A picture containing text, sign, close

Description automatically generated

**IS HR ANALYTICS REALLY IMPORTANT?**

To answer the above question, we 1st need to understand what is HR Analytics?

**HR analytics** is defined as the company’s process of measuring the impact of **HR** metrics, such as time to hire and retention rate, on business performance.

**HR analytics** is the science of gathering, organizing and analysing the data related to **HR** functions like recruitment, talent management, employee engagement, performance and retention to ensure better decision making in all these areas.

**BENEFITS OF HR ANALYTICS**

* Improves HR performance.
* Identify best performance talent.
* Identify Attrition and its cause.
* Predict in demand skills and positions within organisation.
* Transforms the role of HR as a strategic partner.

**WHAT IS EMPLOYEE ATTRITION?**

Employee Attrition refers to the reduction in the number of employees working in a company, business organisation. The reason of leaving the company can be professional growth, better opportunities, health issues, retirement, etc. There are multiple factors which affect the attrition rate in a company. HR analytics plays a major role in predicting employee’s attrition, it helps to predict whether the employee stays or leave the company in the long run?

**IMPORTANCE OF HR ANALYTICS IN EMPLOYEE ATTRITION?**

For any company hiring employees is not an easy task, there is a huge cost involved for hiring employees to a company, business organisation. Company not only spends for hiring an employee, but it also arranges different training sessions for the existing ones by which the employees can gain knowledge, increase efficiency, etc. This is an investment that company do on their employees to get better return on investment. So, the aim of every company is to hire those who will not leave the company in early stage of joining. For this HR Analytics plays an important role at the time of hiring process. There are many factors by which attrition rate in a company can be predicted and considering those factors hiring of employees can be done. This helps in reducing the cost to a company and retaining the best talent.

A picture containing text, city, sign

Description automatically generated

**NOW LET’S TAKE YOU TO THE JOURNEY OF DATA ANALYSIS OF THE SAME.**

1. **PROBLEM STATEMENT**

So, I will be working on a data set provided by [IBM](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset) , thanks to them for making this data set publicly available (this is a fictional data set just to show how things work in predicting attrition). This is a data of former employees working in a company.

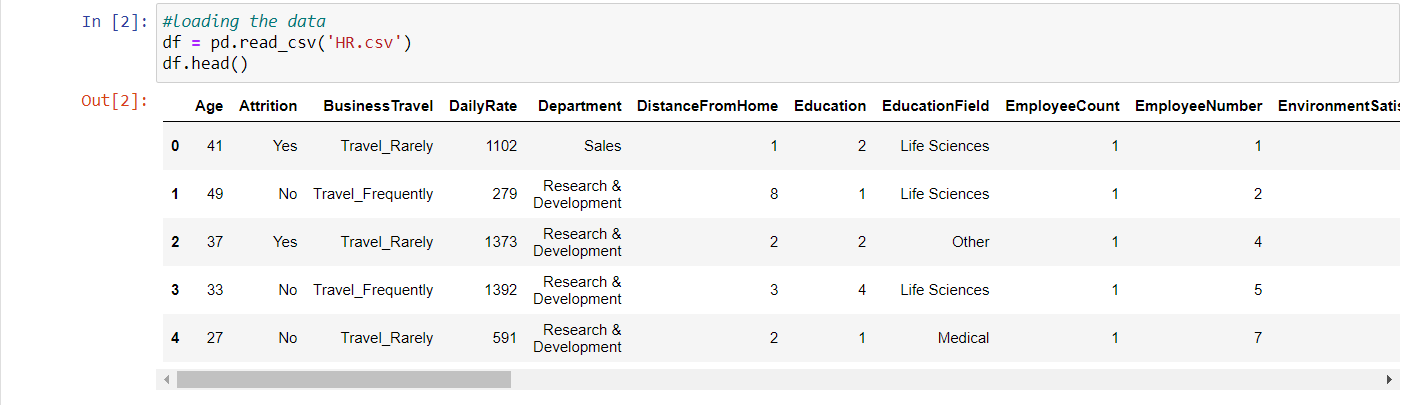
In this project work we will try to predict our Target variable which is Attrition (Probability of an employee leaving the company or not?)

I will also try to show what factors affect the attrition rate and at what stage an employee is more likely to leave a company.

This is a Binary classification problem where we will be classifying Attrition(‘YES’) as 1 and Attrition(‘NO’) as 0.

