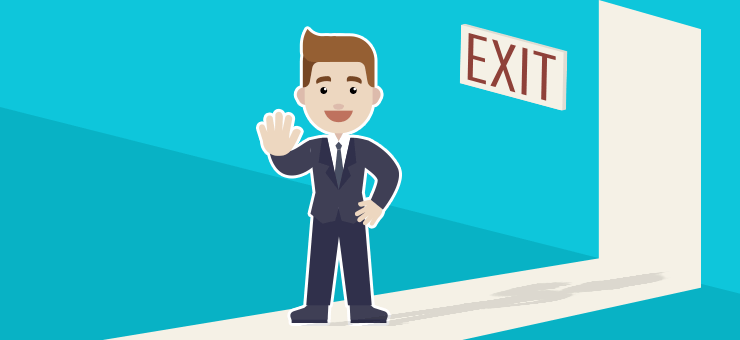
**HR Analytics Project- Understanding the Attrition in HR Using Machine Learning.**



In this blog, We will be going through the whole process of creating a machine learning model on the [**IBM HR analytics employee attrition**](https://github.com/dsrscientist/IBM_HR_Attrition_Rate_Analytics/commit/9ec6236290b60a0cbf510d7ada30d84888635692)**.** Nowlet us quickly go through our problem statement.

**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.

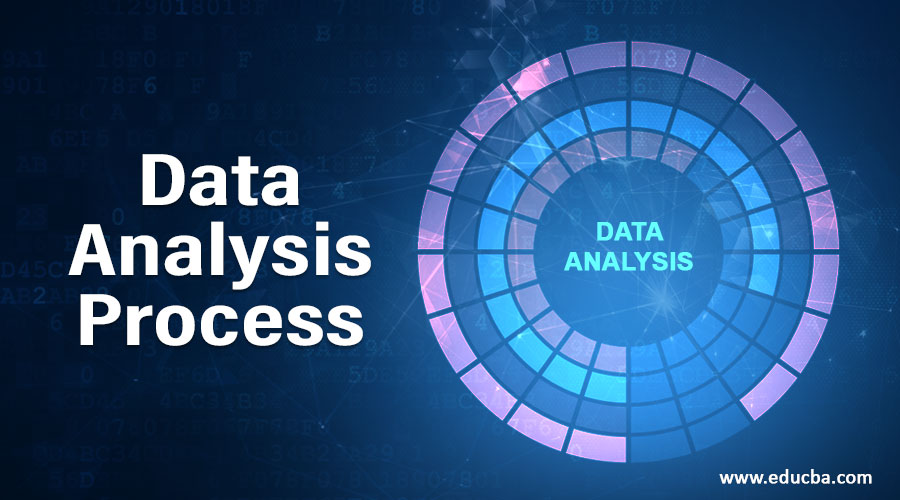
How does Attrition affect companies? and how does HR Analytics help in analyzing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.



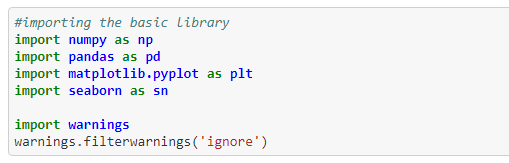
Attrition affecting Companies

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.

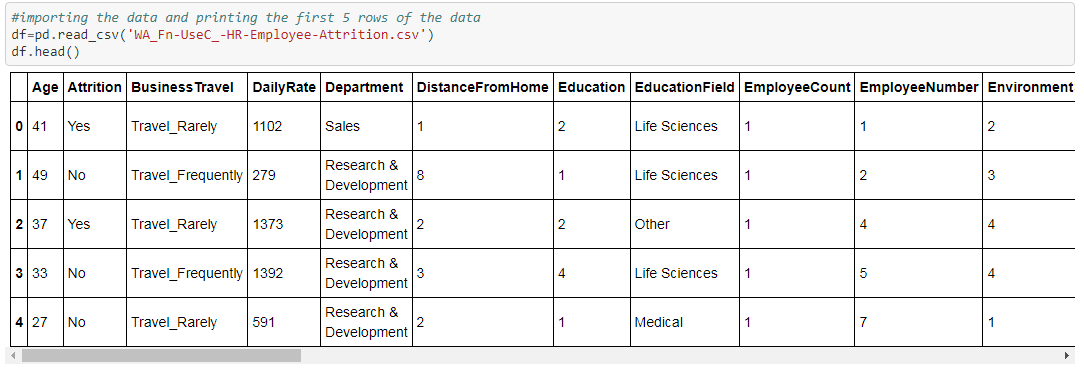
**Data Analysis:**

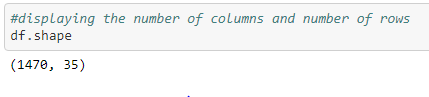
****

Data analysis helps us to get the insight of our data set apart from this we can clean, transform data to discover useful information which in turn help us to make Business decision. So before making any decision let us go through our dataset. Now lets us try understanding our columns, before that lets us import the entire basic library that is needed to carry out our analysis.



