### IBM HR Employee Attrition Prediction Using Machine Learning

# 

In this blog article we will be discussing the major problem of the organization which is Employee attrition.. Attrition is the situation faced by an organization when an employee leaves the organization. It costs organizations time and money spent on the employee for interviews and trainings. It impact all the businesses and organizations irrespective of the location, type of the company and industry. Now days it become a major problem for any organization. According to a study by the Centre for American Progress, cost of rehiring an employee can range from 10 to 30 percent of the first year’s salary of that employee. These cost includes: Advertising, Screening and Interviewing, Trainings for new skills.

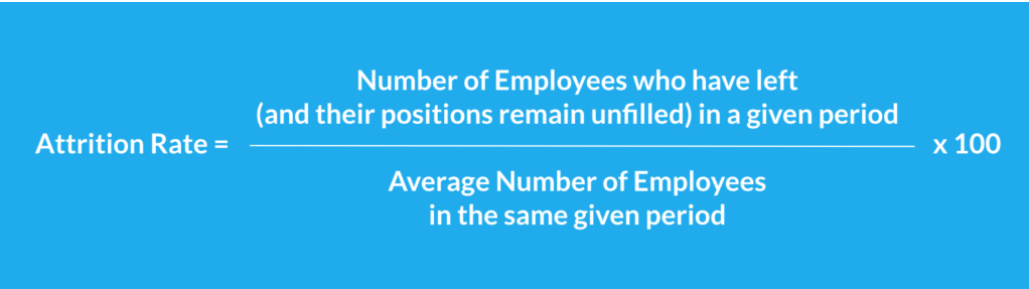
Employee switch jobs to other organization in search of satisfaction and greater facilities for their career growth. In a general if we talk about the cause of attrition it can be dissatisfaction from current job profile, less facilities, no growth, less salary or it can also due to co-worker and most important future stability. Other reason is high demand of employee in particular domain. The attrition will be high with increase in the employment opportunities in the market. We have seen this problem in software industry.

## HR Analytics

To cope up with this situation organizations are taking help of Machine learning techniques to predict the attrition and employee turnover. And here comes the term HR Analytics an area of analytics which refers to applying various analytical processes to the human resource department of an organization in the hope of improving employee growth and getting better returns on the investment. HR analytics is not just gathering the data on employee performance and efficiency, it also provides information and insights of each process and used gathered data to make most efficient decision to improve the process.

According to Heuvel & Bondarouk. HR analytics is defined as the systematic identification and quantification of the people drivers of business outcomes (Heuvel & Bondarouk, 2016).

**Calculation of Attrition Rate:**



# Problem Definition

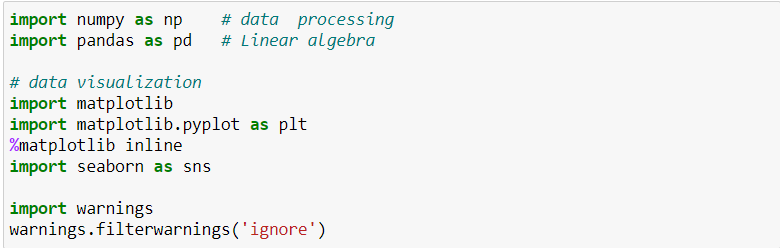
Here our aim is to analyse the employee survey from IBM, which indicates whether there is attrition or not. The data contains 1470 rows and 35 columns. The dataset is limited in size and the model is expected to provide improvement in predicting the attrition of the employee. Classification model can be used to improve the HR’s ability to identify the attrition and taking remedial decision to prevent it.

Here our aim is to predict the employee attrition on the basis of the survey data form IBM. Our main question is how we can use HR Analytics to analyse the attrition. In further steps we will be answering this quetion using some code and data analysis. Let’s start

# Data Analysis

In this section we will analyse and visualize our data using tools available in python for analysis like pandas data frame for data processing, Numpy for linear algebra and Matplotlib, Seaborn for visualization. Let’s start analysis

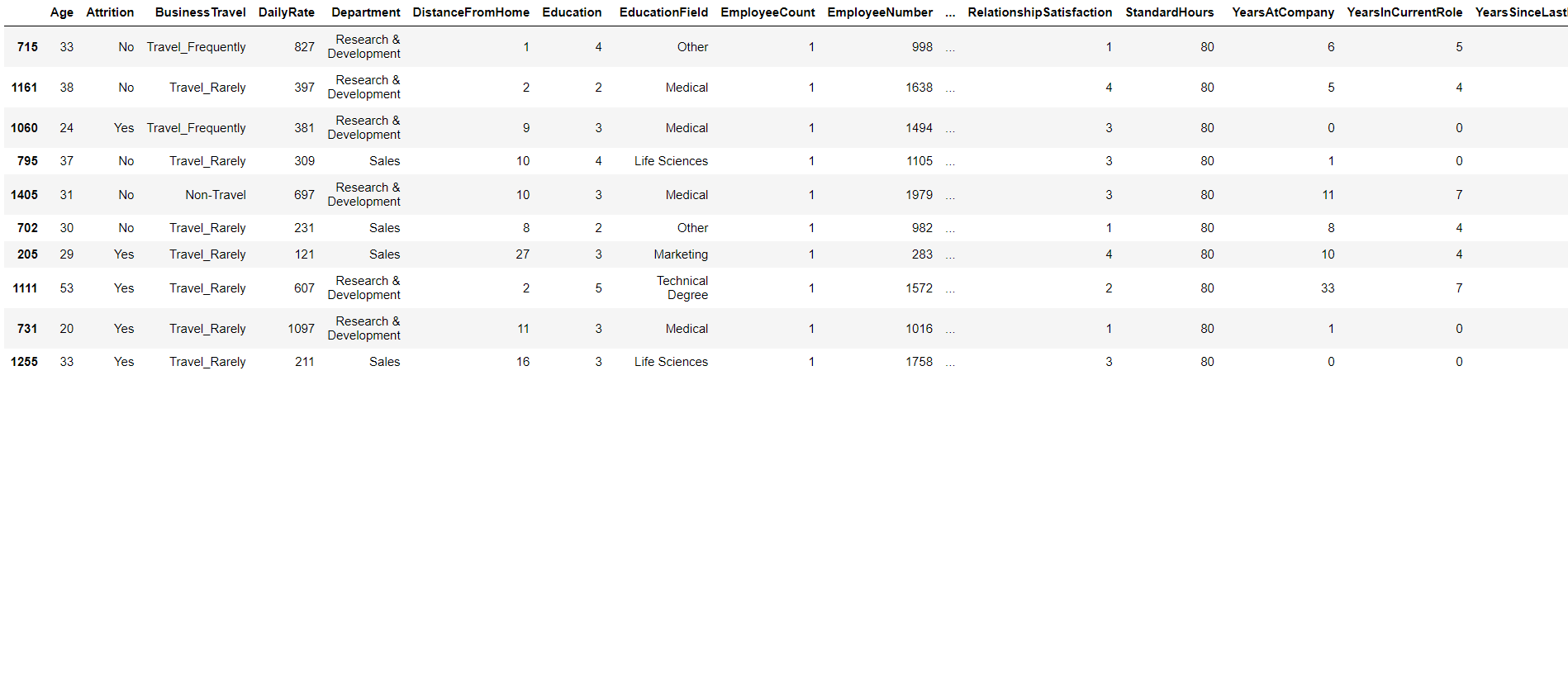
## Importing all the required libraries:



