

Example of an end-to-end machine learning project in Data Science for beginners.

I am going to write about complete end-to-end HR Analytics project which would truly a guidance for all new comers who is aspirant of data science . when I accepted to go with data science and started learning about it I was in huge dilemma what to do, how to learn , how grab the knowledge, these are utmost common question that come to the everyone’s mind first who is willing to proceed their career with data science but since there so many helping blogs already written on internet by so many distinguished expert that helped me more than my expectation.I know there lots of blogs already there to help but in my project I will emphasized over some different points. Maybe,perhaps , my blog that I am going to write will help atleast a bit for all those who wants to be a data scientist .

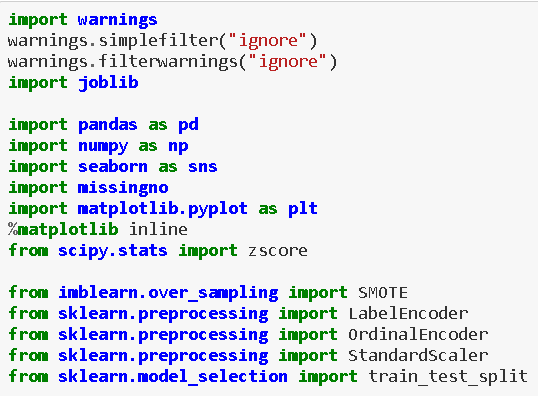
So, please allow me to explain the agenda for this blog post. In this article, I have put down all the techniques in the form of sub-topics that I will be explaining one by one. And those pointers are as follows:  
  
1.      Problem Definition  
2.      Data Analysis  
3.      EDA  
4.      Pre-processing Data  
5.      Building Machine Learning Models  
6.     Concluding Remarks

Let’s start with the problem definition or a short introduction on the project that I have chosen to elaborate and why it was made in the first place.

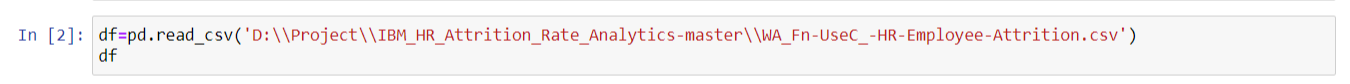
**1.Problem Defination**

The Project I have opted to write over is HR Analytics Project that is basically fictional Project and It was created by Data Scientist In IBM. You can find the whole dataset in kaggle .Basically there so many companies who hired so many new employee every year and organise training program inside the company to the new comers and also for old employee to enhance the efficiency. But my concern through this project is where does HR fit in all of these.

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.  


Attrition in human resources refers to the gradual loss of employees’ overtime. In general, relatively high attrition is problematic for any company. 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 analysing 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 is a major problem since high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork and new hire trainings are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits an organization from increasing its collective knowledge base and experience over time. Errors and issues are more likely if you constantly have new workers too.  
  
Therefore the major goal of this project is to identify the “Attrition” rate as a simple Yes or a No tag making this to be a classification problem!  


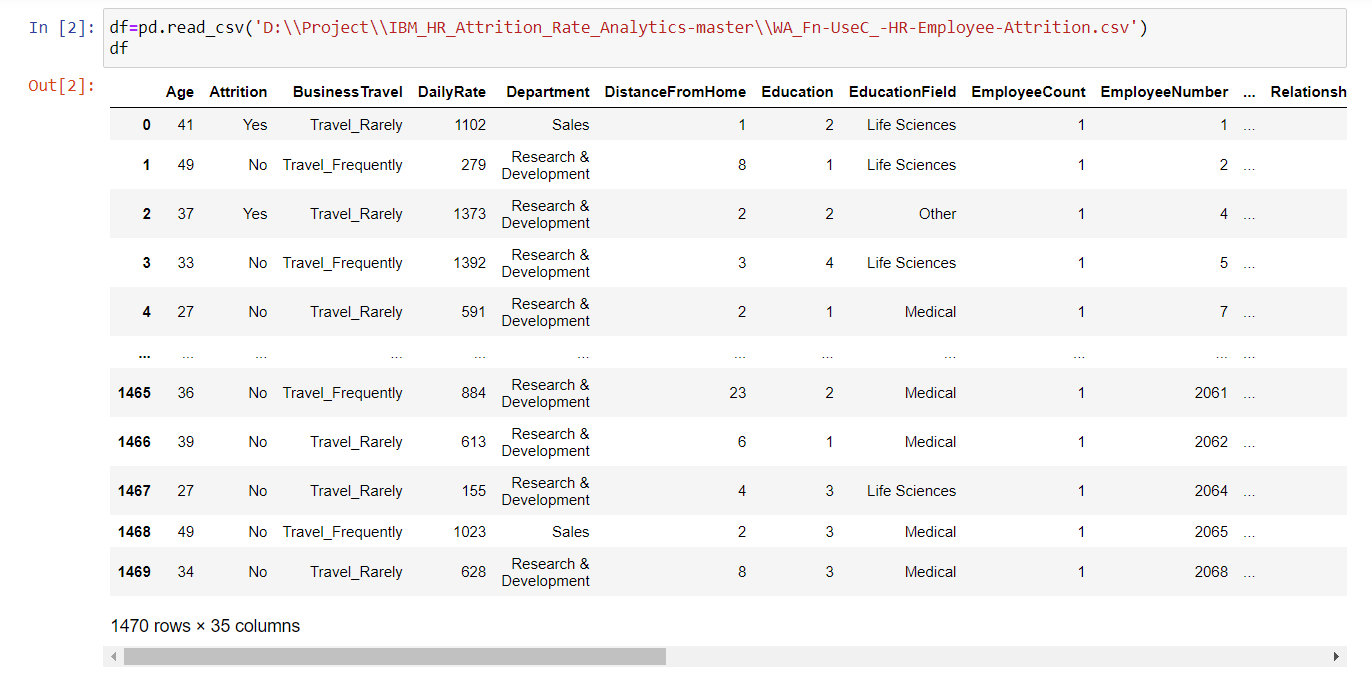
First of all I have imported all the necessary libraries or tools that is sure going to come into help of building best machine learning model. I have to import whole the dataset into single jupytor notebook that we can apply upon various method and machine learning Algorithm.