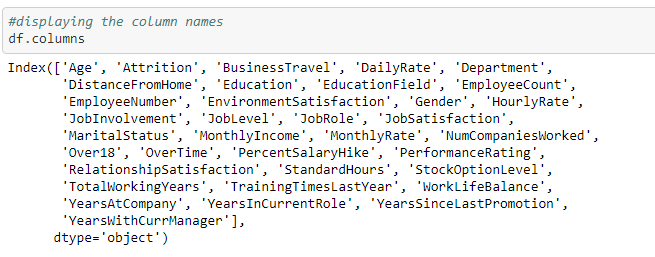
Now let us import our data set.

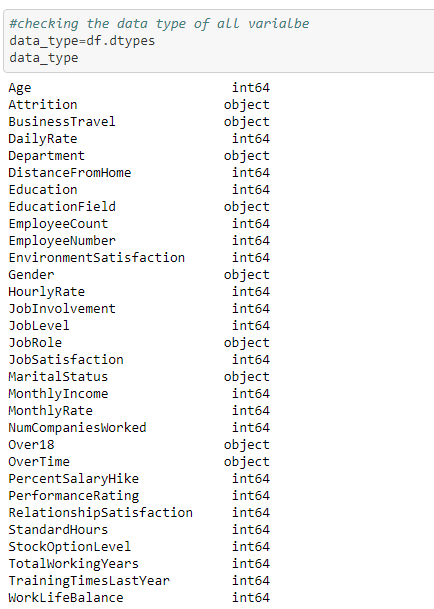


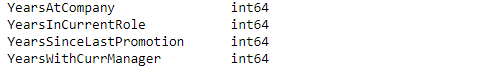


* We have 35 columns and 1470 rows, We can say it’s a small data set.
* In columns we have got both Dependent variable and Independent variable, so it’s a Supervised machine learning.
* Our Dependent variable is “Attrition” and it’s a Categorical Nominal type. This in turn tells us that it’s a classification problem.

Now let’s study our column in detail, by studying the column we will get a lot of information and understanding about our data set.

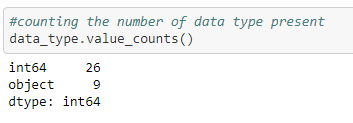






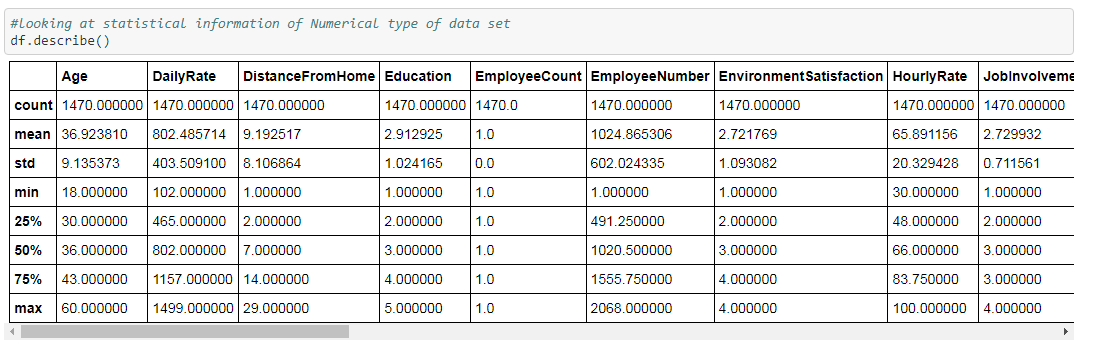
From above we can see that

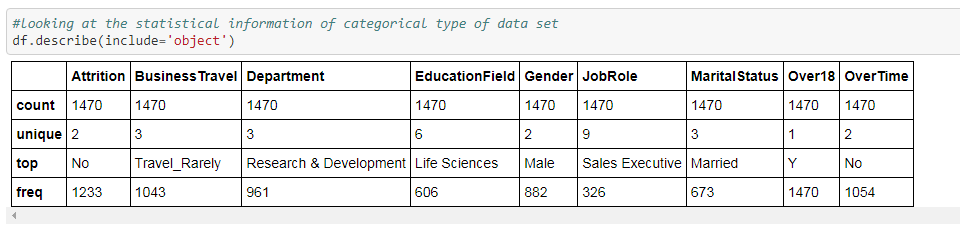
|  |  |
| --- | --- |
| **Name** | **Description** |
| AGE | Numerical continuous Value |
| ATTRITION | Employee leaving the company (yes, no) |
| BUSINESS TRAVEL | Non-Travel, Travel Frequently, Travel Rarely |
| DAILY RATE | Numerical continuous Value |
| DEPARTMENT | Research & Development, Sales, Human Resources |
| DISTANCE FROM HOME | Numerical continuous Value |
| EDUCATION | Numerical Discrete Value |
| EDUCATION FIELD | Life Sciences, Medical, Marketing, Technical Degree, Other, Human Resources |
| EMPLOYEE COUNT | Numerical Value |
| EMPLOYEE NUMBER | Numerical Value - EMPLOYEE ID |
| ENVIROMENT SATISFACTION | Numerical Value - SATISFACTION WITH THE ENVIROMENT |
| GENDER | FEMALE, MALE |
| HOURLY RATE | Numerical Value – HOURLY SALARY |
| JOB INVOLVEMENT | Numerical Value - JOB INVOLVEMENT |
| JOB LEVEL | Numerical Value - LEVEL OF JOB |
| JOB ROLE | Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative, Manager, Sales Representative, Research Director, Human Resources |
| JOB SATISFACTION | Numerical Value - SATISFACTION WITH THE JOB |
| MARITAL STATUS | DIVORCED, MARRIED, SINGLE |
| MONTHLY INCOME | Numerical continuous Value |
| MONTHY RATE | Numerical Value - MONTHY RATE |
| NUM COMPANIES WORKED | Numerical Discrete Value – Number of companies worked |
| OVER 18 | YES, NO |
| OVERTIME | YES, NO |
| PERCENT SALARY HIKE | Numerical Value - PERCENTAGE INCREASE IN SALARY |
| PERFORMANCE RATING | Numerical Value - PERFORMANCE RATING |
| RELATIONS SATISFACTION | Numerical Value - RELATIONS SATISFACTION |
| STANDARD HOURS | Numerical Value - STANDARD HOURS |
| STOCK OPTIONS LEVEL | Numerical Value - STOCK OPTIONS |
| TOTAL WORKING YEARS | Numerical Value - TOTAL YEARS WORKED |
| TRAINING TIMES LAST YEAR | Numerical Value - HOURS SPENT TRAINING |
| WORK LIFE BALANCE | Numerical Value - TIME SPENT BEWTWEEN WORK AND OUTSIDE |
| YEARS AT COMPANY | Numerical Value - TOTAL NUMBER OF YEARS AT THE COMPNAY |
| YEARS IN CURRENT ROLE | Numerical Value -YEARS IN CURRENT ROLE |
| YEARS SINCE LAST PROMOTION | Numerical Value - LAST PROMOTION |
| YEARS WITH CURRENT MANAGER | Numerical Value - YEARS SPENT WITH CURRENT MANAGER |



