HR Analytics Project- Understanding the Attrition in HR

**Problem Statement:**

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.**

**Attrition in HR**

Attrition in human resources refers to the gradual loss of employees overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

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.

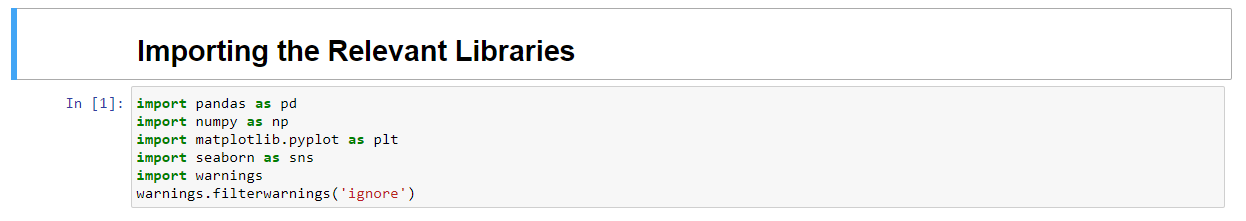
**Attrition affecting Companies**

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.

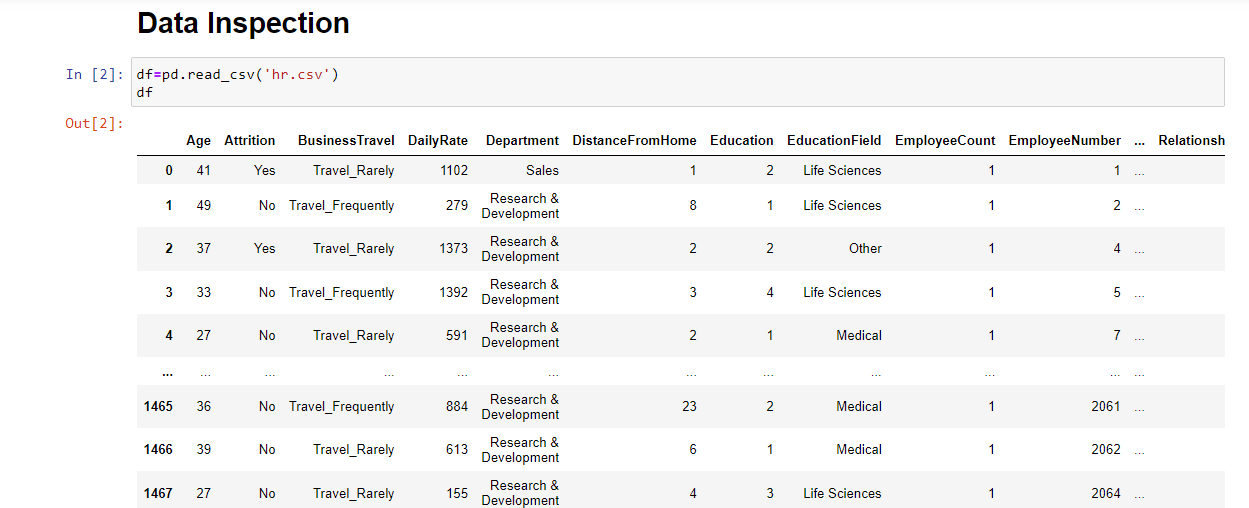
Now let’s start our project you can download the dataset which I’m using from the following link:-

https://github.com/SumeetSingh-R/Projects/blob/main/Dt%20Evaluation%20project/hr.csv