1. **DATA ANALYSIS**

Here I will be loading my data set, in our data set we have 1470 observations and 35 columns (26 are int format and 9 are object format). And luckily there are no missing values in our data set.

Table

Description automatically generated

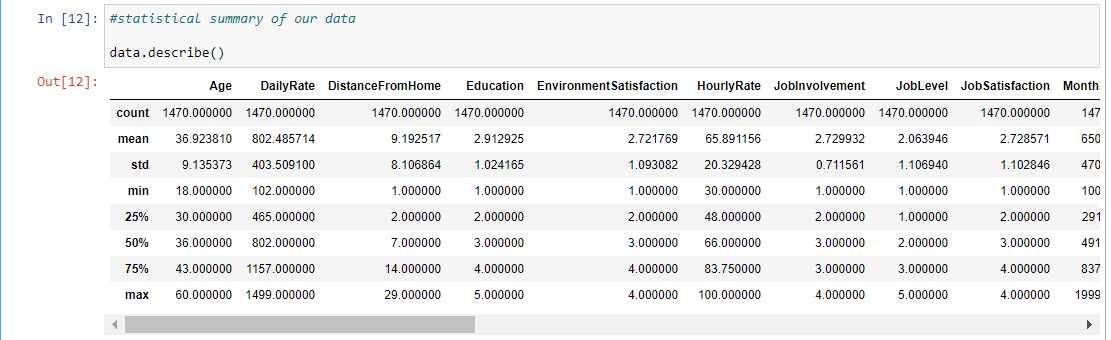
Above you can see that there are many details of an employee available to us. These details will help us in predicting the attrition.  
We will be doing analysis on all of the above features visually as well.

After checking the features, I see that there are some irrelevant features which should be removed, Here I will be removing 2 columns Employee Count and Employee Number, as they are not affecting the change in the number of Attrition, they are just used to refer an employee uniquely and also removing Standard hours, Over18, because there is no variance in these columns, 80 hrs are fixed value for all the employees, and all employees are Over18 so it will have null effect on our model.

Graphical user interface, text, application

Description automatically generated

Now we are left with 31 columns including our target variable.  
From statistical description of our data set, I found out that Age ranges from 18 to 60, there is a high difference in monthly rate and monthly income from 75 percentile to maximum value. Maximum years with current manager are 17 years and Maximum years in a company is 40 years.



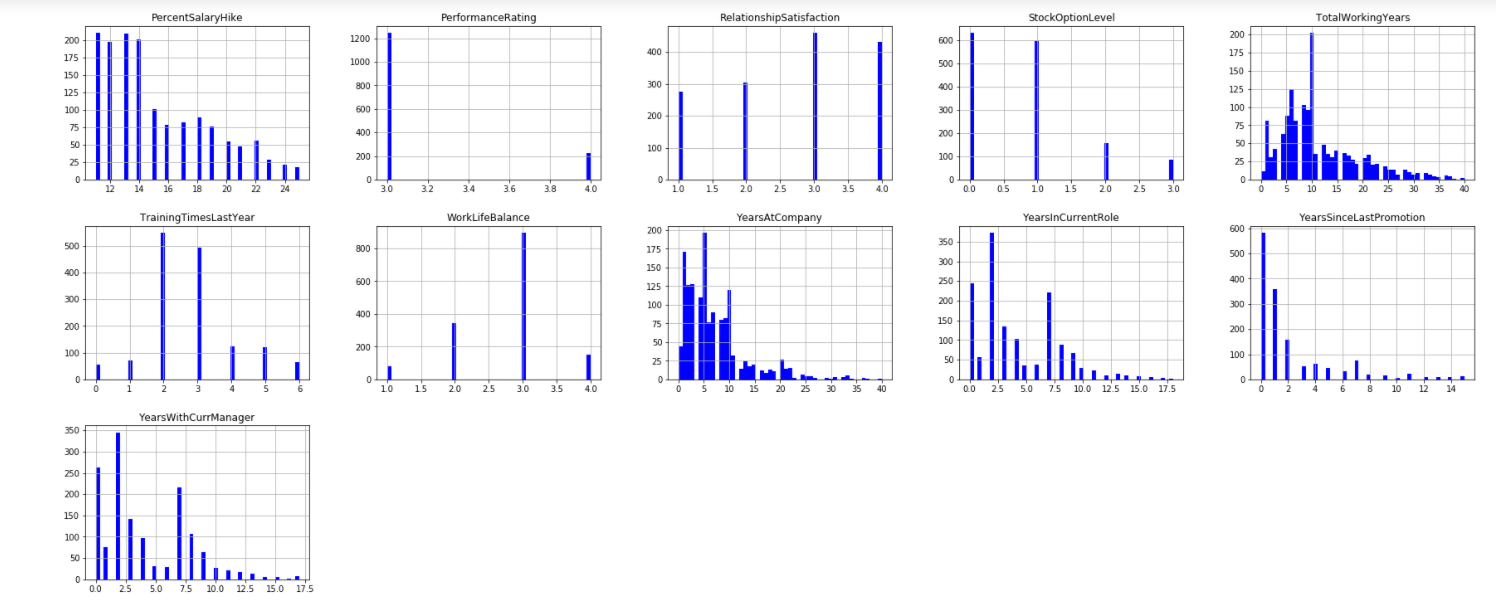
Now as we also have categorical variables, I will be using Label encoder for that to convert them into numeric format. And after converting the Object type columns lets look at the distribution of the variables. Here we can see that there are many numeric features which are right skewed those are (Years with current manager, Years at a company, Total working years, Years in current role, Years since last promotion, Monthly income, Distance from home). There are some columns which are approx. normally distributed those are (Age, Daily Rates, Monthly Rate, and Hourly Rate).