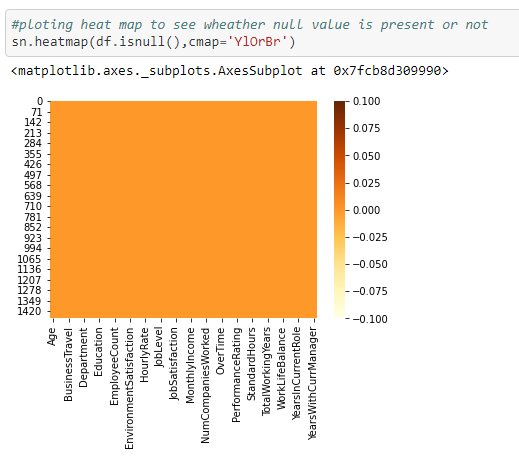
We can see that our data set has dependent variable of integer type and object type apart from that we don’t have any other type of data types.

We can also check the **statistical data** of our data set, here I will be using describe() function from which we can get almost all the statistical information and it has become one of my fovourite function to carry out data analysis.

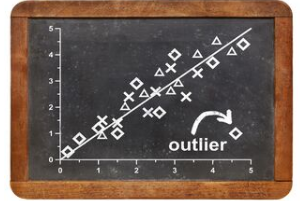




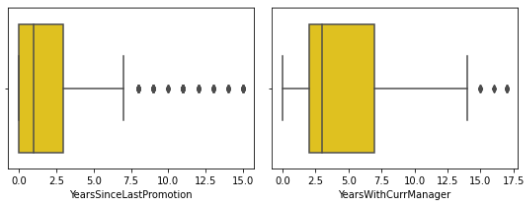
From the function isnull() we can see that we don’t have any **null/missing value** present in our datat set. Lets plot the heatmap to display the null value, the below heat map is plain with only one colour from this we can say that our data set has no missing values.

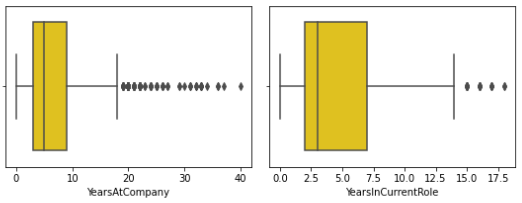


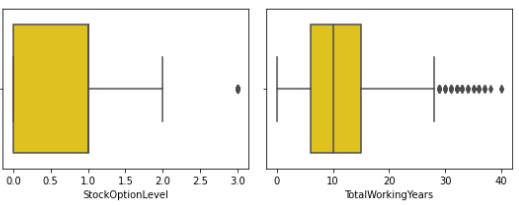
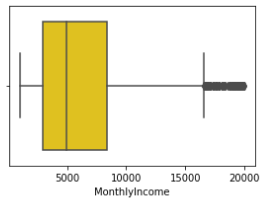
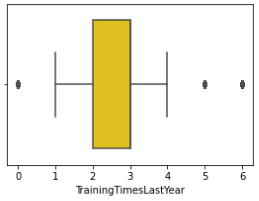
Now let’s check for **outlier** in our data set.