From the above read\_csv method we have imported entire dataset into jupyter notebook and stored into single variable name as df.

**2.Data Analysis-**

In this section of our project we will look at every columns of our data set and we will try to figure out which column is truly related to solve of problem statement or not .



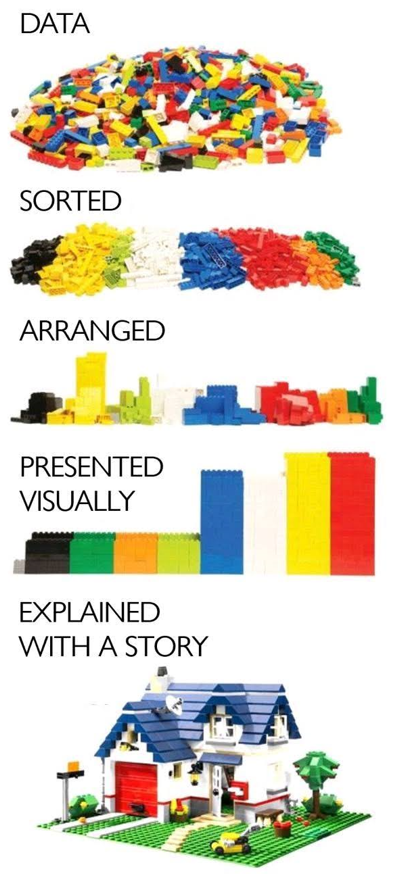
As we have dataset having 1470 rows and 35 columns below .in this 35 columns one is our target means Attrition and rest are features.

**3.EDA**

Without any hesitation we can say that EDA, that is Exploratory Data Analysis is heart for building of machine learning model . This is the utmost important part that every dataset must be go through it . EDA is process use various of produre or method to make our dataset into appropriate format so we can achieve our real target. In this EDA we do entire dataset analysis via using various using tools and python libraries.

From the below picture, i can explain:

1. For any sort of Machine learning Model data is the most important thing that you must have .There are so many way to collect the data. I will not go into depth here but I can say primary and secondary there two are the main sources of data collection
2. Then I sorted the entire data as per their features and Arrange into some format
3. By the different means of visualization technique, I have visualized the data and find a way to explain the whole story.