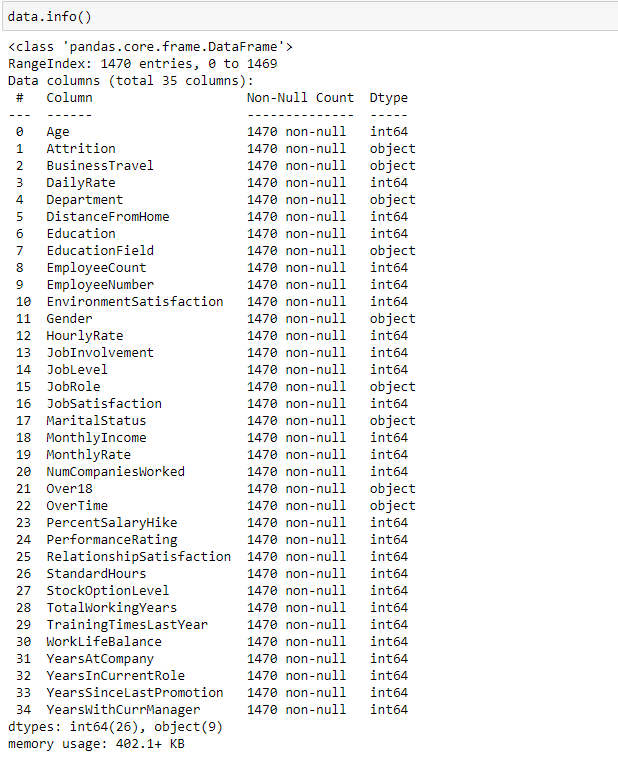
## Loading the dataset:



Now our data set is loaded into our data variable, let’s check the ten sample for our data.

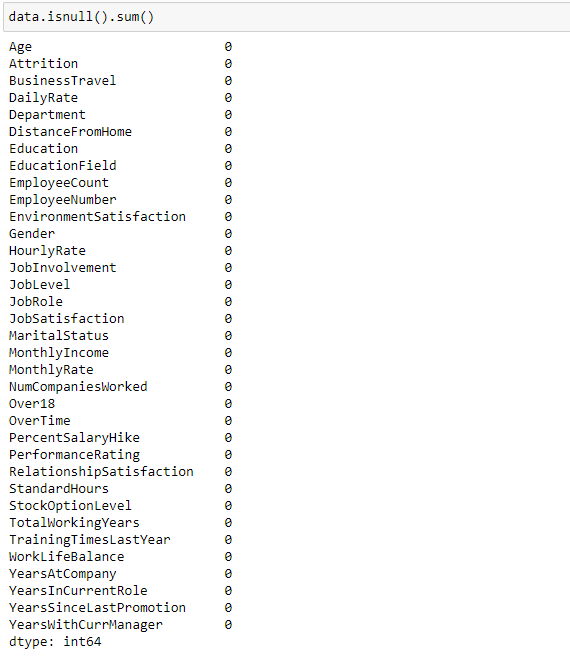


From sample we can see we have columns like Age, Attrition, Daily rate, Department, Distance from home, Education etc. Attrition is our target column and all other columns are feature columns. We have only two type of values in Attrition ‘Yes’ and ‘No’, which indicates it is classification problem.

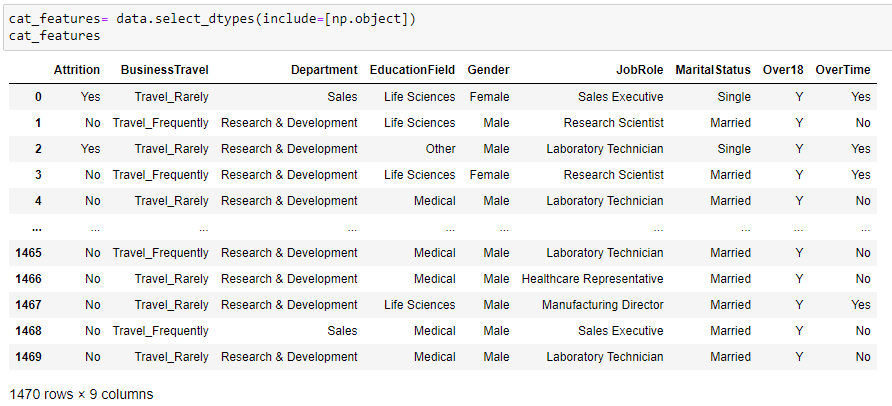


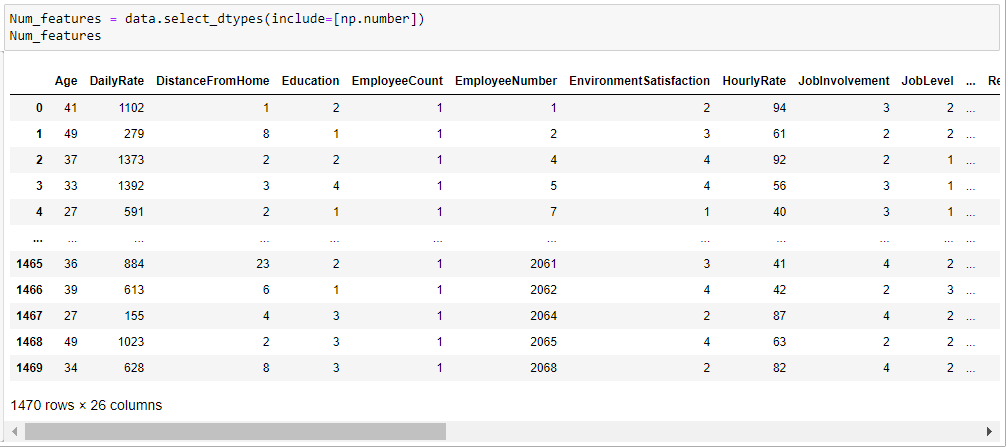
From data information we can see data we have total 1470 entries, it is having 1470 Rows and 35 columns. All the columns have 1470 non-null values. **Here Attrition is dependent variable and other columns are independent variable.** We have 26 integers and 9 object type values. Let’s check the null values present in the data.

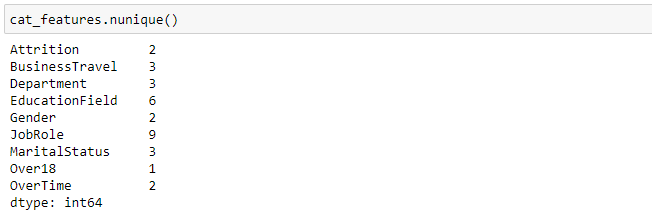
## Checking Null Values:

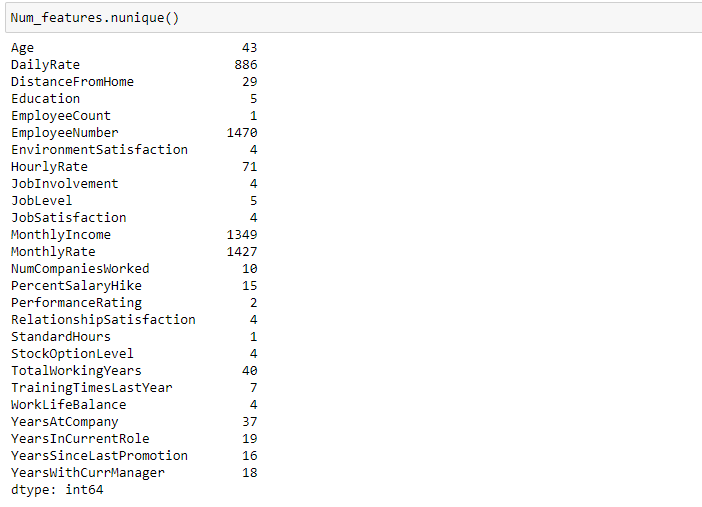


Luckily our Data is not having any null values. Let’s divide the dataset into categorical and numerical columns, for better data handling.



