells the gender of the employee

13.HOURLY RATE

Numerical Value -tells the hourly pay of employees

14.JOB INVOLVEMENT

Numerical Value - tells the level of job involvement of the employee

15.JOB LEVEL

Numerical Value - tells the level of job of the employee

16.JOB ROLE

Categorical Value-tells the various designations of the employee

17.JOB SATISFACTION

Numerical Value - tells the level of job satisfaction of the employee

18.MARITAL STATUS

Categorical Value-tells the marital status of the employee

19.MONTHLY INCOME

Numerical Value - tells the monthly take home salary of the employee

20.MONTHLY RATE

Numerical Value - tells the monthly gross pay of the employee

21.NUMCOMPANIES WORKED

Numerical Value - tells the number of companies at which the employees have worked earlier

22.OVER 18

Categorical Value-tells whether employee is over 18 years of age

23.OVERTIME

Categorical value-tells whether the employee works for overtime

24.PERCENT SALARY HIKE

Numerical Value - percentage of salary increase for the employees

25.PERFORMANCE RATING

Numerical Value - tells the performance rating of the employee

26.RELATIONSHIP SATISFACTION

Numerical Value - tells the level of relationship satisfaction with the company

27.STANDARD HOURS

Numerical Value - tells the standard working hours in a month

28.STOCK OPTIONS LEVEL

Numerical Value - tells the level of stock options offered to the employee

29.TOTAL WORKING YEARS

Numerical Value - tells the total work experience of the employee

30.TRAINING TIMES LAST YEAR

Numerical Value - tells the number of trainings provided to the employees

31,WORK LIFE BALANCE

Numerical Value - tells the level of work life balance of the employee

32.YEARS AT COMPANY

Numerical Value - tells the total number of years the employee has worked in the company

33.YEARS IN CURRENT ROLE

Numerical Value - tells the total number of years the employee has worked in the current role

34.YEARS SINCE LAST PROMOTION

Numerical Value - tells the number of years since the last promotion of the employee.

35.YEARS WITH CURRENT MANAGER

Numerical Value - tells the number of years the employee has worked with the current manager

OVERVIEW OF THE MAIN TOPIC

Hiring and retaining top talent is an extremely challenging task that requires capital, time and skills. Small business owners spend 40% of their working hours on tasks that do not generate any income such as the hiring process for new employees.

Employee Turnover or Employee Turnover ratio is the measurement of the total number of employees who leave an organization in a particular year. Employee Turnover Prediction means to predict whether an employee is going to leave the organization in the coming period.

A Company uses this predictive analysis to measure how many employees they will need if the potential employees will leave their organization. A company also uses this predictive analysis to make the workplace better for employees by understanding the core reasons for the high turnover ratio.

APPROACH

- Data Preprocessing and Exploratory Data Analysis
- Visualisations using various plots with the aid of packages like matplotlib.pyplot and seaborn
- Creating different data frames so as to provide the information of employees of different categories(left,staying,hard working,best performers,hybrid of both best and hard working employees)
- Finding various insights from the visualisations and statistical descriptions.
- Removing the unnecessary columns from the dataset so as to fit into the Machine Learning Model
- Splitting the dataset into train and test data.
- Using SMOTE as the dataset is having an imbalance set of target variables(y) and it is mainly used with ensemble models and xgboost models
- Building various Machine Learning(ML) Models using various algorithms
- Fitting dataset into various algorithms
- Predicting the values using the ML model.
- Finding the accuracy of various models
- Finding the confusion matrix and classification report for each ML model.
- Selecting the ML model with the highest accuracy.
- Finding the importances of features
- Finding out the employee in various categories who are more probable to leave the company.

MACHINE LEARNING ALGORITHMS USED

- Logistic Regression
- Ridge Classifier
- XGBoost Classifier
- RandomForest Classifier
- Adaboost Classifier
- Gradient Boosting Classifier

GridSearchCV was used for Ridge Classifier and Random Forest Classifier for hyper parameter tuning

XGBoost Classifier was selected for building the final ML model as it was having an accuracy score of 0.94(94% accuracy).