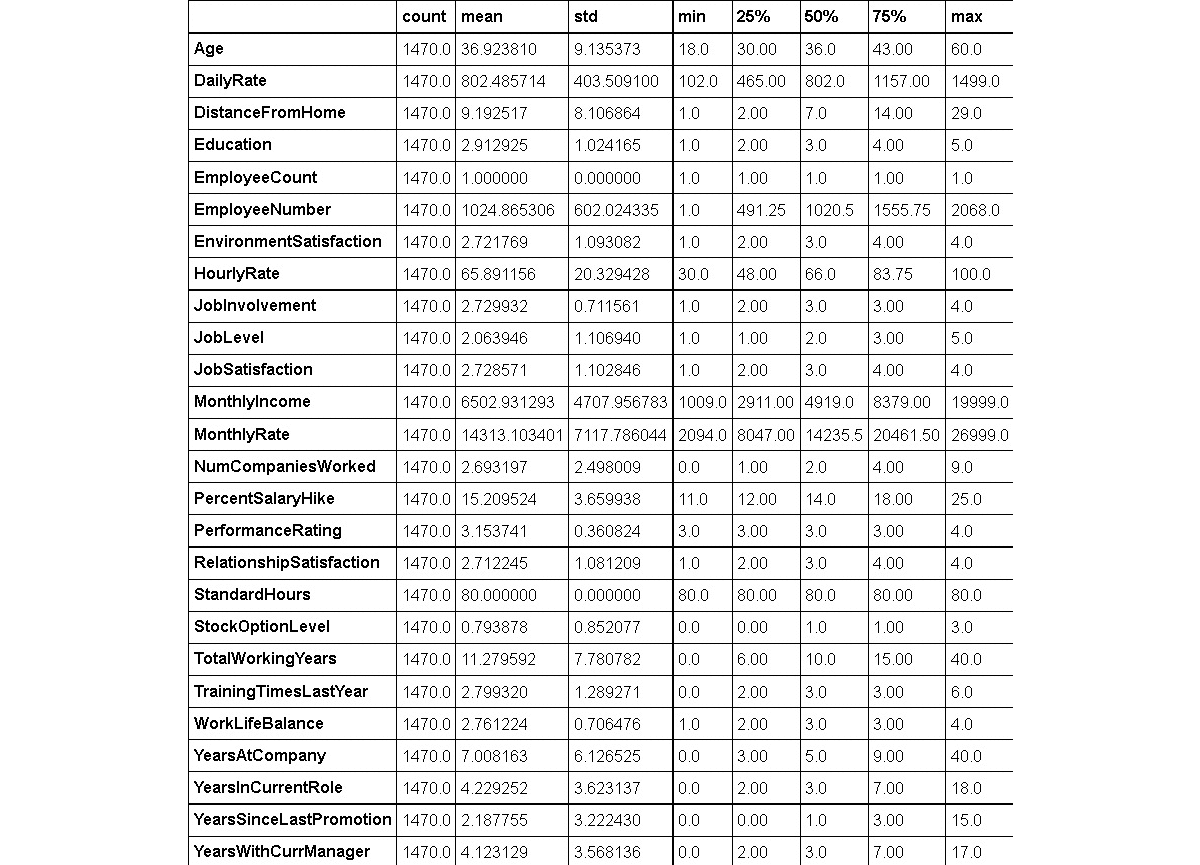
 This isnull() method with sum function allow us to figure out the missing value present in our dataset .



Now,fortunately we dont have any null value present in dataset

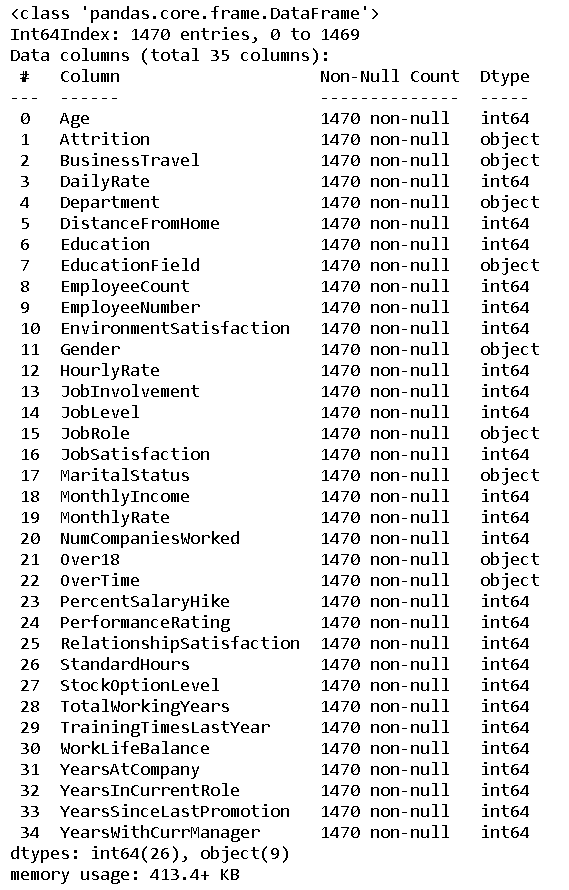


As using to this method to know the type of data and we can say that most of our data are in int form but our target is in object form.

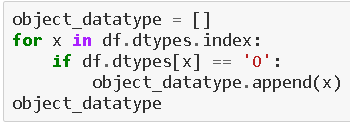


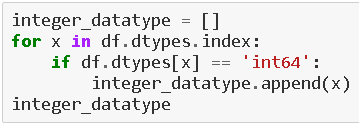
from the above describe transpose method it is displaying that there is no missing data as every columns have same count (1470) And maybe columns like Monthlyincome, Totalworkingyears,YearsAtCompany,YearsinCurrentRole,YearSinceLastPromotion,YearWithCurrManager has huge difference in their 75 % and 100% so they must have some amount of outlier that is need to be treated.





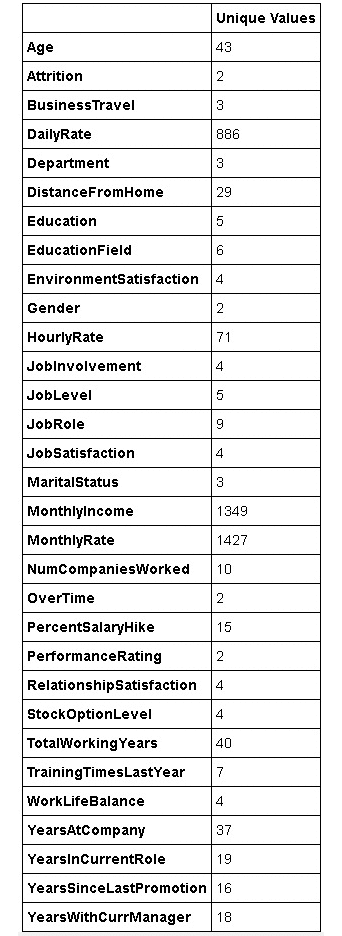
This df.info () method gave us all information regarding of types of our dataset.