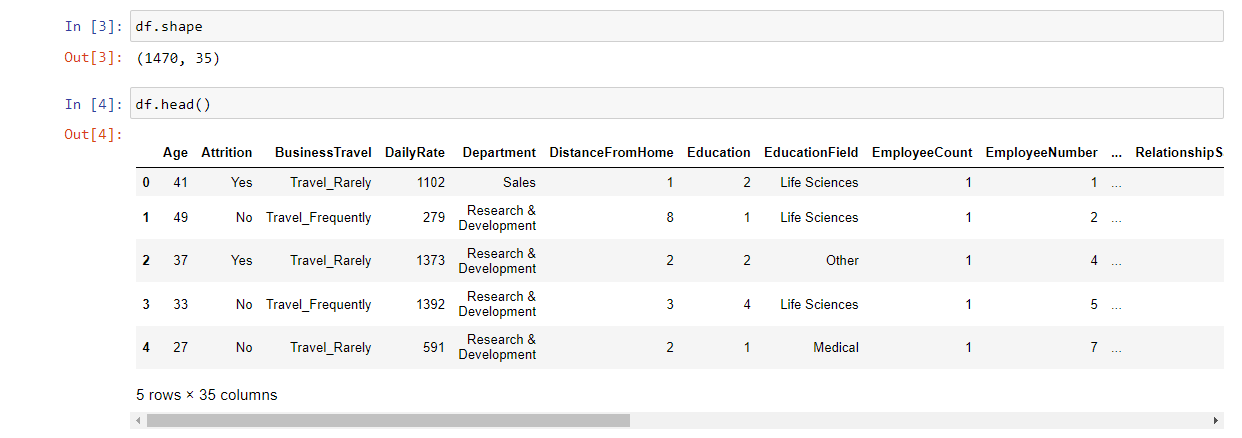
1. We will start by importing the relevant file



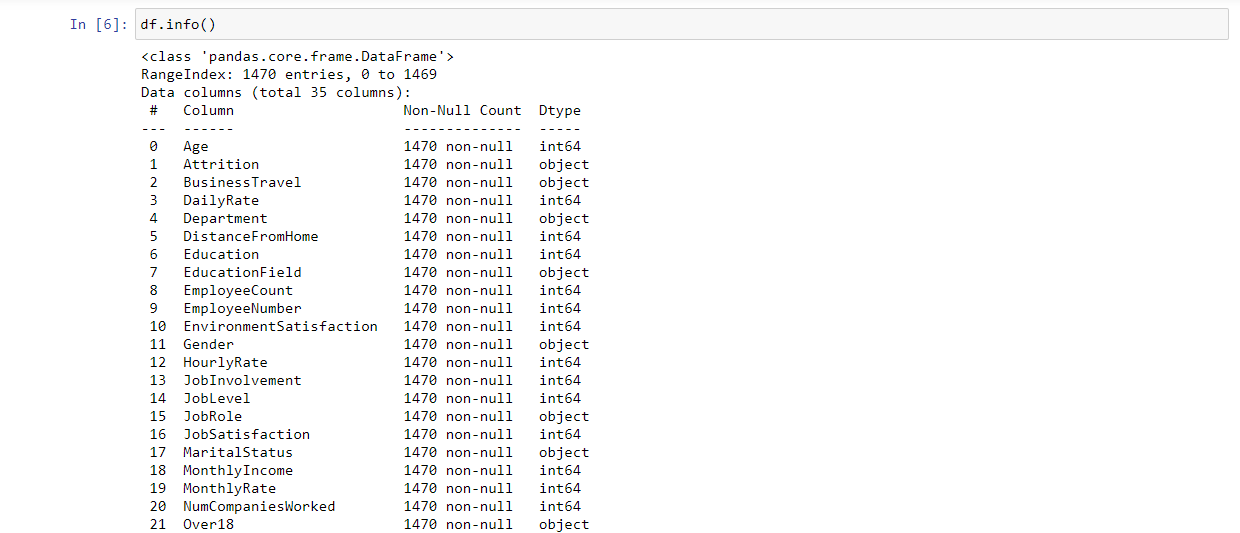
1. After we have imported our libraries let us import our dataset and don some analysis.



As we can see we were having our data in csv file so we used pandas to read our csv file by pd.read\_csv and name the file.

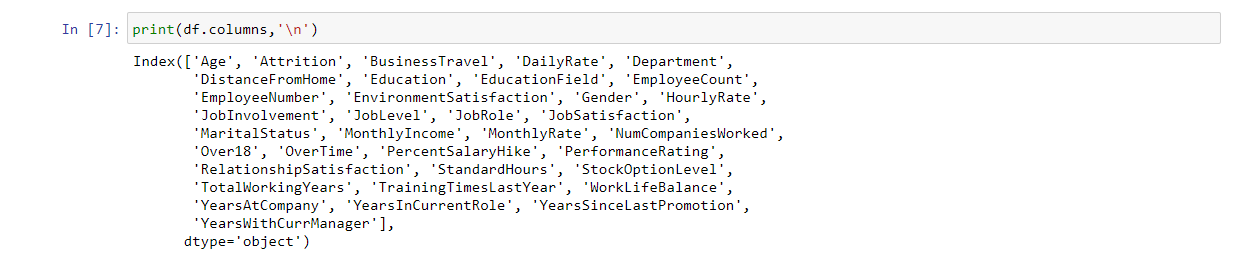


There are total of 1470 rows and 35 columns In our dataset & we are also checking the head od the dataset which gives an info about the data.