RESULTS

The following results were found from the statistical description and visualisations:

- Majority of employees who are working in the company are aged 37 and the minimum and maximum age of employees are 18 and 60 respectively.
- The average daily rate of the employees is 802.5 dollars with minimum being 102 dollars and maximum being 1499 dollars
- Most of the employees are living within the range of 9.2miles from the company
- Most of the employees are having decent education level of 3
- Majority of the employees are earning 66 dollars per hour with the employees earning 100 dollars as highest and 30 dollars being the lowest
- Environment Satisfaction,Job Involvement,Job Satisfaction,Relationship Satisfaction and work life balance are having the similar mean value measure which infer that majority of the employees are happy to work and are more productive
- Coming to the job level,majority of the employees are having level of 2 with the highest being 5 and lowest being 1.It would be better if the average job level was raised to 3
- Most of the employees are having an average take home of 6503 dollars with the highest being 19999 dollars and lowest being 1009 dollars.

- Majority of the employees have working in about 3 other companies before joining the company which tells that the workforce is having a lot of experienced employees
- All the employees are over the age of 18
- The average salary hike for employees is 15 percent with the lowest being 11 percent and highest being 25 percent which is also a good indication and company should retain the present levels
- The performance levels of employees are very good and the best performers are having a score of 4
- The Standard working hours of the employees is 80 hours per month
- Only a few of the employees are getting company stocks and this should be a case which is to be looked upon by the management.Better stock option level should be offered to more number of employees
- Majority of the employees have working for more than 11 years which further supports the statement that the workforce is rich with experienced employees
- The company used to provide 3 trainings in an year for majority of the employees and there are certain employees who are given 6 training in an year and there are employees who haven't got any trainings.The company should at least provide 1 training to all the employees as more trained employees are necessary for increasing the productivity
- Most of the employees are working in the company for over 7 years with the highest being there working for 40 years and majority of employees wish to work for over 4 years in their present role.
- Majority of employees got promoted 2 years back
- Most of the employees wish to work with the same manager for over 4 years and there are employees who are working for over 17 years with the current manager

From the correlation heatmap,the following results were found:

- Age is having high correlation with Job Level,monthly income and total working years
- Job level is having high correlation with monthly income,total working years and years at company
- Monthly income is having high correlation with total working years and years at company
- Percent salary hike is having high correlation with performance rating

- Total working years is highly correlated with years at company,years in currentrole,years since last promotion and years with current manager
- Years at company is having a high correlation with years in current role,years since last promotion and years with current manager

From the count plots and kernel density estimate area plots ,the following insights were obtained:

- Younger employees of age in the range 18-21 tend to leave from the company.Employees of age 26,28,29,31,32,33,35 also tend to leave the company
- Comparing the count plots,more employees working in the Sales department tend to leave from the company than the other two departments
- Employees working as Sales Representative are more probable to leave.Sales executives,Research Scientist and Laboratory Technicians are also probable to leave
- There are more number of attritions of employees in the educational background of Marketing and Technical Degree
- Graphs indicate that single employees tend to leave more compared to married and divorced
- Employees having work life balance level 1 are more probable to leave
- The employees who are least involved in the job tend to leave
- It is clear from the chart that when the number of trainings given to employees increases,the attrition rate of employees decreases
- Employees of job level 1 are more probable to leave
- When the stock option level is higher,less employees tend to leave
- The blue area plot supersedes the red one at the beginning which indicates that the employees who are closer to the company are less likely to leave and as the red area plot supersedes the blue one,which represents the employees who are far away from the company are more probable to leave
- Employees who are working under the current manager for more years(more than 5) are less likely to leave and employees who are working for less than 3 years with the current manager are more likely to leave
- Experienced employees are less likely to leave when compared to employees having less years of experience(less than 5 years)

- Employees who are working more than 5 years in the current role are less likely to leave and the employees who are working less than 3 years in the current job role tend to leave
- Both male and female employees are getting decent remuneration and female employees even though they are less in number are paid more compared to male.
- Managers and Research Directors are highly paid when compared to employees in other designations.Sales Representatives are the ones who are paid the least when compared to other employees.Research Scientists and Laboratory Technicians are almost getting equally paid

From the three categories of employees(best performers,hard working employees and hybrid of best performing and hard working employees),the first 10 employees who are more probable to leave the company are found.

Dataframes are created with the employees of the three categories in the decreasing order of their probability to leave the company.