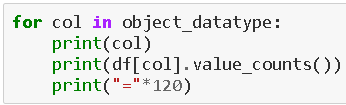


I have seprated our columns as per their type they are having. Means I have created two types of list namely object datatypes and integer datatype. I have kept all object type data into object datatype list and all numerical type data into integer datatype.





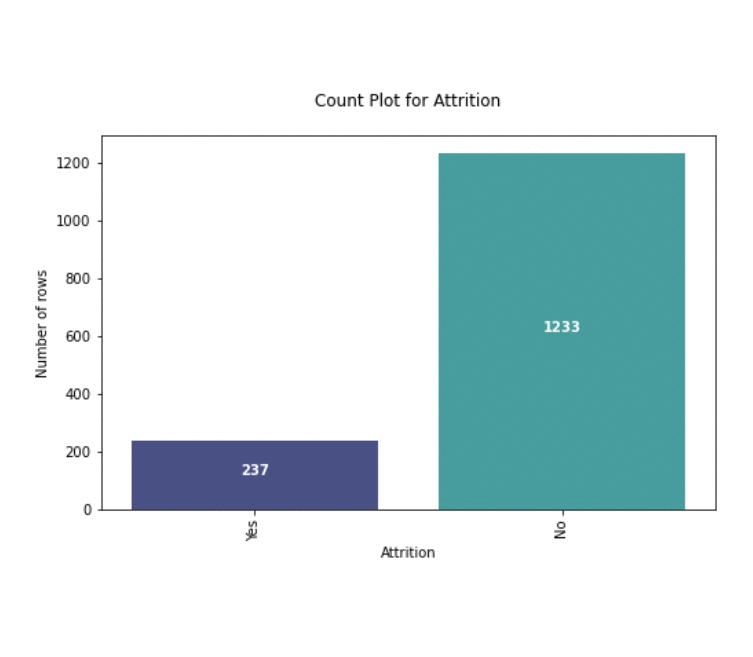
This df.nunique method has given us all the unique value present into each of columns of our dataset.



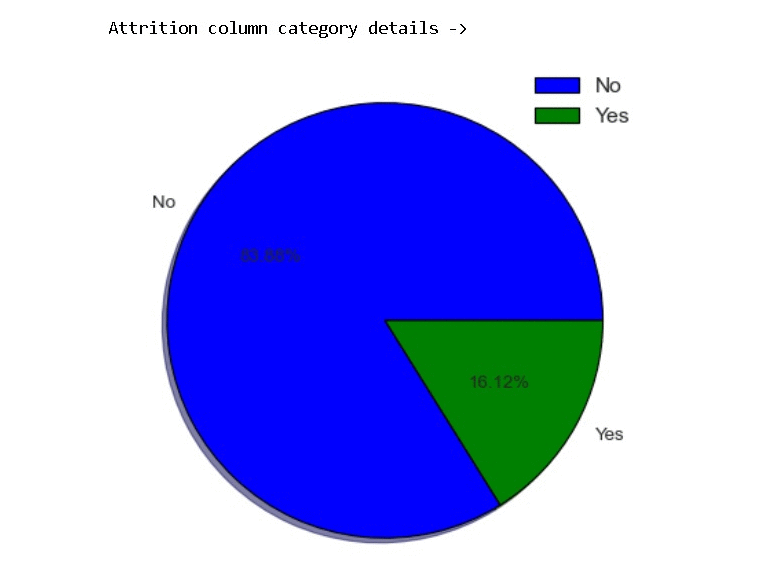
Above the method I have used to find the uniqueness present in our dataset but this method has some limit means it does not give the information of uniqueness for categorical data. That method was only useful for the numerical data.

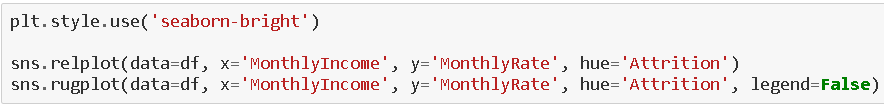
So, avoiding that limitation I have used loop method for the object data type object.