For that we can make use of the box plot, Very simple and most effective plot to find outlier





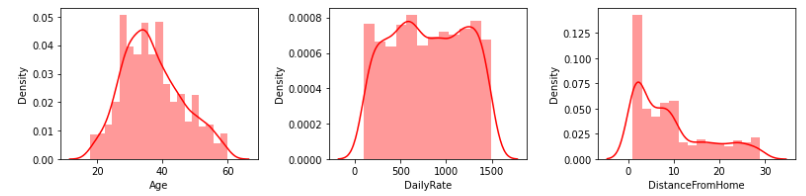
  

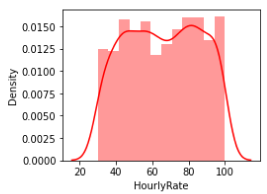
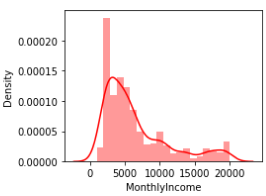
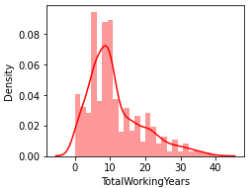
Didn’t I tell you that it’s very simple to identify the outlier with box plot! The points beyond the whiskers are the outlier. We can see that some outlier is present and outlier should be treated or removed as this are the unwanted/misleading data points.

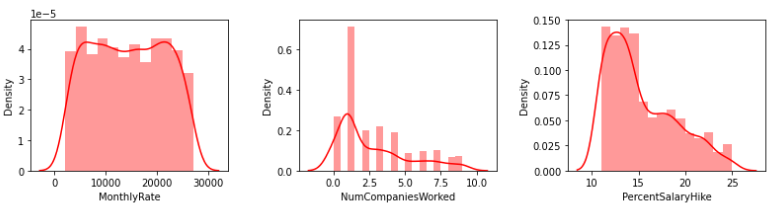
Most of the Machine learning Model assume that our model has no **skewness** and Tail data will act as the outlier this will affect the performance of the data. So we must minimize the skewness as much as possible.

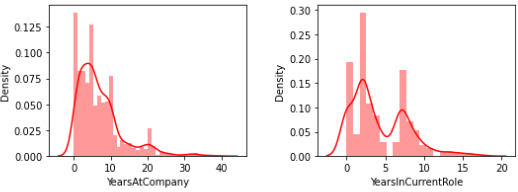
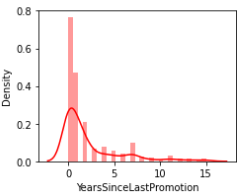


We can make use of dist plot and see how much our data is skewed



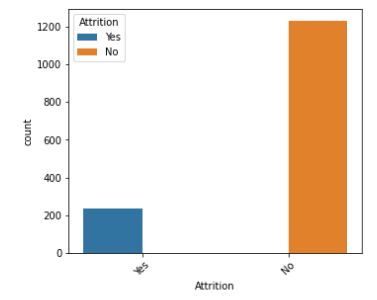


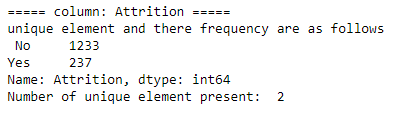
From the above graph we can say that we have some skewness in our data set and we need to treat/remove it before model building.

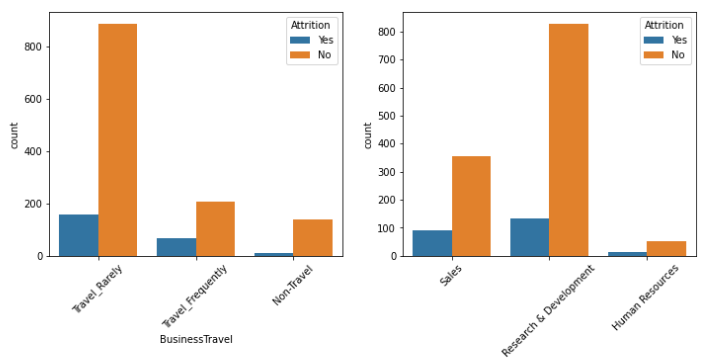
Apart from this we have got some good package in seaborn visualization library. Visualization helps us to get more information from simple graphs and charts, So let’s make use of it in order to get some knee important information from our data set.

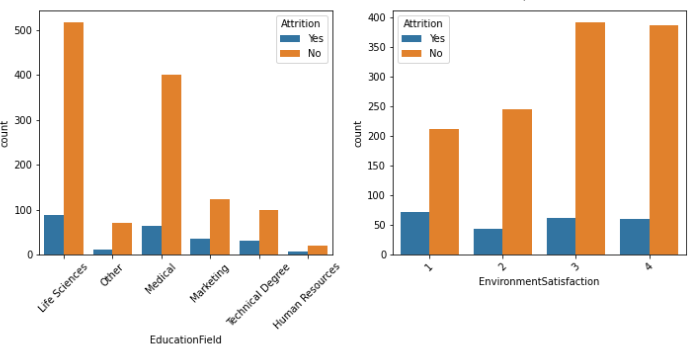
Firstly if we have classification problem we need to check whether we have **class imbalance** or not. Because if we have class imbalance our model, model will become biased to a particular class which has got more records.