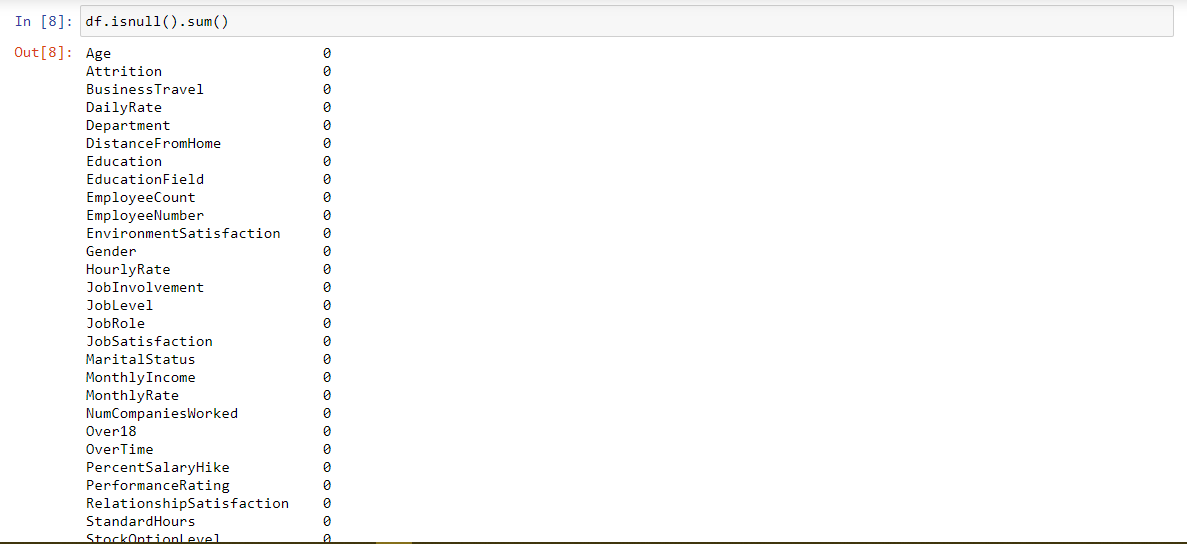
After that we are checking the data’s info which will let us know about the data type & no nulls.

As a next step we are checking the columns of our data so we can come to know what are the attributes we are dealing with and how they are helping us in prediction.



As we want to predict the attrition of the employees we will be making attrition column in our dataset as our target variable.

Checking null values in our dataset is essential, because most of the machine learning algorithms throw error while dealing with null value data except some algorithms like k-nearest neighbours (KNN) so we are checking for null values below.

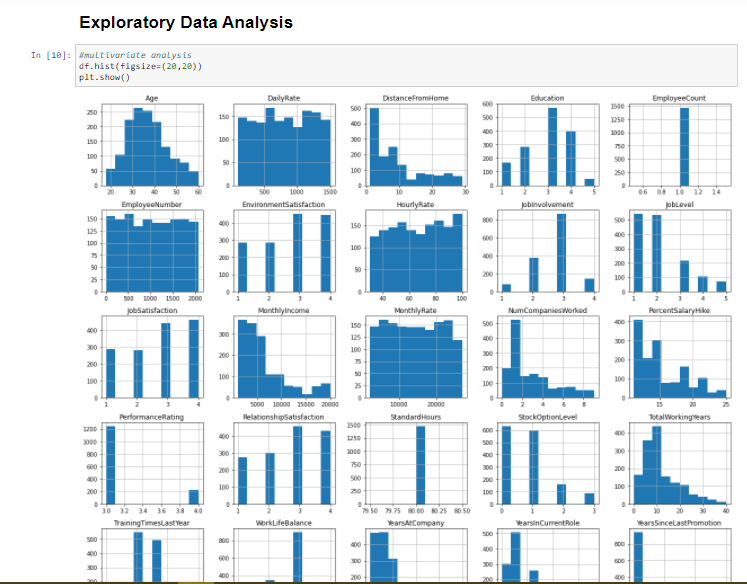


In our data set we don’t have any null values so we don’t need any treatment for it.