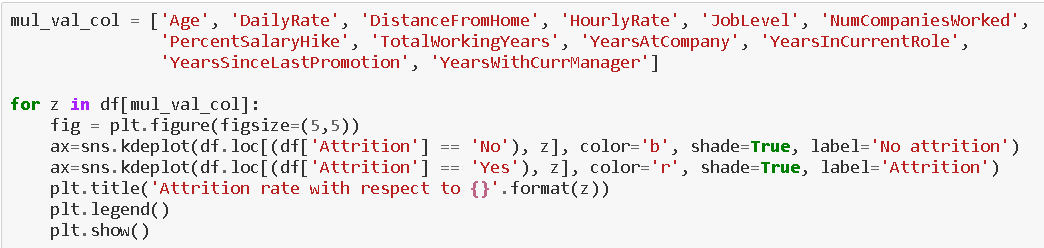


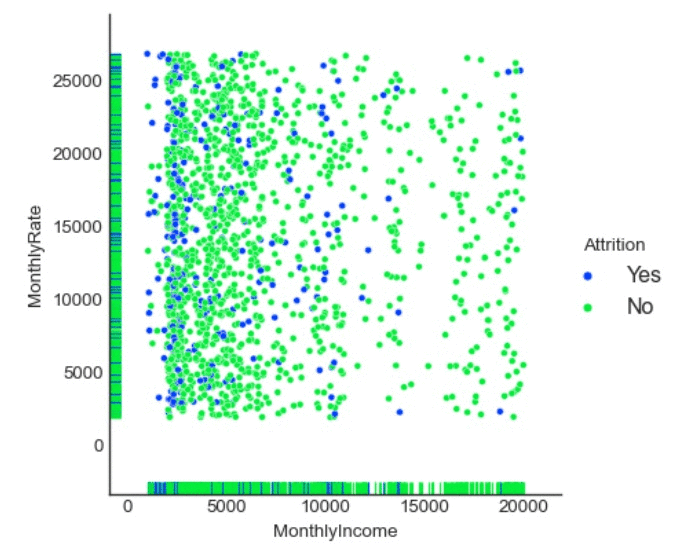


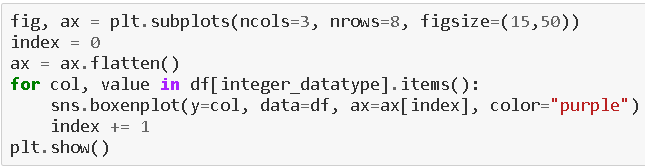


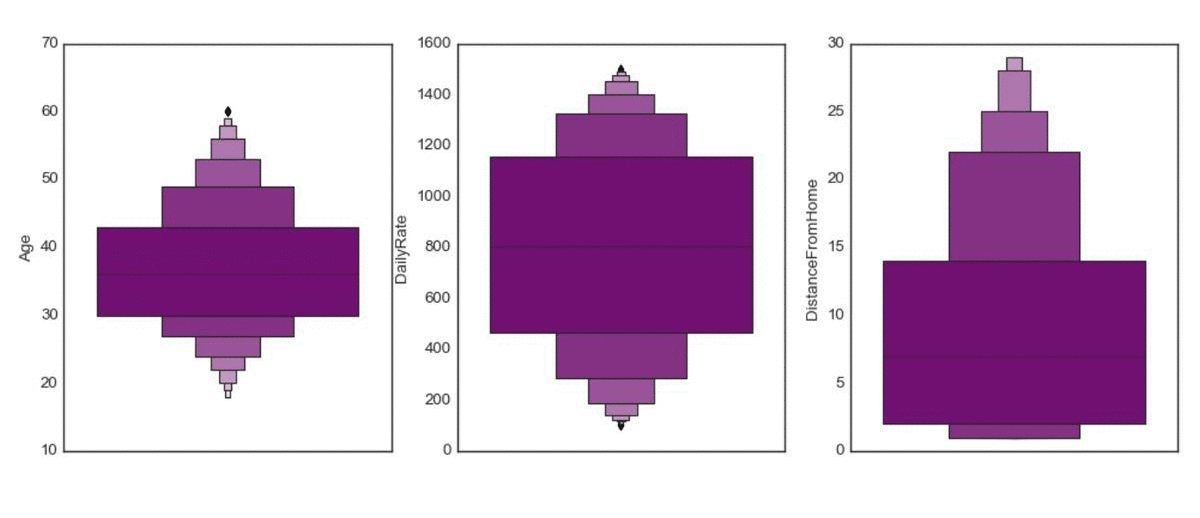


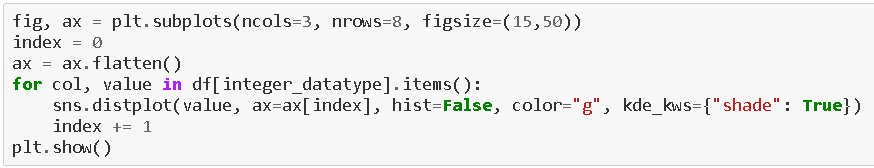


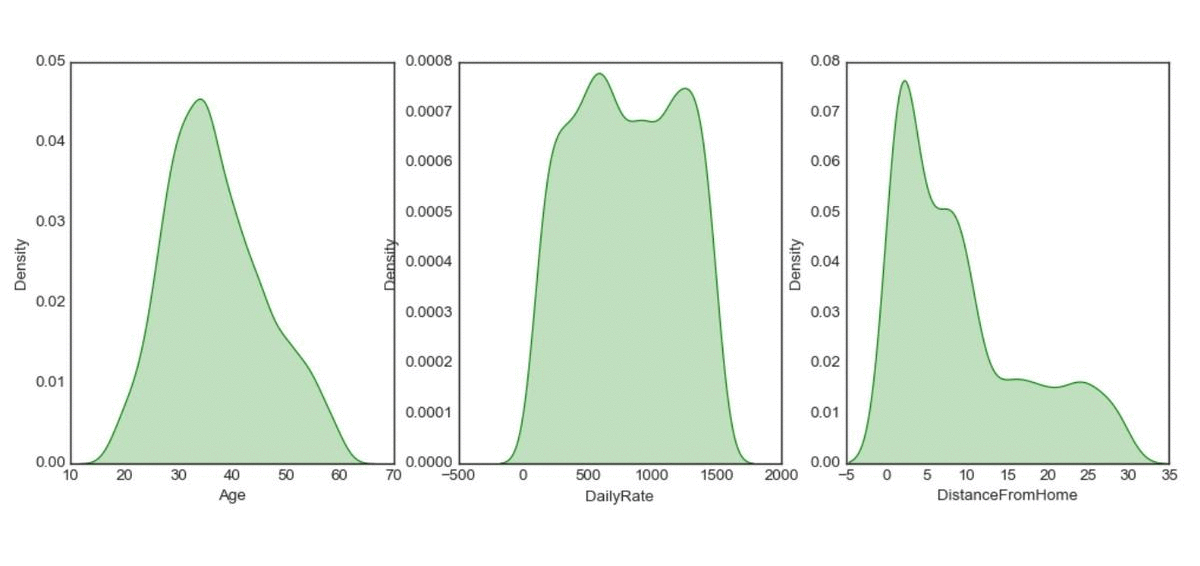








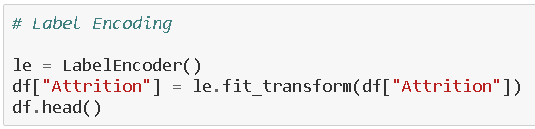


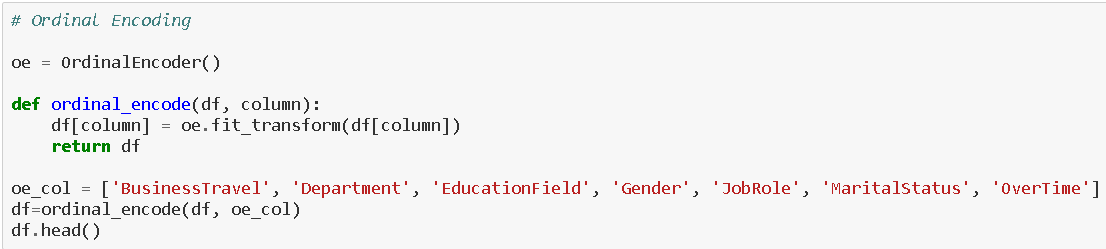
  
You can see that with the help of above codes and getting the outputs I was able to take a look at all the column values/counts, the boxen plots gave me a view on the presence of outliers and the distribution plots showed me the skewness information that will needed to be treated. These are like the challenges