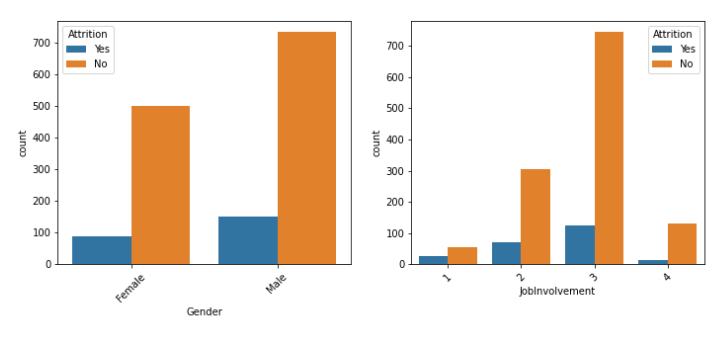


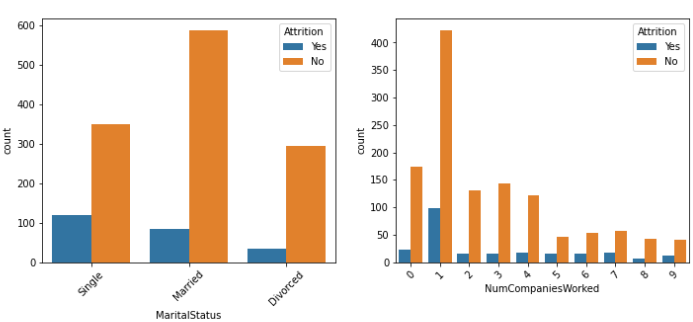
'No' class is dominating compared to 'Yes' class.

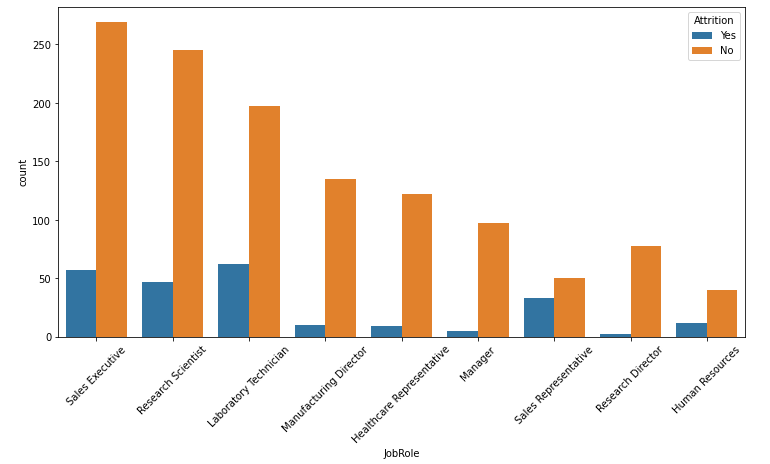


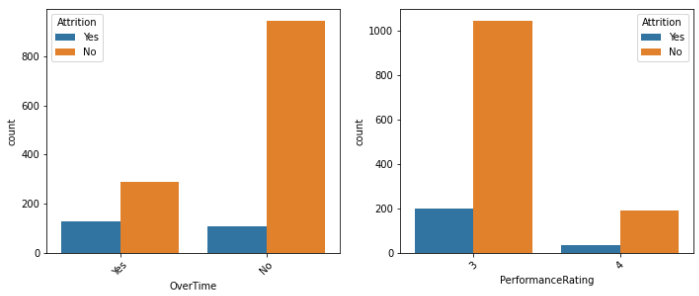


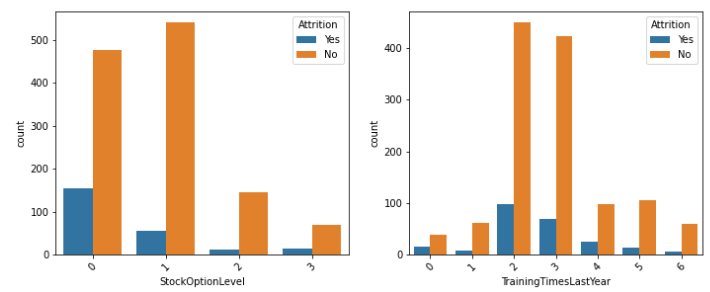






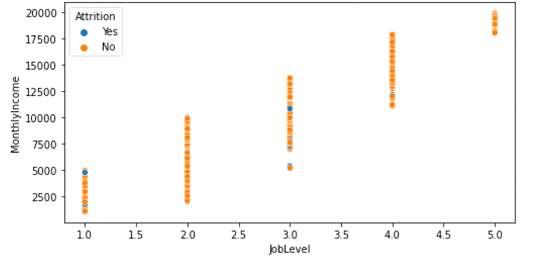






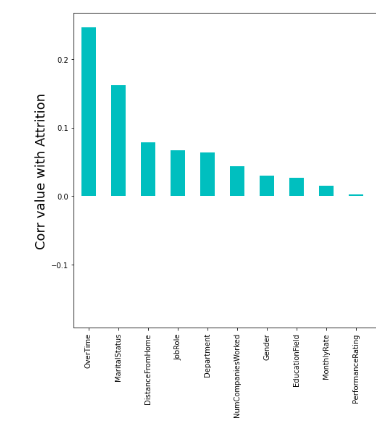
**EDA remarks** from above plot :