1. Now we have an basic idea of our data, that what are their datatype, is there any null values, what are their columns, so now let’s us start with Exploratory Data Analysis (EDA).

Trust me analysis can become so easy with EDA because it is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

* Maximize insight into a data set;
* Uncover underlying structure;
* Extract important variables;
* Detect outliners and anomalies;
* Test underlying assumptions;
* Develop parsimonious models;
* Determine optimal factor setting;



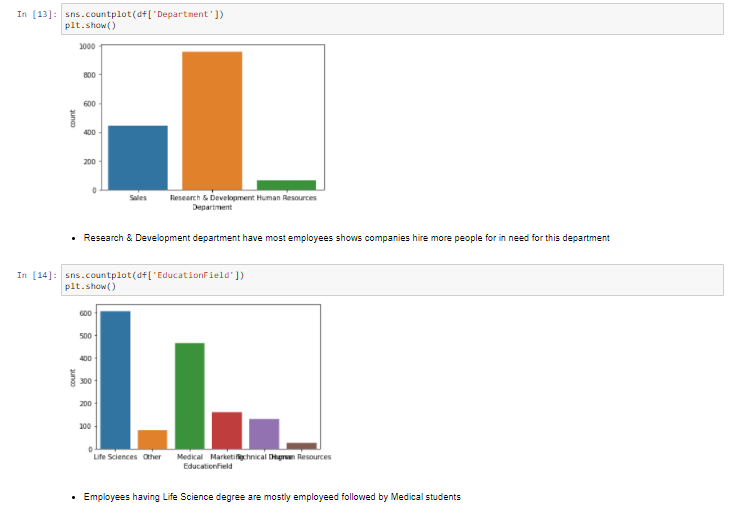
Above we have plotted a histogram diagram of all the data you can try that yourselves.

It’s always good to write the observations at the same time, so from this diagram we can get lot of data insights that are as follows below.

* Employees are between the age of 25-45.
* Maximum employees have spent between the year 0-5 with current manager.
* Major employees are spending between 2-3 years in current role.
* Employees tends to spend 10 years in one company.
* Most of the employees stay near to they work place.



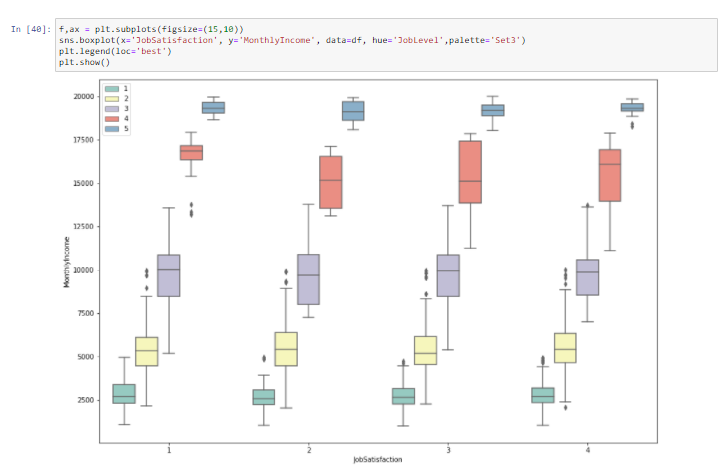
Above we are plotting count plot which gives us a really good idea to know the ration of anything we can easily come to know the difference b/w variables. In the first diagram we have plotted attrition and we can easily come to know attrition rate in our dataset in very low most of the people leave the organisation instead of staying. In the Second diagram we have plotted Business travel column which shows how frequently employees are travelling. So after plotting we come to know that majority of the employees are rarely travelling.



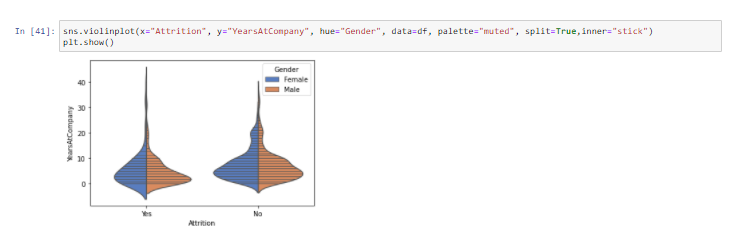
Here we have plotted two other columns which is department and education field so from that we come to know that most of the employees are working in Research & Development department followed by sales department, and in Education Field most of the employees have studied life science followed by medical.