that will need to be deal with before I even think of building my Classification Machine Learning models.

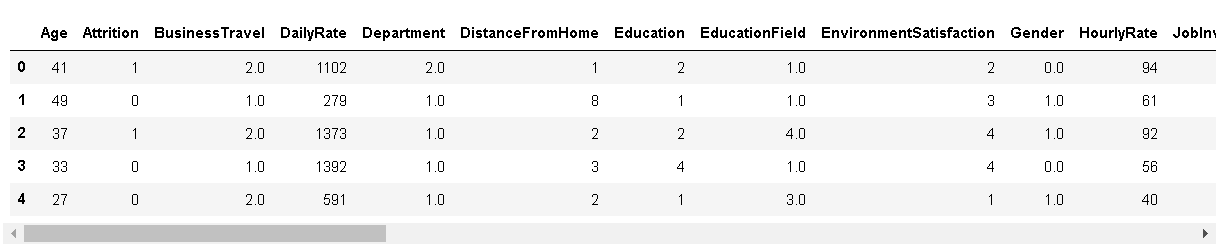
1. **Pre-Processing**

Data processing is a processes of preparing the raw data and making it suitable for it a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating machine learning project it is note always a case that become across the clean and formatted data and while doing any operation with data it is a mandatory to clean it and put in formatted way. So, for this we use data pre-processing task.

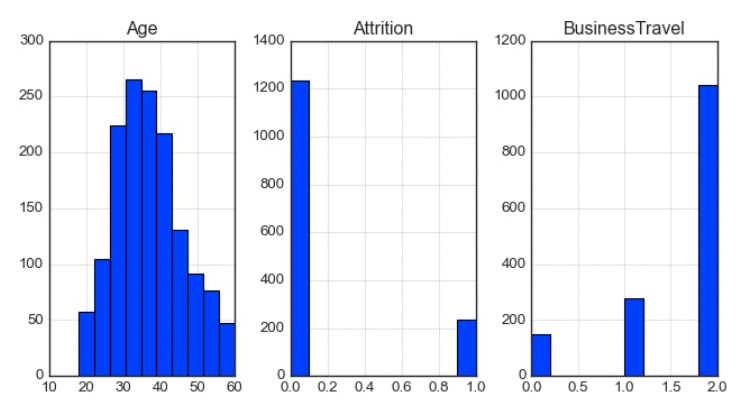






Here you can see all categorical data has been encoded.