Also, if we look at our target Variable which is Attrition, we can see that our target variable is imbalanced. The data states that there are more than 1200 employees still working in the company and about 250 employees left the company. This shows abnormal distributed data.

Graphical user interface, text, application, email

Description automatically generated





1. **EDA**

* **1st lets check the relation between employee age and attrition.**

Chart, bar chart, histogram

Description automatically generated

Here we can see that the attrition is maximum between the age groups 26-33. At a very younger age, i.e. from 18-20, the chances of an employee leaving the organization is more- since they are exploring at that point of time. It reaches a breakeven point at the age of 21. At old age above 50 the attrition is there, and the major reason could be retirement or health issues.

* **Relation between Gender and Attrition**

Chart

Description automatically generated

the Attrition count is more in case of Male; however, this is not justified as we had a data with more counts of Males.

* **Relation between Distance from home and Attrition.**

A picture containing text, writing implement, screenshot

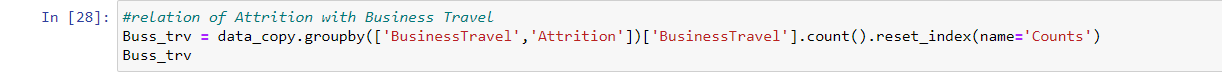
Description automatically generated

As the Distance increases the Attrition rate also increases, breakeven point at 24, when distance is 24 km the attrition is the highest. There is a positive Linear relation between attrition and distance from home.

* **Relation between Business travel and Attrition.**

Chart, line chart

Description automatically generated

Here, we see the relation between attrition and employee’s business travel for work. I have created a Data Frame (Buss\_trv) which is the main data frame grouped by Business Travel and Attrition.  


Here we can see that the rate of Attrition is high when there is more travelling. At 1 the attrition rate is highest 25% approx.

* **Relation between Monthly income and Attrition.**

Graphical user interface, application

Description automatically generated  
Here we can see a good relation of Monthly Income and Attrition counts. We can see when the income level was around 2500 (from starting) there are more Attrition, this could because the employees are looking for more better opportunities in the starting. And then again at 10000 there is a spike, that might be because the economic expectation of employee rises, and he desire for a better Pay and shifts to another company. When an employee gets good pay, he tends to remain in the company as shown in the graph.

* **Relation between Education and Attrition.**

**Chart

Description automatically generated**From the above graph we can see that at education level 1 the rate of attrition is highest, which is 18.23% (That I have calculated separately by grouping Education and Attrition, and at education level 5 the rate of Attrition is minimum at 10.41%)

* **Relation between Environment satisfaction and Attrition**

**  
Chart

Description automatically generated**From the above graph it can be seen that at the initial phase the employment there is more rate of Attrition, at 1 level the attrition rate is 25.35% highest from all other levels. A new employee takes time to keep up with the office Environment, from 2 to 3 we see that there is an increase at attrition rate, the reason could be if an employee is shifted to a new process or workplace.