* Employees who travel rarely is more in count as compare with Frequent Business travel and non traveller.
* Also Attrition is high on the employees who travel rarely compared to that of others.
* In Research and Development department employee count is high with high count in Attrition when compared with other department.
* People travelling for shorter distance are more in count. And people travelling from longer distance are more likely to leave the company.
* More number of employees are from 'Life Sciences' field. Next is from Technical Degree. Attrition is more in count in Life Science dept as compared to other dept.
* Number of Male employees are higher than Female and the male count for Attrition is slightly higher than females.
* Job Involvement we can find its moderate as more counts in 2nd and 3rd level. Number of counts with 2nd and 3rd level seems to have high without Attrition but when compared with level 1 and 4 , it has higher attrition.
* The count is higher without attrition in 'Sales Executive' and 'Research scientist' as compared with other roles of job. But the attrition is higher with Laboratory technician and Sales executive. Research Director and Manager are the two roles who has the less attrition comparatively on other roles of jobs.
* Higher number of counts in Married peoples compared to Singles and Divorced. Attrition count is higher in Singles.
* Large number of employees in Low scale salary between 2500 to 5000$.
* We can see that 9 is the maximum number of companies worked by an employee.
* Number of count is high for the people not doing overtime.
* Maximum counts of no attrition are with 11,12,13,14 percentage . i.e with minimum salary hike percent.
* Number of counts in performance rating is less with 4 ratings and more number of counts in 3 rating.
* We can see that there is a wide range of working experience with employees (from 0 to 40). More number of counts are there in 10 years of experiences. And Attrition is high with more counts in 1 year experience.
* We can see that male has higher work life balancing as compared with female employees.
* Number of years working at present company is more in the range of 5 years and 1 year.
* Most counts of people in current role is 2 years and 7 years.
* Promotion are with higher counts of people with less than 2 years.
* There is a large count in people working with current managers for 2 year and 7 years. Very minimum number of years for long term working with same manager.

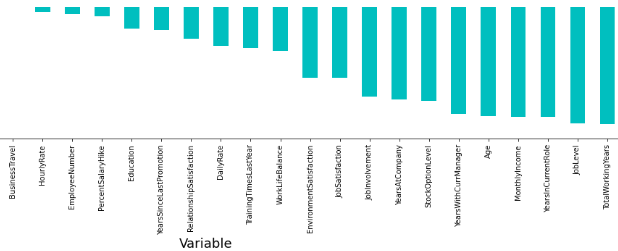


* From the above plot we can tell that Monthly income increase with increase with increase in Job level.

We should also see **correlation** of our dependent variable with independent variable and also the correlation with dependent variables with other independent variable.



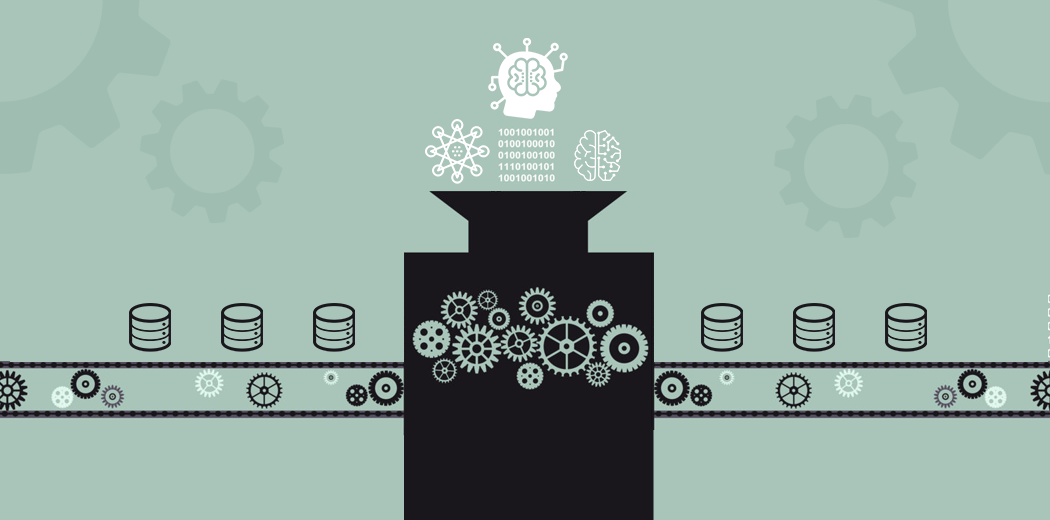
From the above graph we can see that which all variable has got positive correlation with our data set. Performance Rating has very low positive correlation compared with other variables.



From the above graph we can see that which all variable has got negative correlation with our data set. Hourly Rate has got low negative correlation and Total working Years has got high negative correlation with our dependent variable.

Apart from all this getting rid of **Multi colinearity**  is also very much important because most of the machine learning algorithm assume that there is no colinearity with each other among the dependent variable.

**Pre-processing Pipeline**



From EDA we have got to know that our data is semi-structured so we should make it structured before building the model.

We have come across some less useful variable, class imbalance, outlier presence, skewness in our data, huge scale difference between each variables. Lets Treat one by one

**Dropping the variables/columns:**

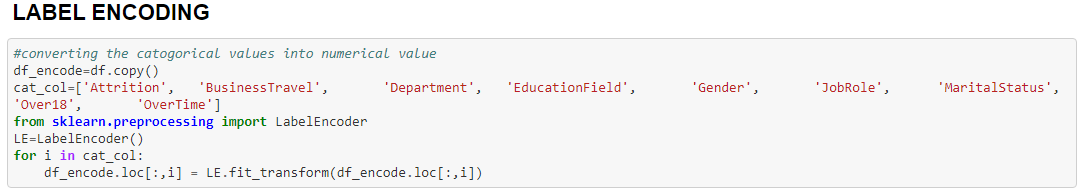
****

I have basically dropped 4 variables Employee Count, Employee Number, Standard Hours and Over 18.