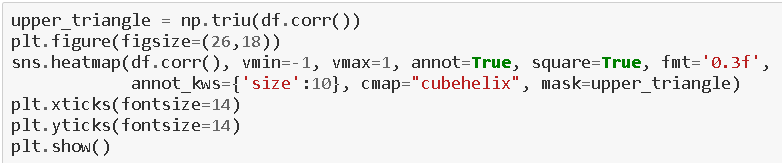


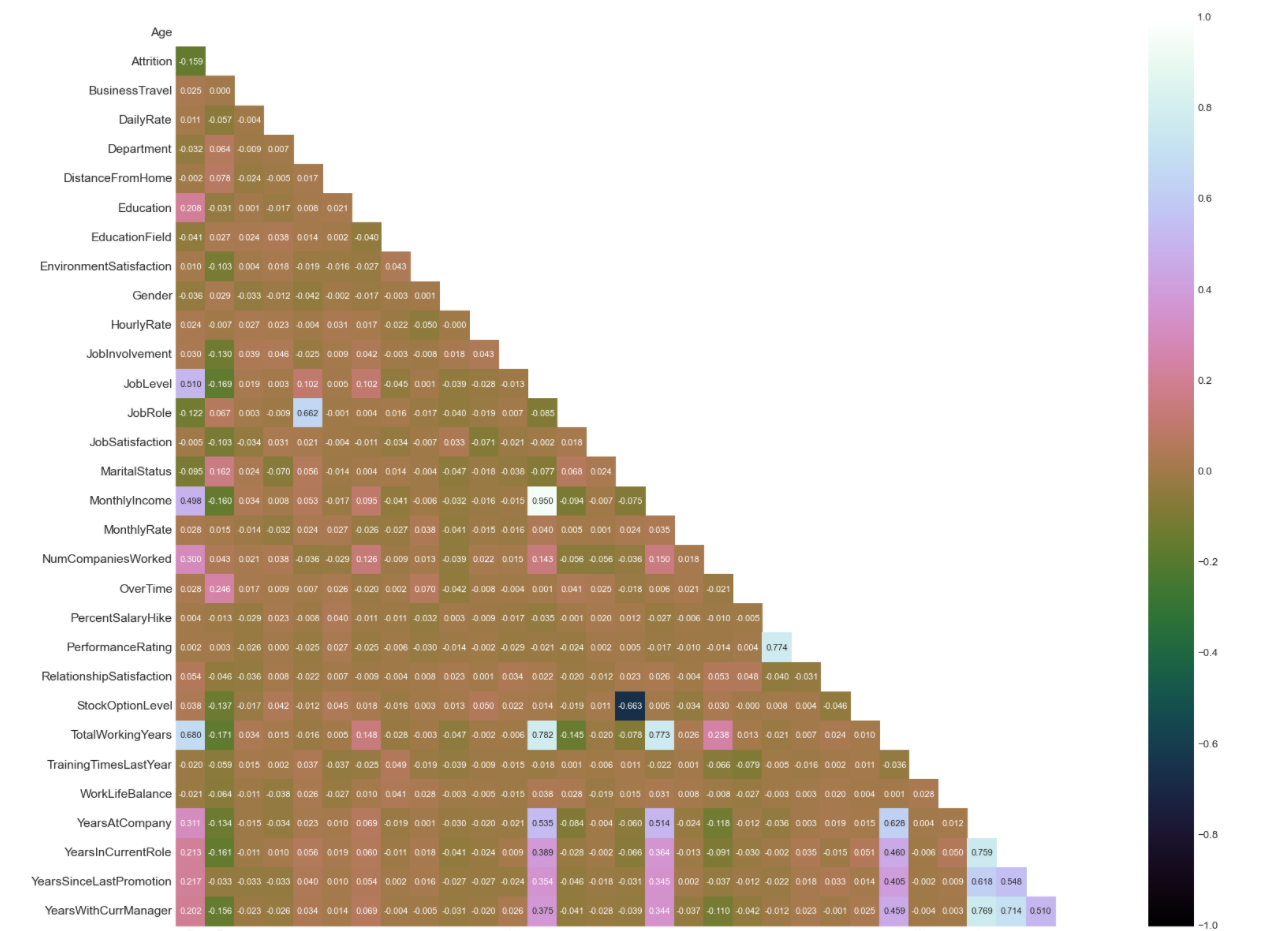


I have plotted histogram after using Encoder and it only give the distribution of numerical

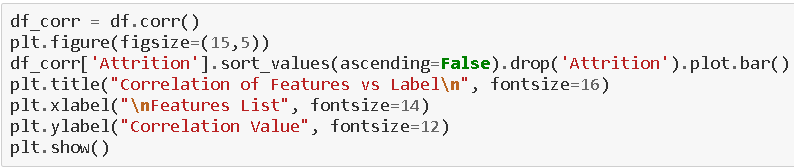
Columns present in our dataset.