* **Relation between Marital status and Attrition.**

**Chart, box and whisker chart

Description automatically generated**Attrition is higher in case of single employees and lowest in case of Divorced at 10%. Single employees show the largest proportion of leavers at 25%.

* **Subplots of remaining categorical variables.**

Diagram

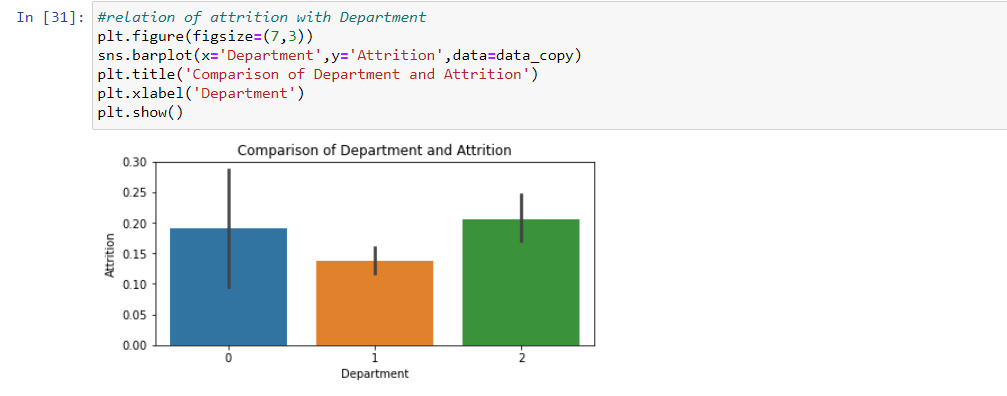
Description automatically generated  
I have 1st created a separate grouped by data frame for all the remaining category variables.  
After that I used subplots from matplotlib to plot all the above line plots.

From above plots we can see that Job Level, Performance rating, Stock Option, Percentage on Salary hike, Years at company, Years with current Manager, Years in Current role, Total working Years, Years since Last promotion is negative related with Attrition. We can see that initially when the level of above features is low then there is high rate of attrition, and some sometimes in the middle phase there is spike in Attrition but then again decreases.

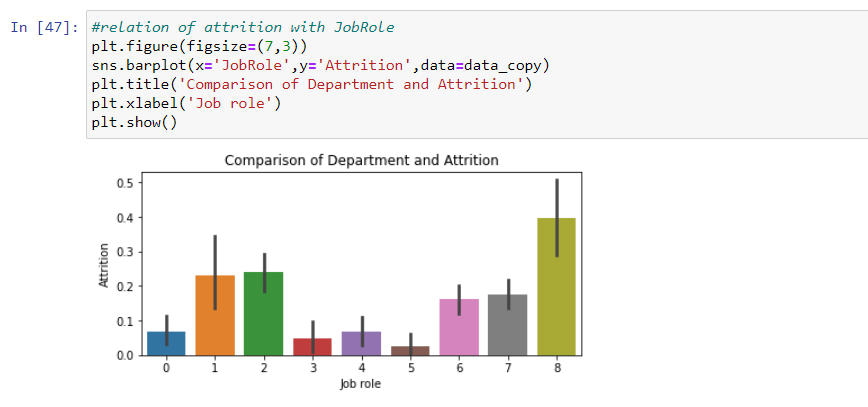
Initially, people get adjusted with their normal work life balance, but later at some level the expect for more and then Attrition comes. But when the work life balance is at its best then the attrition rate decreases.

With having more experience in work, people tend to remain in the company. The initial phase is the one where there is an increase in the rate of Attrition. This can be seen in NumCompaniesworked graph (Last one).