To get more precise insights let us do some detail analysis so above in the first bar chart we are plotting attrition rate vs monthly income and having gender as our hue. So, we come to know that employees whose salary are less than 45K are not leaving the organization and male are tending to be in the organization more than females. And employees above the salary of 45k are leaving the organization and in that female tends to leave more than males.



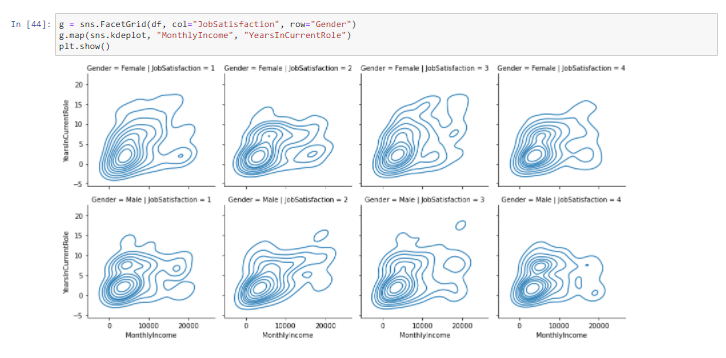
We can clearly observe the difference in monthly income across different job levels. The difference in the monthly income is observed by the median value for different job satisfaction level.



Here, we have plotted the number of years spent in an organization based on gender. The middle dashed line shows the median. The lines above and below the median show the interquartile range. The denser part shows the maximum population falls under that range and thinner part shows the lesser population



Above factor plot shows that monthly income plays an important role in retaining the employees in an organization. It can be observed across job levels and different age groups.



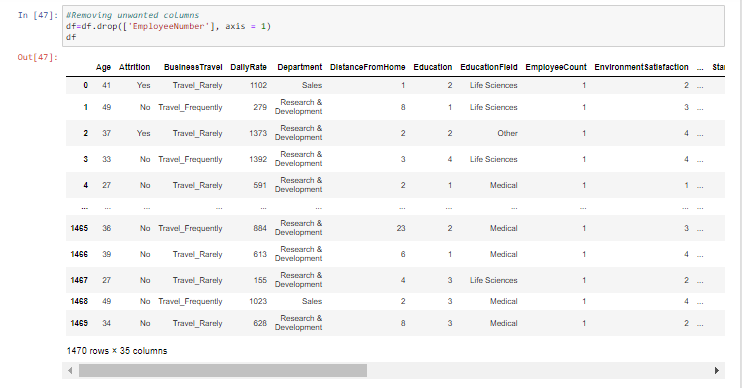
We have created kernel density estimation plot. It displays the density distribution of two continuous variables (namely, Monthly income and years in current role). We have created facets according to different job satisfaction levels and gender.



Pairwise plots between continuous variables show the relationship between them. For example. observing the relationship between Age and Monthly Income, we can find that with age, monthly income has increased but the increase is not similar for both groups (Attrition and Retention).

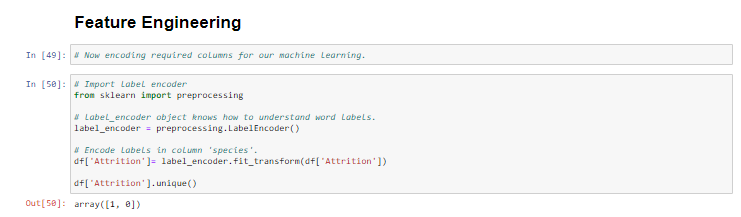


The above plot does not convey much of any relationship between variables across gender. This shows that hourly rate, daily rate, monthly rate and percent salary hike is same for both female and male employees.

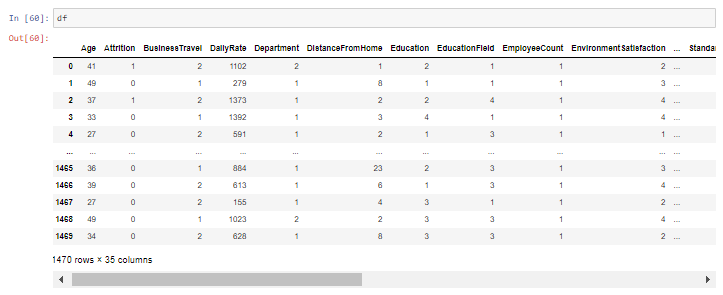


We are dropping Employee Number which will not effect our machine learning model.