Employee Count, Standard Hours and Over 18 is same for all the rows so the information from this to our model will be null so its better to get rid of such variables.

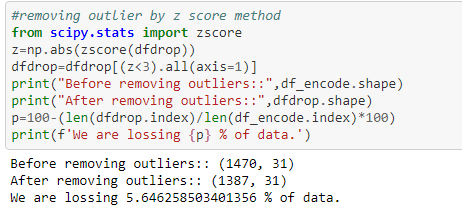
Employee Number is unique for everyone and it’s just like a serial number. So I have dropped that column also.

**Label Encoding:**



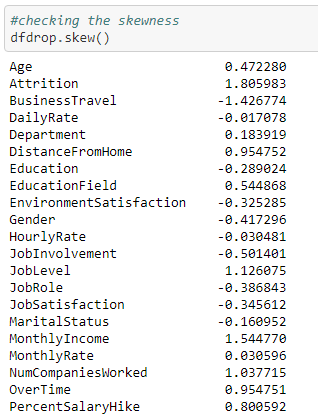
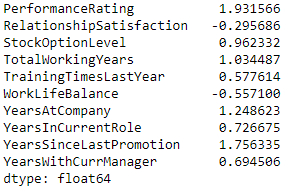
Our models cannot understand categorical values So its very much important to convert them in to numerical form. There are many methods to do that but I have used Label Encoding on 9 columns.

**Outlier Removal:**

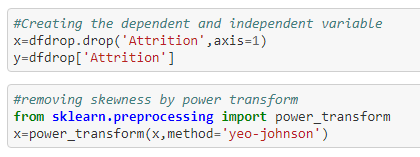


Outlier may affect the efficiency of model so I have removed it by using Z-score method. By this method we would be losing 5.6% of data this -means that we would be losing 83 rows which is less.

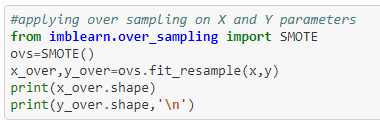
**Skewness Removal:**

** **

Above we can see how much skewness is present. This skewed data act as outlier which in turn decreases the efficiency of our model so I have used the power transform to remove the skewness and I have applies this method only on the independent variable.

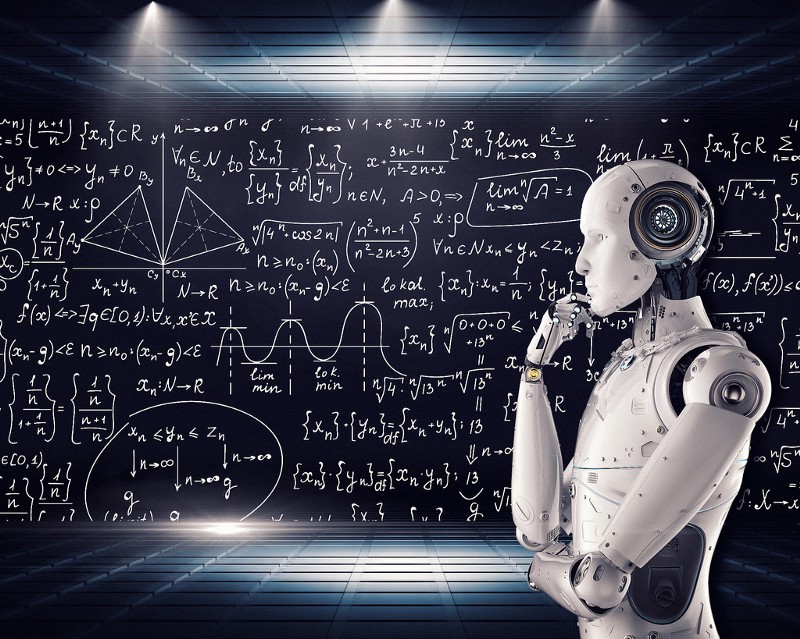
****

**Handling the class imbalance:**

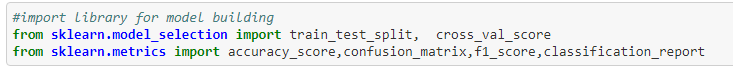
****

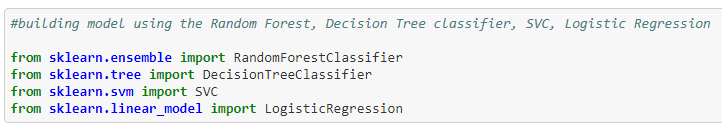
We have seen that in dependent variable ‘NO’ class is dominating compared to ‘YES’ class. So we should make it equal for that I have used Over sampling method.