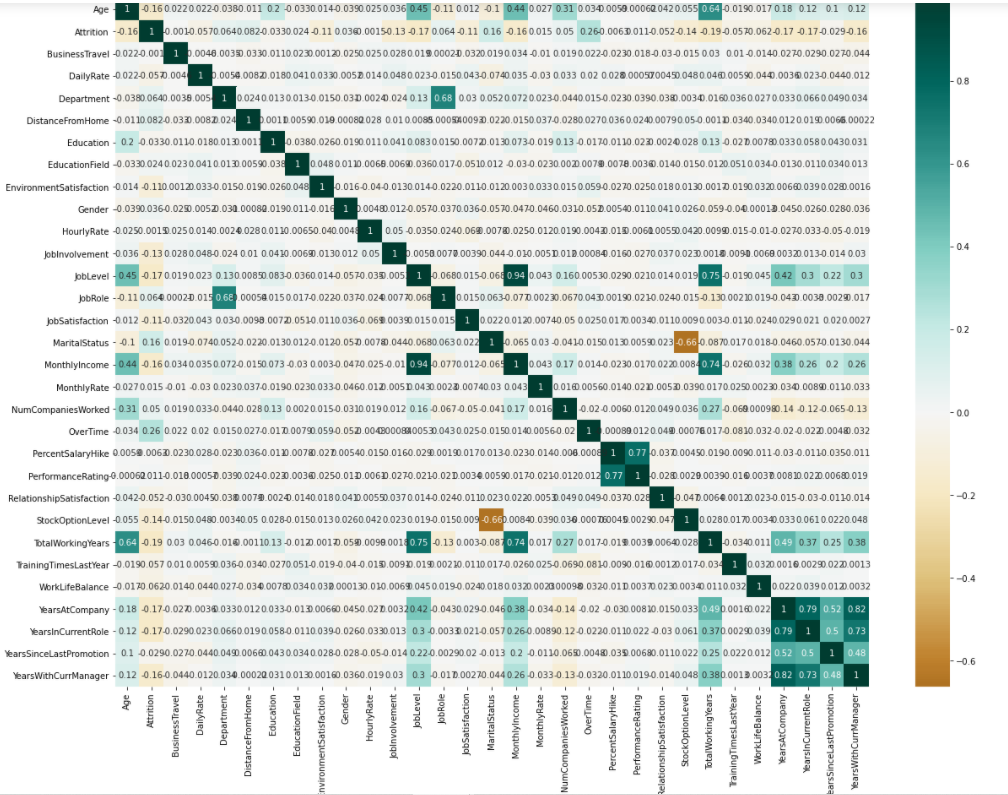
* **Relation between Department and attrition**

   
Here, we can see that in sales department the attrition rate is higher, and in research and development department it is less. The reason being in sales the pressure is higher and also there is a lot of business travel included which is no suitable for the employees.

* **Relation between Job role and attrition**

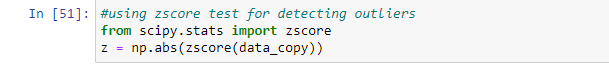
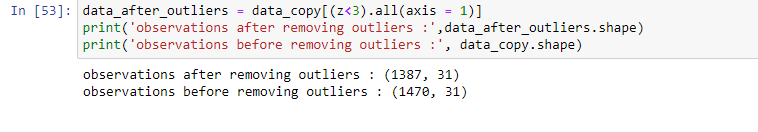
  
Here we see that Sales representatives – 8 have the highest frequency of job role working in a company and higher number of attritions as well. Where Research director –5 shows the least number of frequencies.

* **Correlation metrix**

  
I see that Job level and Monthly Income are very Highly correlated at 0.94, I will remove monthly income for my model building data set from input variable, as With Job level there is more accurate and negative linear relation with Attrition.

some of the Independent features (basically which includes long duration(yearly)) are positively correlated to each other.  
There is high correlation of years at company and years with current manager, years in current role and years with current manager.

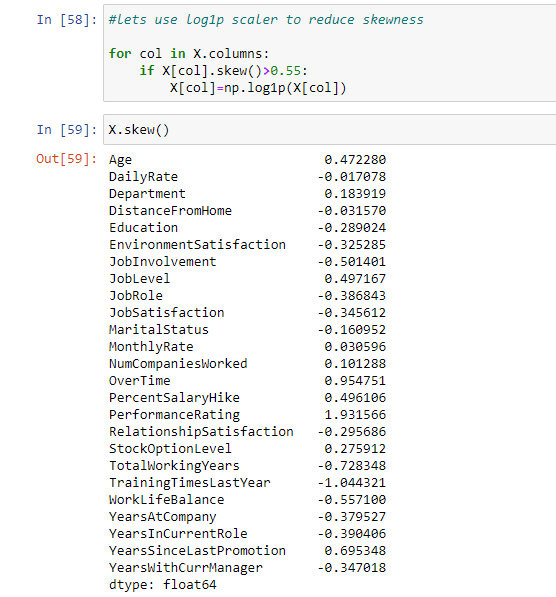
* **OUTLIER DETECTION AND REMOVAL.**

Now, using z-score I have identified the outliers in our data set and removed them.  
  