1. We came to know lot of things about our data and the information from it so for the next step before making our model, we need data which is in numerical format so we can train our model on that. For that we will used Label encoder which will convert our string data into numerical.

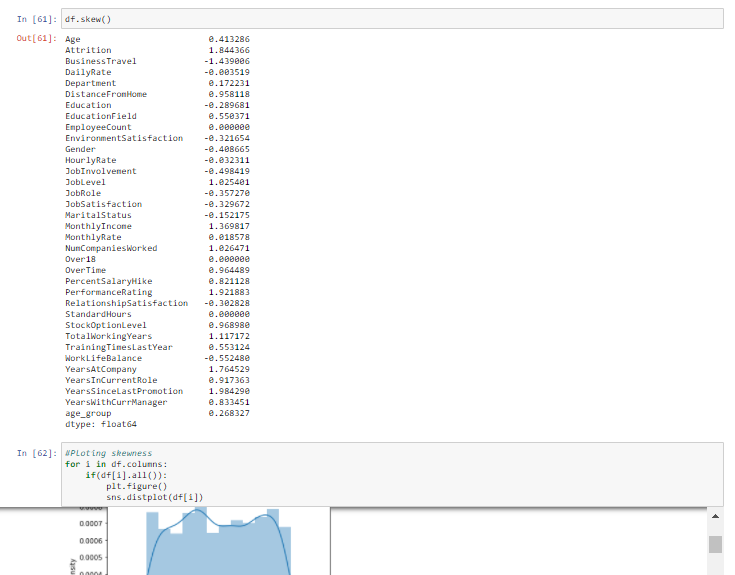


If you remember our attrition column was having data as yes and no so we used label encoder to convert it into numerical value so we got 1,0 as the output. Similarly let us convert all the other character data as well and check our data frame now.

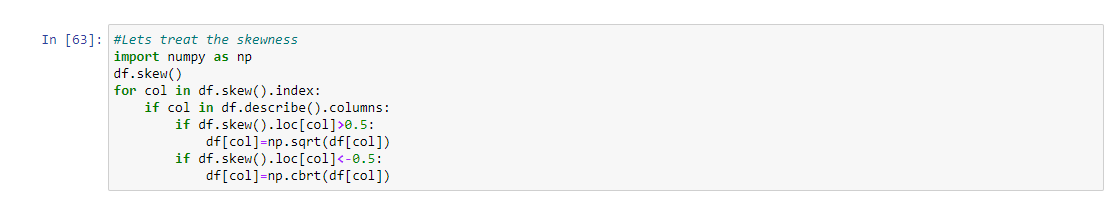


As we can clearly see that all of the data is now in numerical format and it is also ready for training our model.

Now let us check skewness of the data.

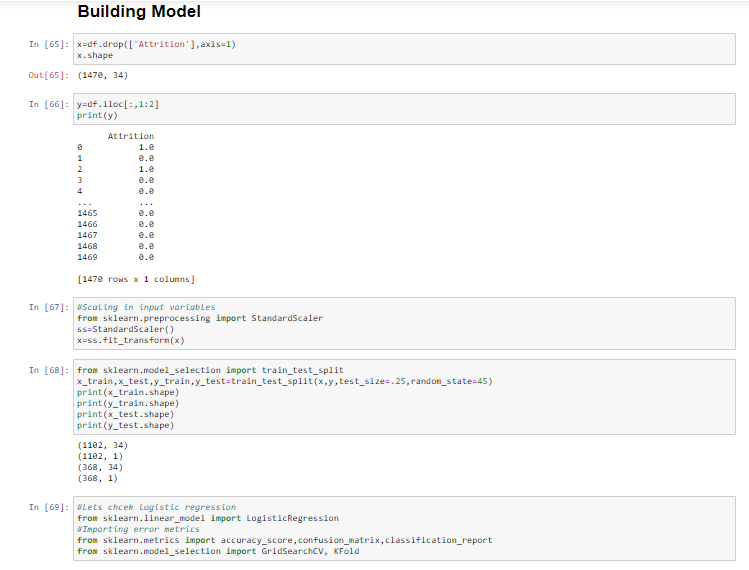


After plotting the skewness I came to know that there are skewness present in our data so to treat that we can use log transformation, square root transformation, Box-Cox transformation. I’ll be using log transformation because that gave the best result when I was trying to treat this skewness for this dataset.



Using this method are treating the skewed data and now our data is almost ready to be used for training and testing in the model.