 Checking unique values present in the Categorical data and Numerical data.

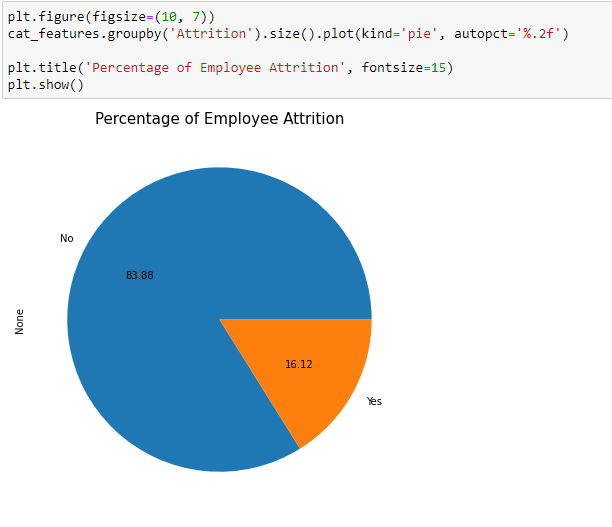




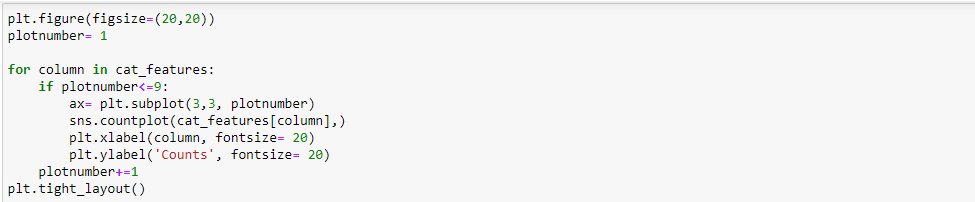
Here we can see number of unique values present in all the categorical and numerical columns. Here nunique() function gives the number of unique values present in data frame. Let’s visualise data using visualisation technique. We will visualise data using univariate, bivariate and multivariate data analysis.

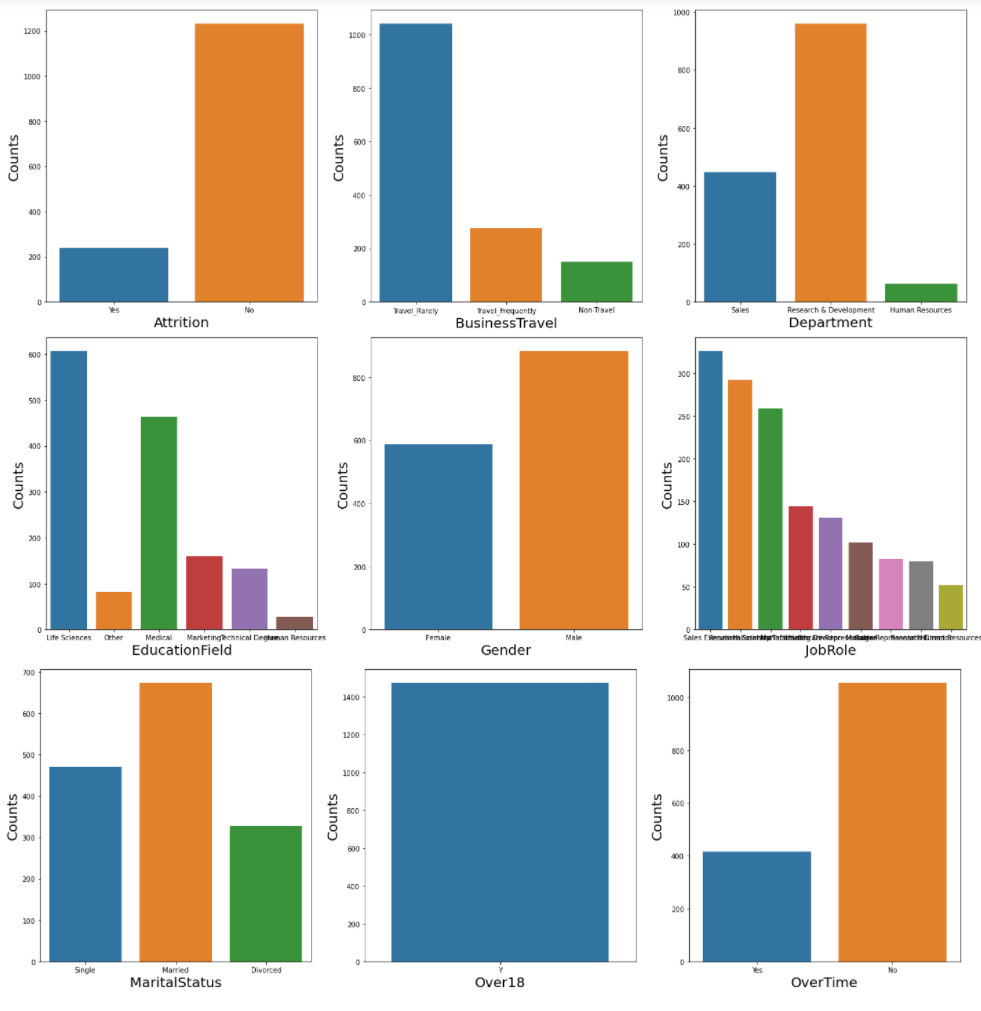
# Univariate Analysis:

Let’s first analyse Attrition Percentage.



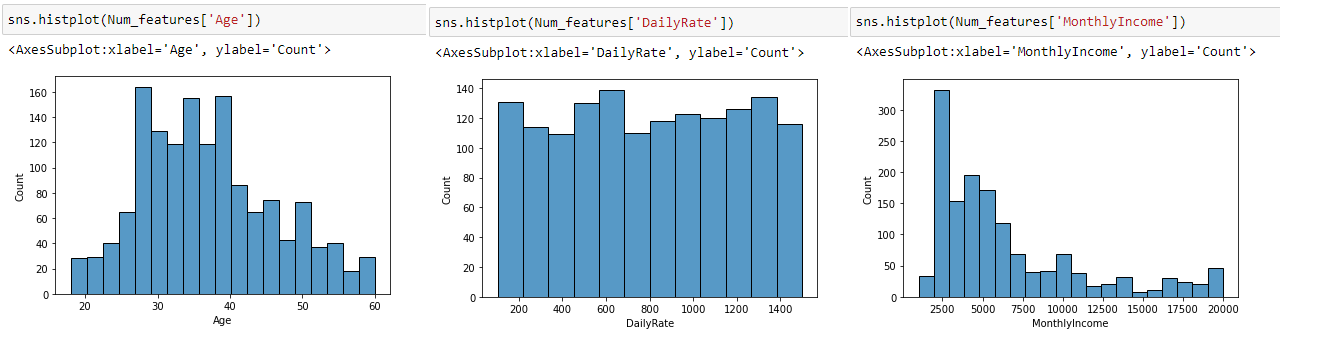
Pie chart indicates 16.12 percent of employee data is having attrition, Means 16.12% are more likely to quite the job.