**Building Machine Learning Models:**

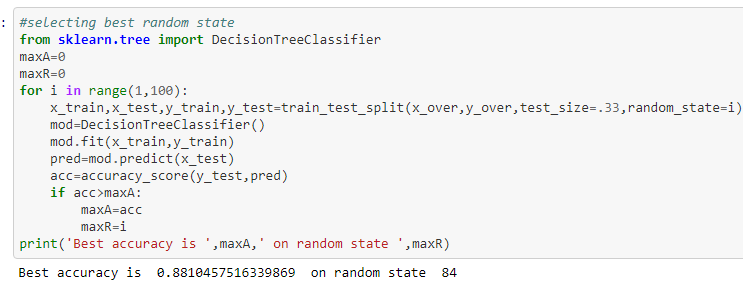


Let’s import the entire library that is needed to build our model





Now next step is to select the best random state



We have got 84 as our best random state and we will be using this random state itself in our model building.

## I will be using Random Forest Classifier, Decision Tree classifier, SVC, Logistic Regression models on my data set

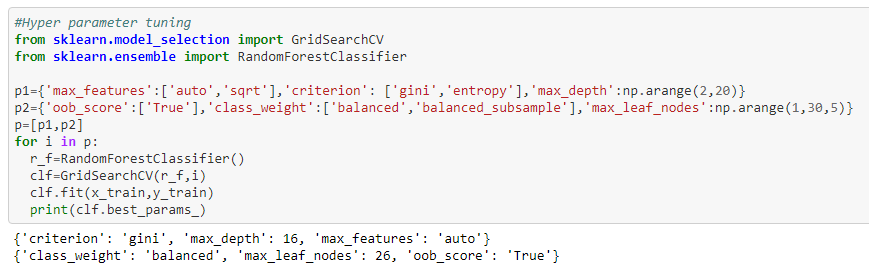
## 

In above diagram we can see the Accuracy score and Cross validation score of all the models that we have build. From that above table we can say that Random Forest Classifier is our best model with **93%** of accuracy with only **0.2%** deviation.

**Hyper Parameter Tuning on Random Forest Classifier:**

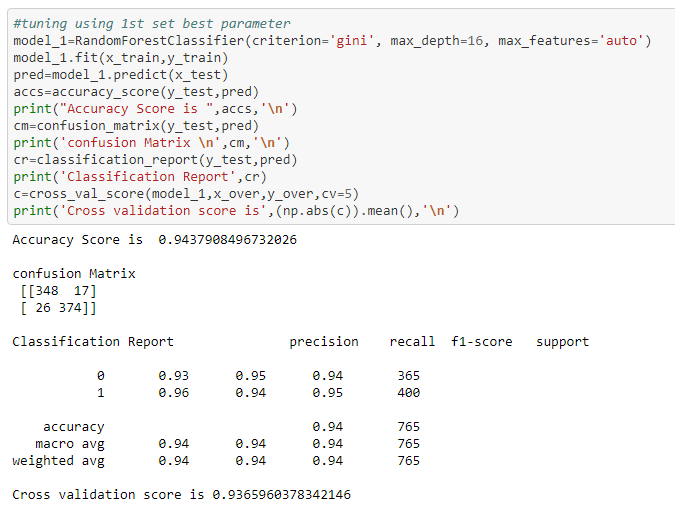
****

Let’s see whether we will be able to improve on accuracy by tuning the parameter of Random Forest Classifier. For this first we should select the best parameter.

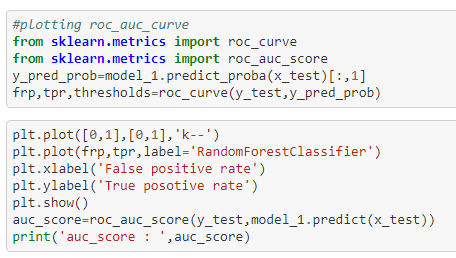


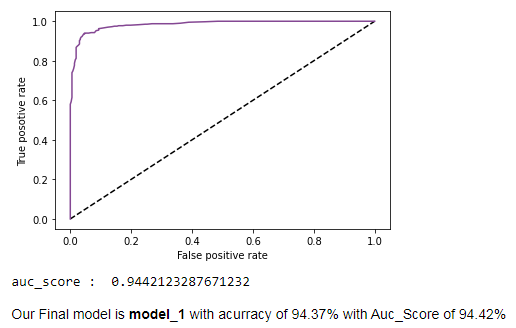
From above we can see that we have got 2 sets of best parameter and let’s build the model using this best parameter.

After building the model we observed that with 1st set of best parameter we were able to improve our accuracy to **94%** which is great so we will be deploying this model.



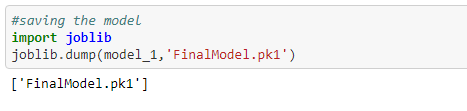
**Let’s plot roc\_auc curve:**

****

****

Our Final model is **model\_1** with accuracy of 94.37% with Auc\_Score of 94.42%

**Saving the model:**

****

**Concluding Remarks:**

* We were able to analyze employee attrition using Machine Learning.
* We were able to identify what all factor influences the attrition of the employees.
* We were able to get the probability of an employee leaving the company
* We were able to see which all area needed to be look after to restore the employee.

**Link to the solution of this problem:**

<https://github.com/rs6044922/DataTrained/blob/main/HR%20Analytics%20Project-%20Understanding%20the%20Attrition%20in%20HR/HR_Analytics_Project.ipynb>

**Link to my LinkedIn profile:** <https://www.linkedin.com/in/ravikumar-s-a874081b8>