  
Here, we can see that initially we had 1470 observations, after removing the outliers we are left with 1387 observations.

* **SPLITTING OF DATA INTO INPUTS AND OUTPUT**

After removing the outliers here comes the step to split our data into inputs (independent variables) as X and output (dependent variable) as y.  
We have already used the label encoding in our data set to change the values of categorical values into numeric one, as our model will only support the numerical values for training and testing.



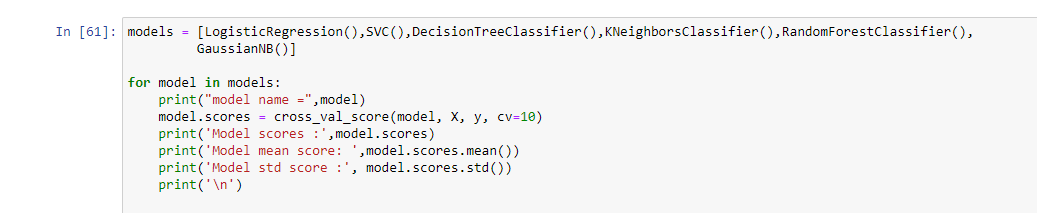
After that I tried to remove the skewness where possible using np.log1p transformation of our inputs data X.  


**4.** **MODEL BUILDING**

* **CROSS VALIDATION TEST**

Now, the steps comes where I will be checking the cross-validation score of different machine-learning algorithms and find the best model which suits our data set and gives us the best accuracy score.

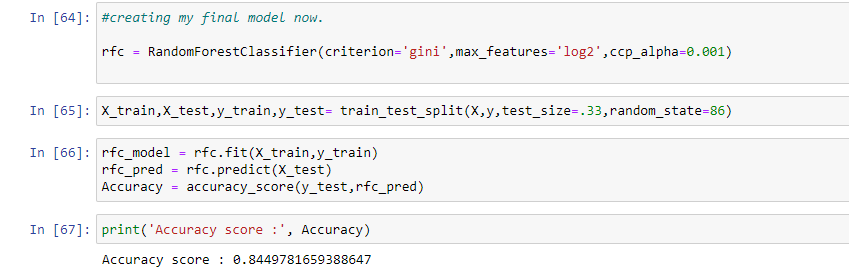
It **is a** technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.



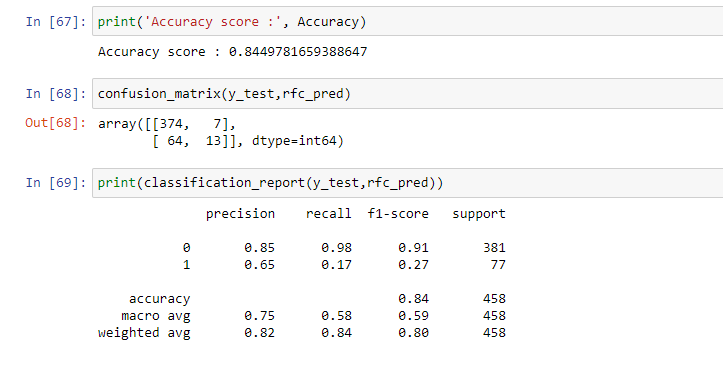
* **FINAL MODEL BUILDING**

From the cross-validation test we get the best means accuracy score from Random forest classification model, so I will be creating my final model building using Random forest classifier.

I have then used grid search CV for parameter hyper tuning for my model, and got an accuracy of 84.49%.