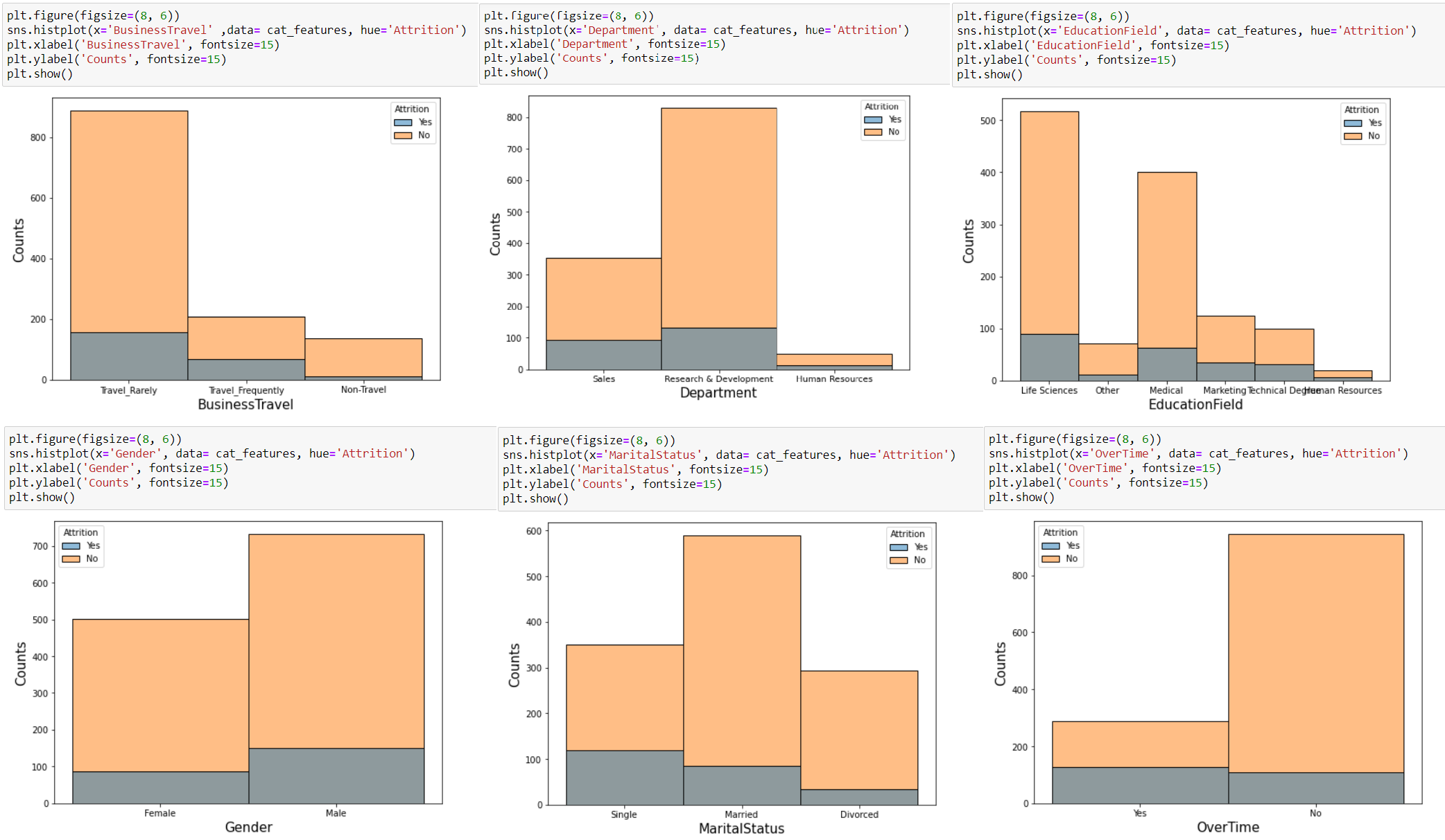


* **Attrition:** columns, distribution of data is not equal, data is imbalanced. Before model building we will balance our data using SMOTE.
* **Business travel:** the frequency of traveling is very less as compare to frequency of rarely traveling job profile.
* **Department:** In department column frequency of Research and Development is high as compare to other.
* **Gender:** Most of the employee are male and Married employee are more as compare to other single and divorced. All the employee are over 18 and Count for employee who do overtime is less.

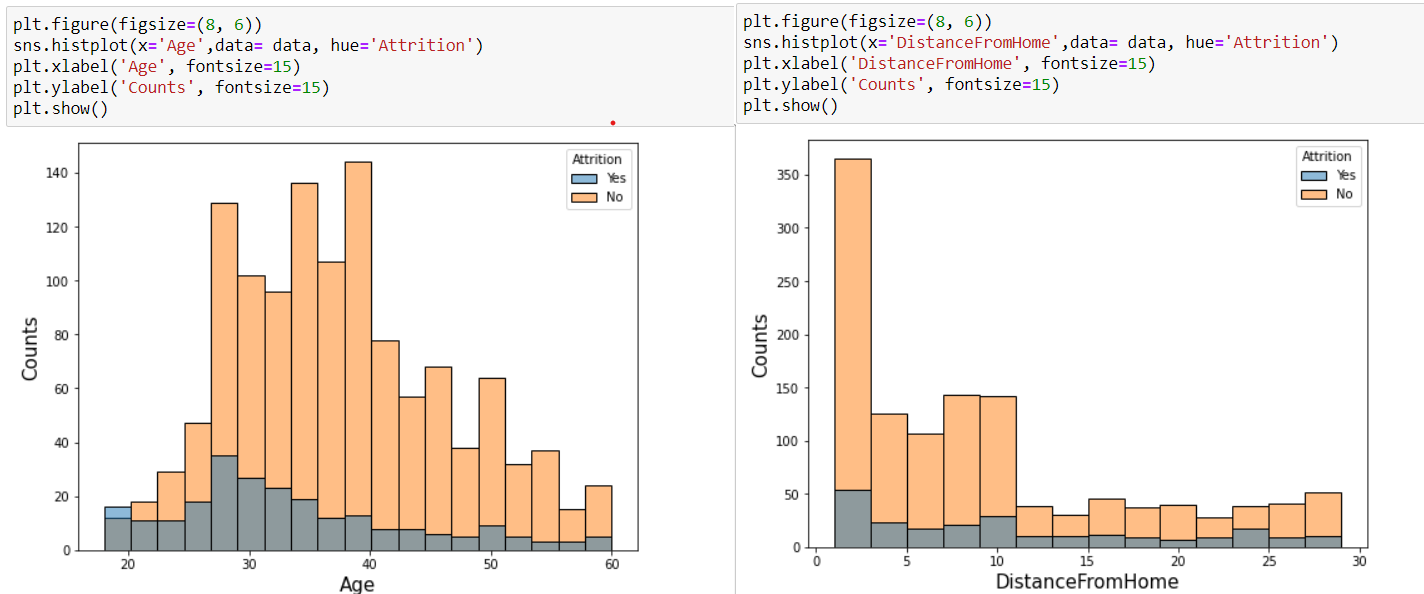
Let’s analyse the distribution of few Numerical columns.