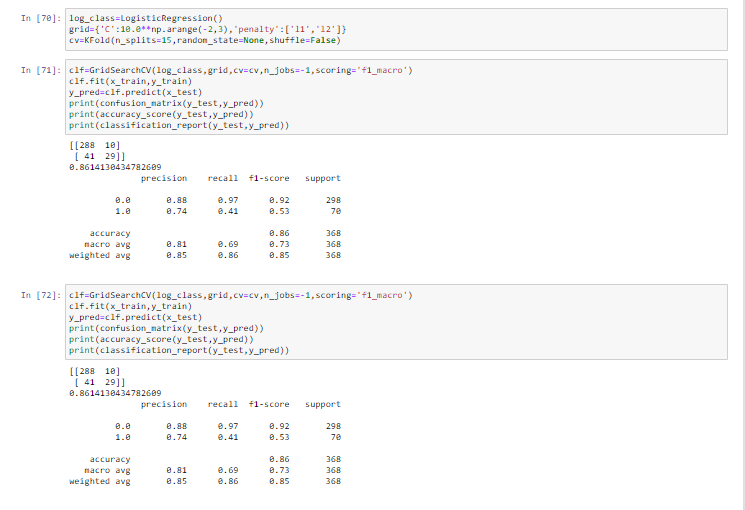
1. Now we have the require data for the machine learning let us start building our model.

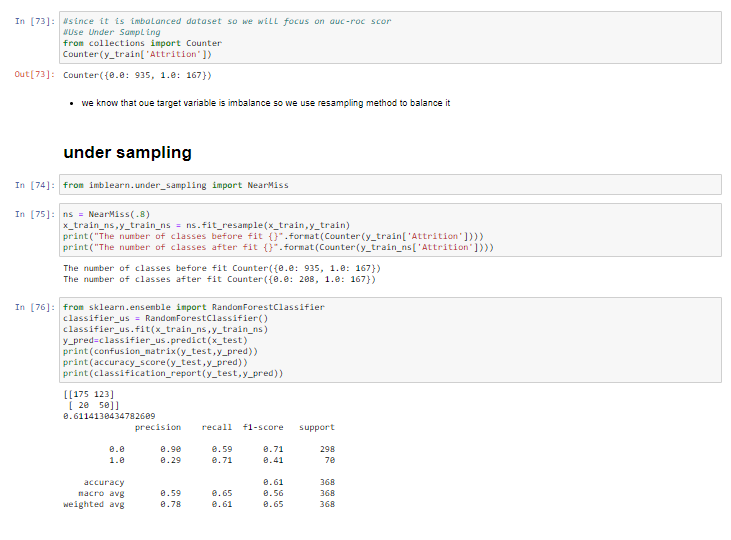


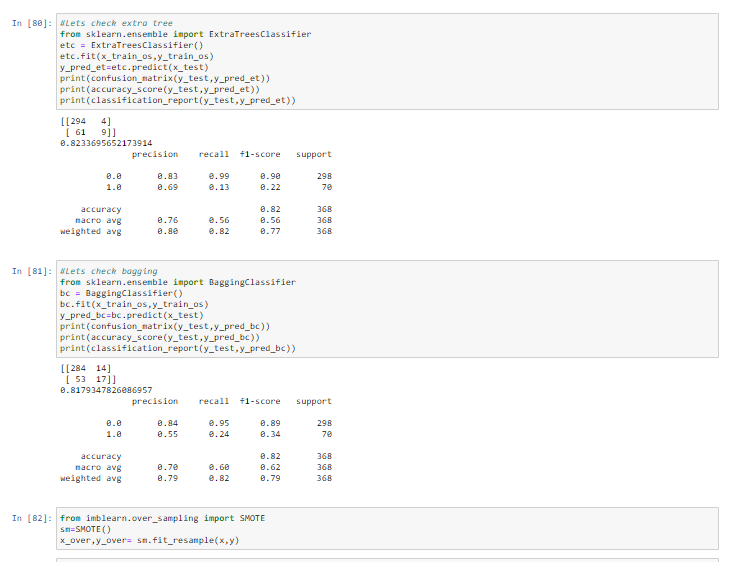
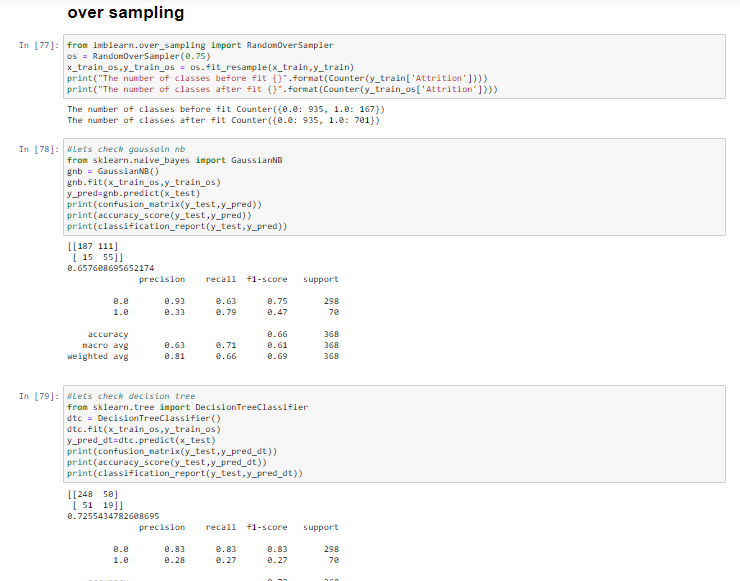
First step what I did was to define x and y variables. For x I’ve remove attrition column and all other column will be in x which will be our data for training and testing. Then for y we are having attrition rate which is our target and which we need to predict. Once we have our data and target variable decided let us scale our data we are using standard scaler for this.