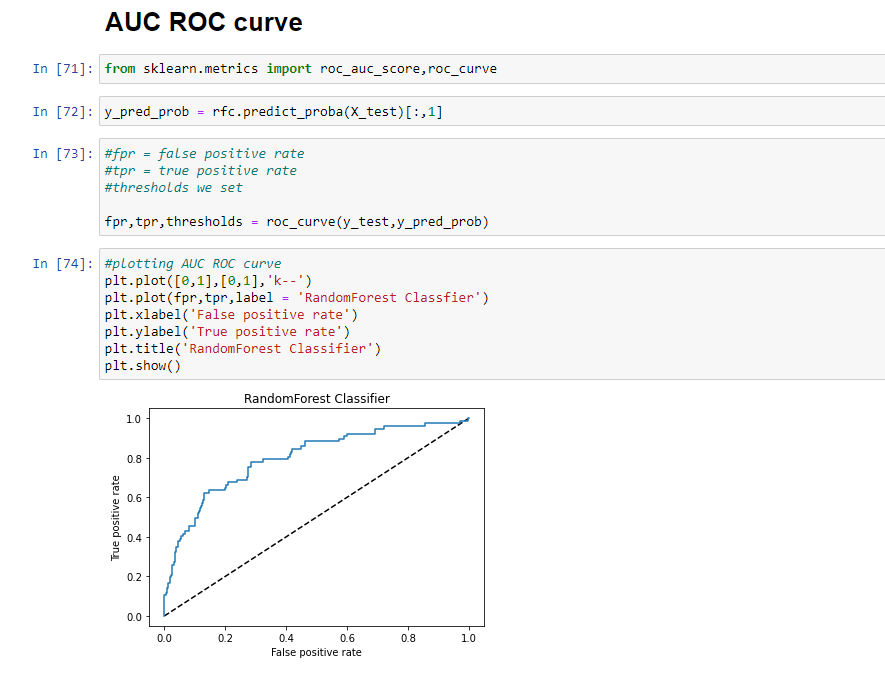


Grid search CV helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. As the model performance always depends upon the values of hyperparameters, we pass predefined values for hyperparameters to the GridSearchCV function. We do this by defining a dictionary in which we mention a particular hyperparameter along with the values it can take. I have searched for criterion, max\_features and ccp\_alpha.  
  
To check our model performance, I also checked the model evaluation metrics and visualise the model performance using AOC-ROC curve.



* **AUC\_ROC CURVE**

AUC\_ROC curve plays a very important role when you have to evaluate your machine learning model visually, the AUC-ROC curve helps us visualize how well our machine learning classifier is performing. It is best for Binary classification problems.  
It uses the values of confusion matrix (true positive rate and false positive rate), here we 1st try to predict the probability value of getting the value of our target variable as 1 and then apply roc\_curve function with our actual and predicted probability values, it will give us the true positive and false positive rate which is used to plot AUC\_ROC curve.



The AUC represents a model’s ability to discriminate between positive and negative classes, and is better suited to this project. As we were required to do a binary classification.

**5.** **CONCLUSION AND REMARKS.**  
  
The main factors that affect the attrition in a company are:

* **Age**  
  we have seen above that employees aged between 26-33 the rate of attrition is higher.  
  Boredom and lack of challenge or growth can also lead to young workers quitting to look for another workplace.  
  These young workers can be retained with the help of employee engagement team, which will motivate them to stay and work effectively and efficiently.
* **Distance from home**  
  Usually as the distance from home to work place increases rate of attrition also increases, to prevent this company should provide cab facilities to their employees or they can provide transport allowance.
* **Business travel**  
  If the travel is more for business purpose then there is high chance of an employee leaving a company. This is a business need, to prevent attrition in this employer should motivate the employees to stay by recognition at work place.
* **Overtime**  
  Employees who are working late hours gets frustrated and actively looking to switch or leave the company, there should be a R&R for such employees and also additional incentives for them to motivate them to stay in the company.
* **Monthly Income**  
  For most of the employees, Salary plays an important role. People with less wages are more likely to leave the company. So, it is required for a company to provide competitive salary/wages to their employees by checking the industrial benchmark in the market. To meet the economic expectation of employee, he desires for a better Pay and shifts to another company. When an employee gets good pay, he tends to remain in the company
* **Percentage salary hike**  
  When the salary hike percentage is more, employees tend to remain in a company, however for every employee the salary hike cannot be the same, it depends upon overall performance ratings of an employee. There should be impartial decision making while rating an employee for its performance.
* **Total working years**  
  As the total working years of an employee increases, the sense of leaving or shifting to a new company decreases.
* **Years with current manager**  
  Here we can see that within 5-6 months there is a huge increase in the attrition rate,  
  this could be because of less interaction with the manager. This can be evaluated by checking up the individual records of the employee under one manager and then taking preventive steps.

Here, we come to end our project discussion.  


I hope the above article was informative and helpful, it helped me as well understanding that HR is not an easy profession and all the question asked at the time of HR discussion are very important.

For complete code, please visit my [git hub repository](https://github.com/sonirl/DT_Evaluation-projects/blob/main/project_4_HRanalytics.ipynb), here I have also created a model using dummy variables for categorical variables and the result was good.

Thanks for your time.

NEVER STOP LEARNING!