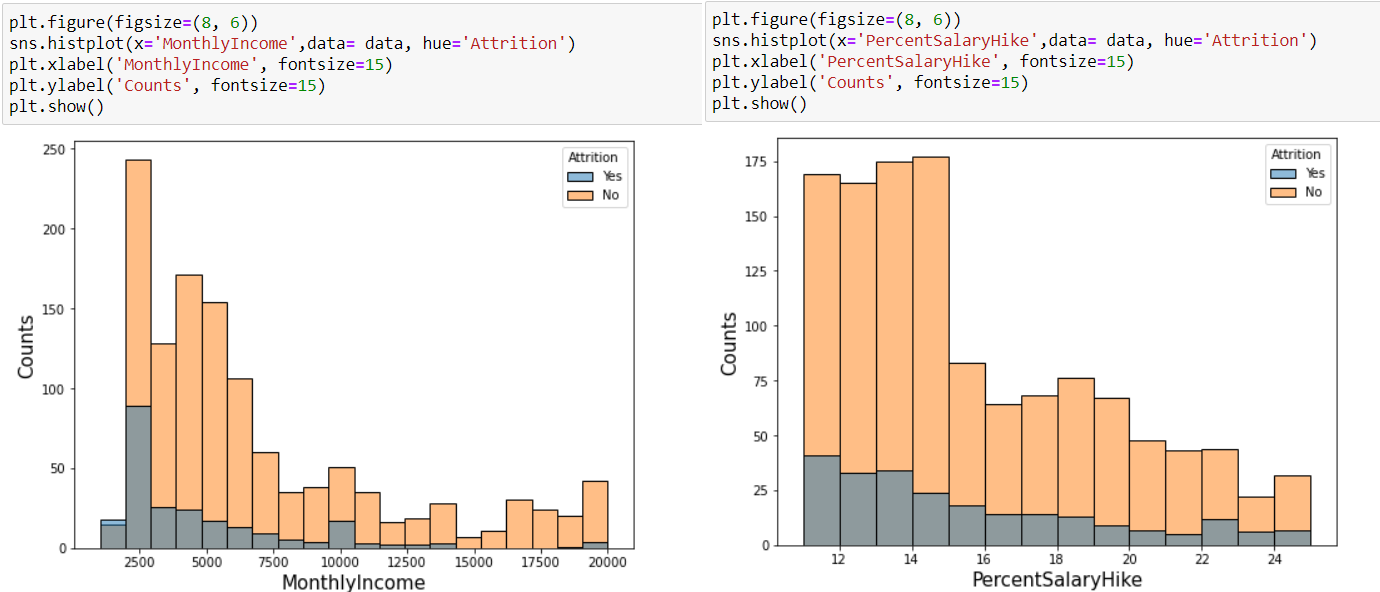
We can see all the employee are above 18 and most of the working employee lie in range between 28-40 and the daily rates are almost equal. If we talk about the income most of the employee earning 25000, most of the employee working in salary range of 2500 to 60000. Let’s analyse our data using bivariate analysis. Here we will analyse out data with respect to our target columns.



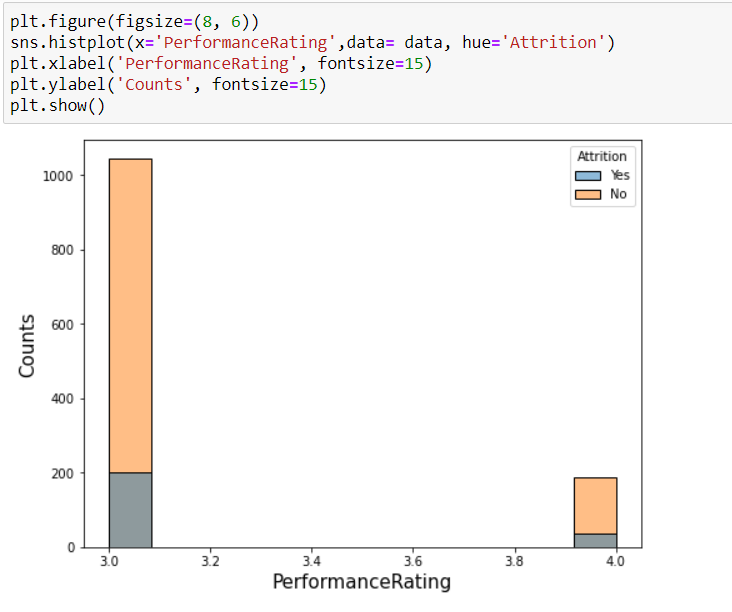
We can see a very good relation between Business Travel and Attrition, employee traveling frequently are more likely to quit the job. It is obvious job which are having high traveling frequency make employee unsatisfied and the security also a big concern in job with high traveling frequency. If we talk about the department employee working in sales department are more likely quit the job. A very good relation we can see overtime and attrition, employee working overtime are more likely to quite the job. We can see single employee are not stable, they are more likely to quite the job, married employee more stable than the single. Let’s analyse relation between Age and distance with Attrition.



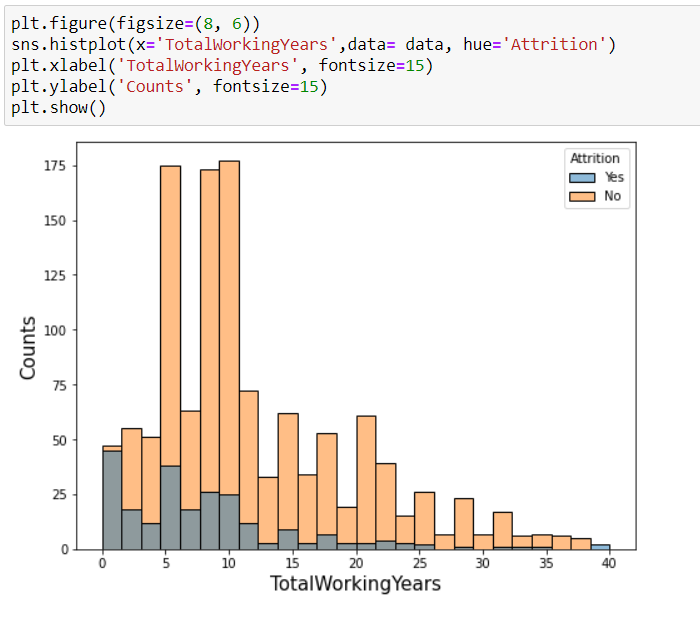
Here we can see attrition is very high between ages 20- 40. It because employee change job in search of stability and satisfaction. The attrition is very high 26-34. If we talk about Distance From home, most of the employee don’t want to work near their home. Employee having very less distance from office are more likely to quit the job. Let’s analyse most important parameter, for job satisfaction that is income and salary hike.



We can see employee having income in range on 20000-25000 are more likely to quite the job. Attrition tends to decrease with increase in the salary. Same we can see with the hike in the salary, lower the percentage of hike in salary higher the rate of attrition, we can clearly see this in the second graph. For percentage hike till 16 percentage, employee are not satisfied with the salary and more likely to quite the job.



Another factor which affect more is performance rating, lower the rating higher the rate of attrition. Let’s analyse total working years. In the below graph we can see with increase in total working year, attrition decrease. Which indicates employee in the early years are more likely to change the job.