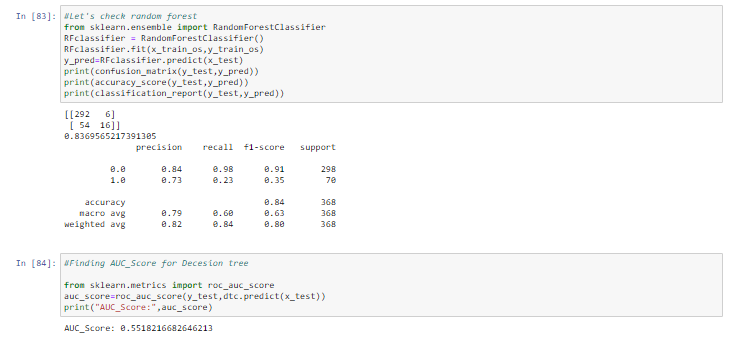
Let us split our data into training and testing I’m dividing 75% in training and we will have 25% in testing.

As this is an supervised learning problem I’m using Logistic regression, Random forest classifier, Decision tree classifier, Bagging classifier, extra tree classifier to predict.

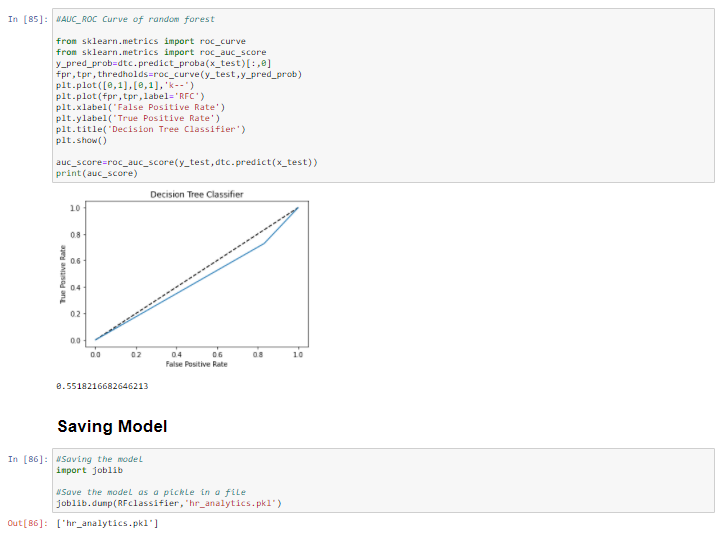








The best accuracy which we are getting is from Decision tree classifier so we are using that for ROC curve (receiver operating characteristic curve) to show the performance of a classification model at all classification thresholds.



Finally we are saving the model using joblib as that it can be used in production environment.

Thankyou so much for Reading till now.

Let’s connect

[Linkedin](https://www.linkedin.com/in/sumeet-singh-r/) [Github](https://github.com/SumeetSingh-R/Projects) **[Tableau](https://public.tableau.com/profile/sumeet.singh4497#!/)