# EDA Conclusion:

* The employee with Low MonthlyIncome, RelationshipSatisfaction, PercentageHike, JobInvolvement, EnvironmentSatisfaction, YearAtcompany, Totalworkingyear are more likely to quit the job.
* BusinessTravel: Employee with higher Business travel are more likely to quit the job.
* DistanceFromHome: Employee living very near to home are more likely to quite the job.
* Department: Employee in sales department are more likely to quite the job.
* Environment Satisfaction: Employee with low satisfaction are more likely to quite the job.
* Marital status: Single employee are more likely to quite the job.
* Number Companies worked: Employee who worked in 1-2 company only, are more likely to quite the job.
* Overtime: Employee doing overtime are more likely to quite the job.

# Data-Pre-Processing:

In this stage we will pre-process our data to make it relevant for model building. We will first drop all the columns which are not relevant. We will encode all the categorical feature columns using one hot encoder and target column using Label Encoder. After that we will check correlation, skewness and Scale the feature data using Standard Scaler.

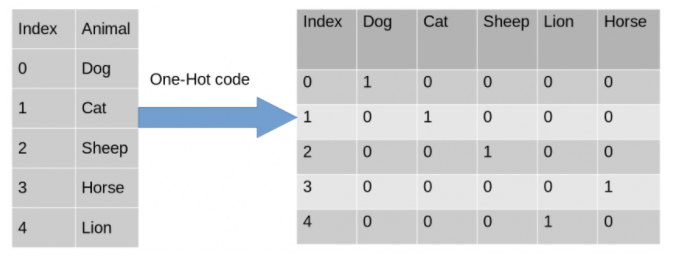
## Dropping Data:

Columns which need to be dropped: Employee Count, Standard Hours, Employee Number, Over18



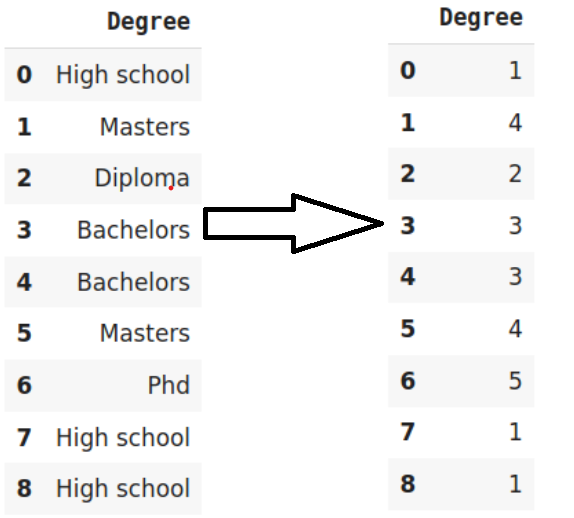
## One Hot Encoder:

We will encode our data using pandas get dummies function. Let’s first discuss how it works. One hot Encoding is used to encode all the categorical data which are nominal. Many machine learning algorithm not performing well on categorical data. So we require all the output variable to be numeric. In this method Each category is mapped with binary variable either 0 or 1. In this method we create dummy variable which represent each category of the data. Presence indicate 1 and absence indicate 0. Let’s take an example of animal category.

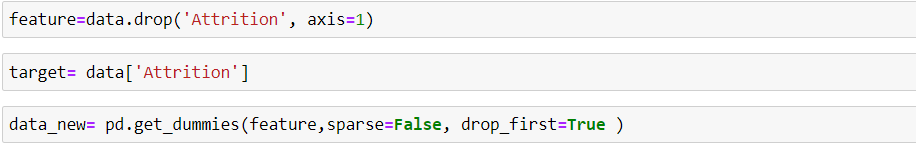


After encoding, in the second table, we can see dummy variables columns represent each variables. In each columns we can see ‘1’ representing presence of that dummy variable and ‘0’ represent absence.

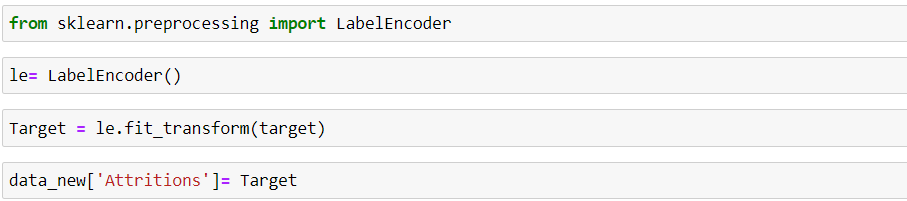
**Label Encoder:** We will apply one hot encoder on all the feature columns, for Target column we will use LableEncoder, as the name indicates it always used for encoding the target label data. In label encoding each label is converted to numerical values. Check the example below.



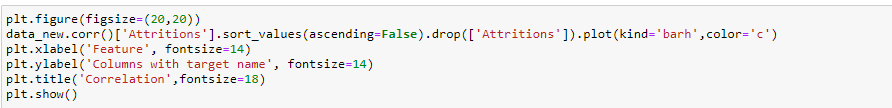
After encoding each category is encoded with a numerical values, like High School encoded as 1,Bachelors as 2 and so on.

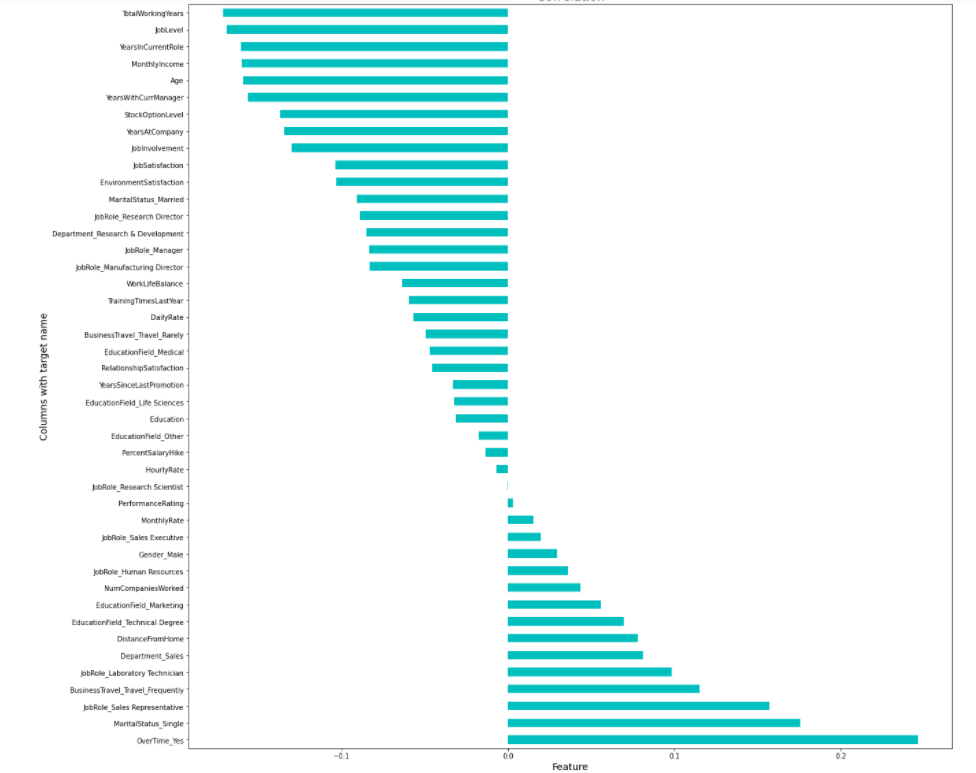


Here we used get\_dummies and we dropped First column to remove any multicollinearity



**Let’s check the correlation:**

****

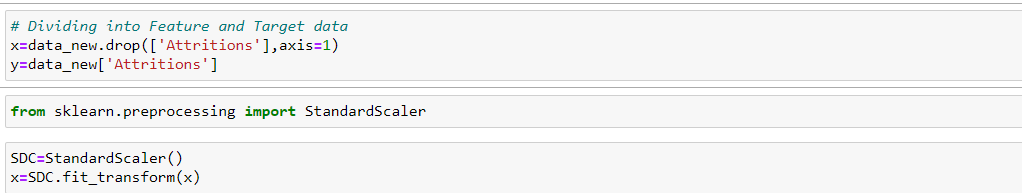
****

From the graph above we can see data on the left of ‘0.0’ is negatively correlated and on the right it is positively correlated.

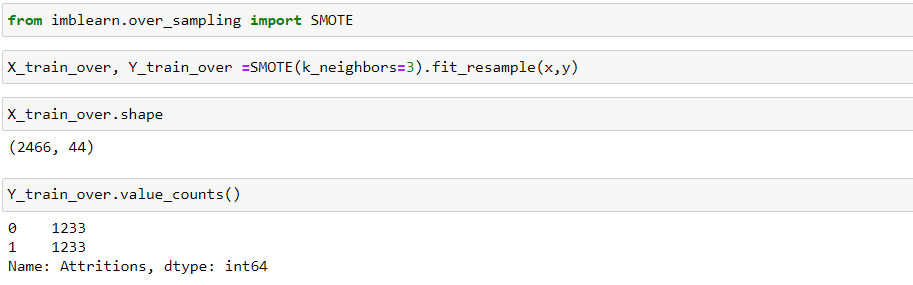
As all the data is categorical we will not remove any outliers.

After checking the skewness we can see DistanceFromHome and MonthlyIncome are having skewness. Lest remove it using numpy log transform.

 Scaling the data Using Standard Scaler:



Earlier we have seen our target column was imbalanced, so we will use SMOTE method for up sample our data.



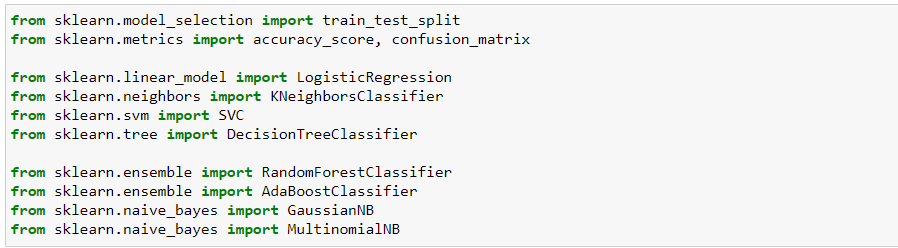
We can see Y\_train\_over is now up sampled, total values for 0 and 1 are equal now.

After checking VIF we have seen ‘Department\_Research & Development’, ‘Department\_Sales’, ‘EducationField\_Medical’ are having multicollinearity. We will drop these columns and proceed for model development.

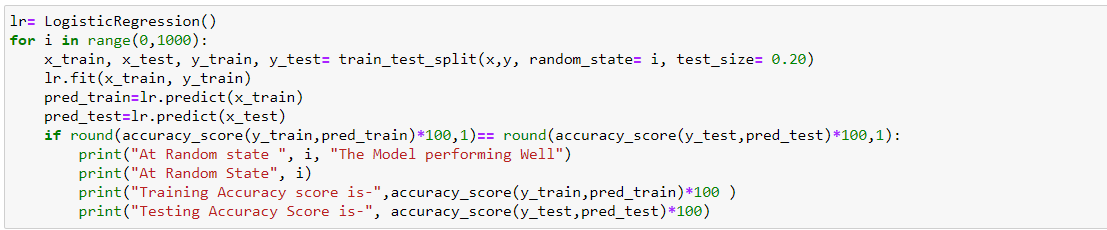
# Building Machine Learning Models:

Our Target column is having two kind of values either yes or no which we encoded using label encoder. So its classification based problem. In model development stage we will start with the basic classification model and end up with complex models like Decision Tree, Random forest and adaboost classifier. We will check the accuracy and cross validation for each model and select the best model which will be having lesser difference in their cross val score and accuracy score.

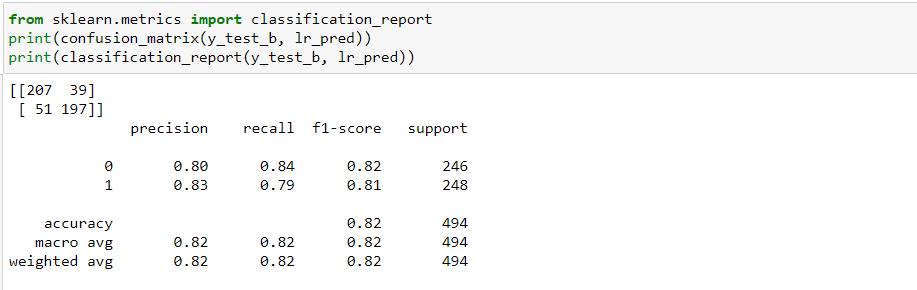
**Importing all the models:**



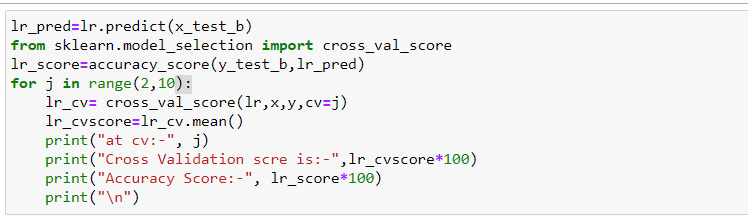
## Choosing Best Random State:







For linear model we have accuracy score 81.78 percent. We will check Cross validation score.



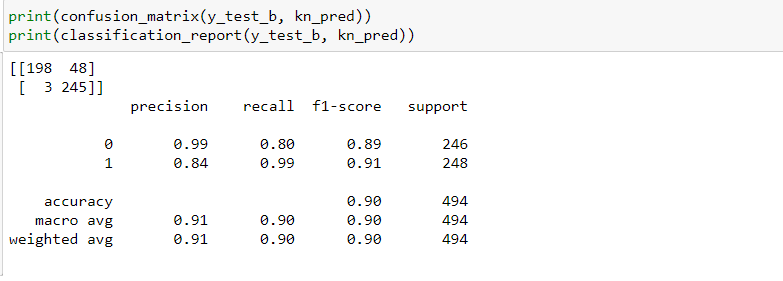
At CV= 6, Cross val score 81.83292781832928 and Accuracy Score: - 81.78137651821862 are almost same.

This is the best model till now. Let’s try another model.



Both the score having bigger difference as compere to logistic regression model. Let’s try another model.

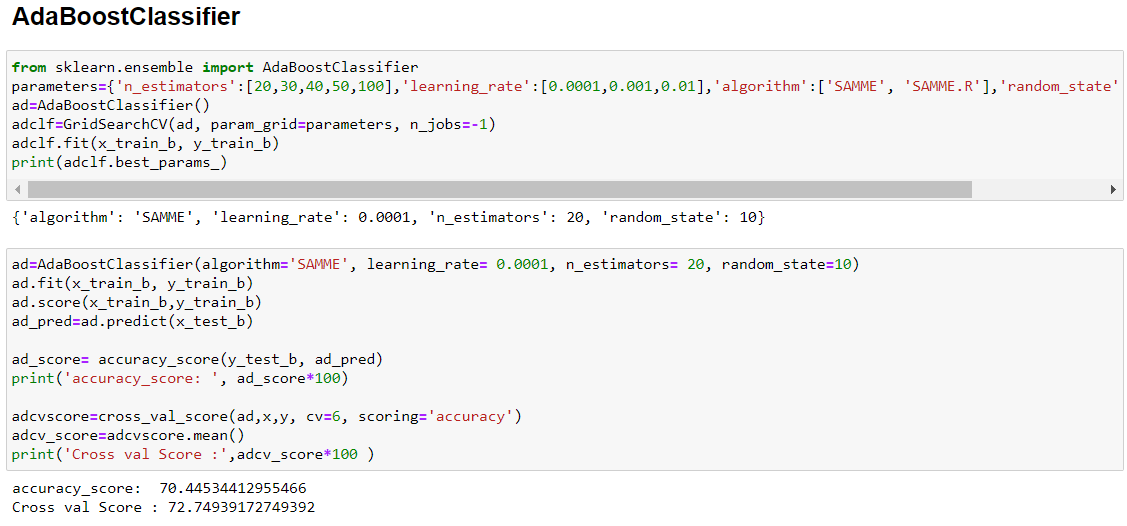


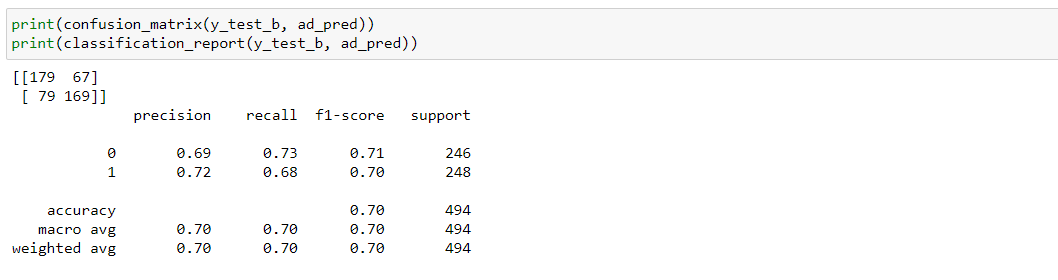


Both the score having big difference as compare to logistics regression, logistic model is still our best model. Let’s tyr Random Forest Classifier



Both the score are almost same but difference is considerable as compare to logistic model.





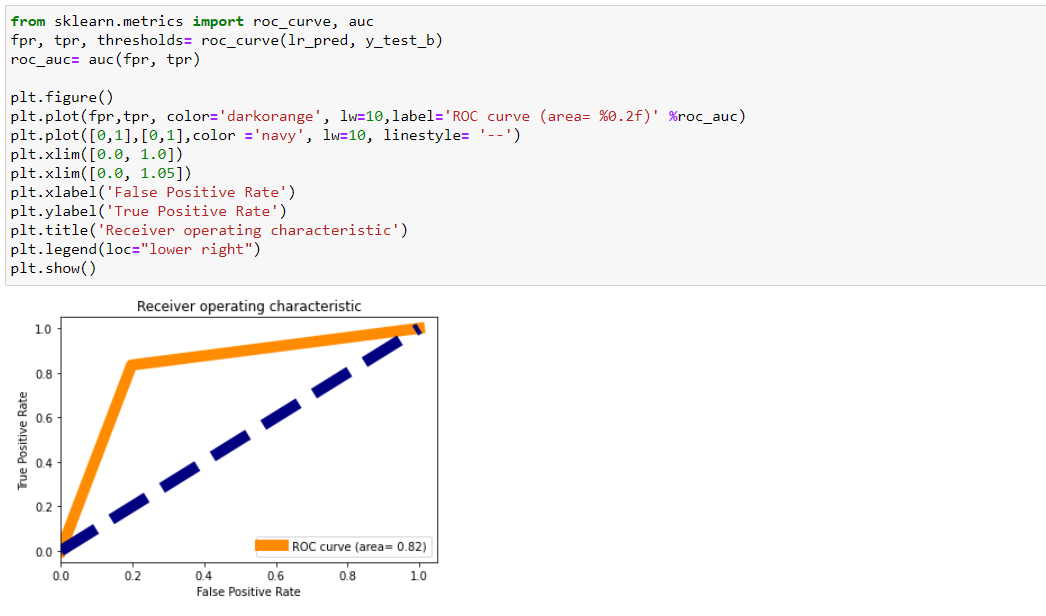
Both the score are less, model performance is bad. Let’s try another model.



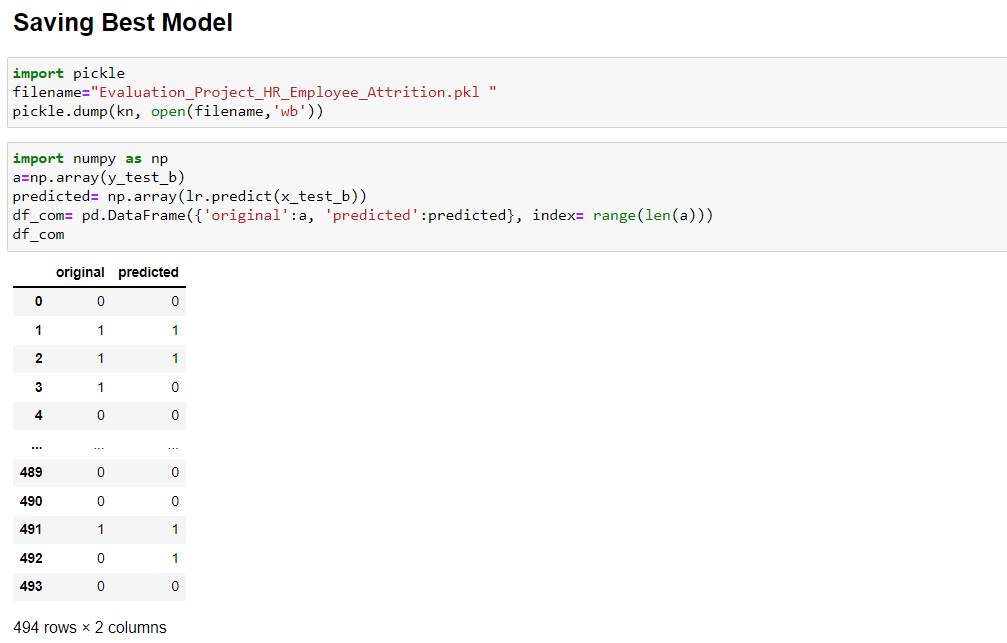
Decision tree classifier performing well but still there is difference in the score as compare to logistic model.

# Concluding Remarks

After checking all the model we came to conclusion that logistic regression performing well. With almost equal accuracy score and cross validation score. Let’s save the best model, before that lets draw the AOC- ROC score.



## Saving best model:



We can conclude that with help of machine learning or HR analytics we can predict the behaviour or employee whether there will be attrition or not, It will help HR to take some remedial steps to stop the attrition.