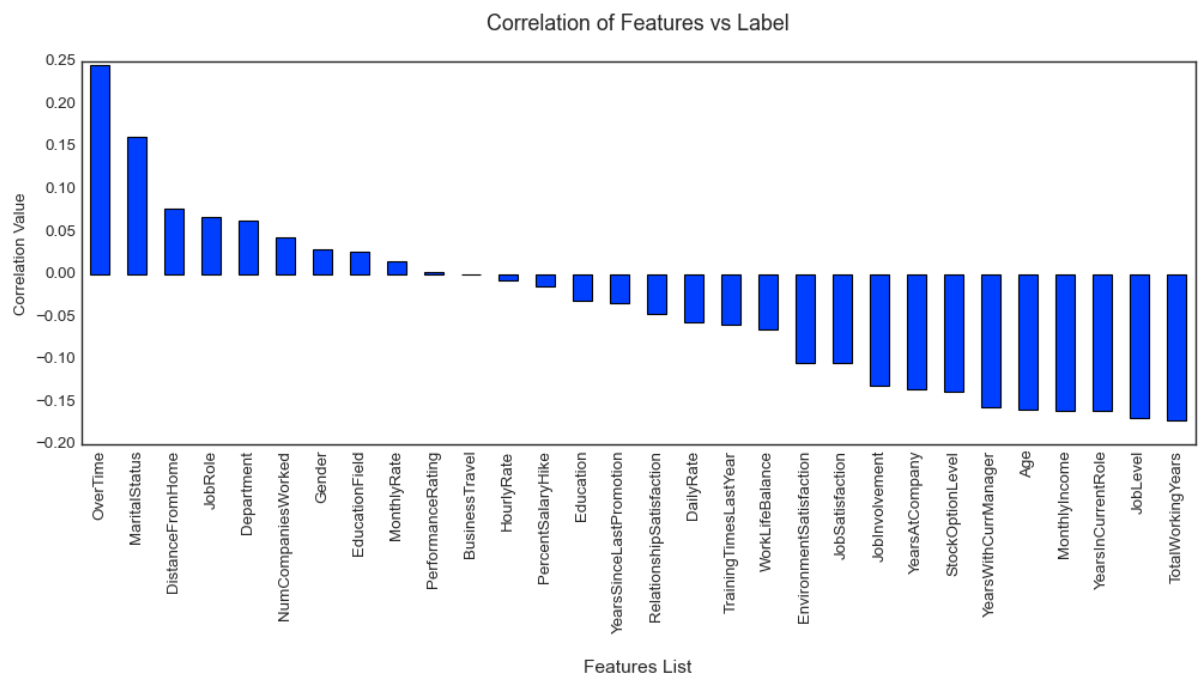
I now feel the need to check for correlation details in our dataset through a Heatmap. For those who still feel a confusion on correlation details let me break it down in two simple points that there are Positive correlation - A correlation of +1 indicates a perfect positive correlation, meaning that both variables move in the same direction together and Negative correlation - A correlation of –1 indicates a perfect negative correlation, meaning that as one variable goes up, the other goes down. The code to see this information is displayed below.



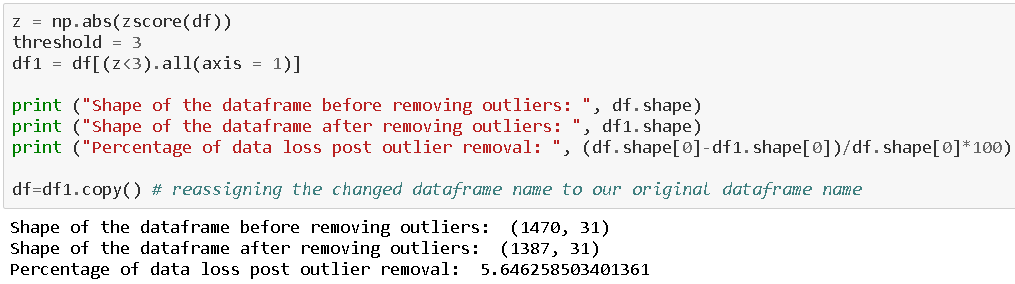


Obiously, because high number of columns present in our dataset I an not able to look at the numeric value of graph generated by the help of sns.heatmap. we can only have some information of multicollearnity by color .

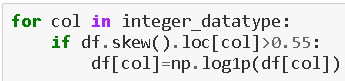




For the above bar graph atleast we are to find which of our columns of our dataset either negatively or positively correlates to our label target and also have some numeric information too.



From the above zscore function we have removed outliers present in our dataset .Since outlier was also part of our dataset. So, In this way we have lost our 5% data of our dataset. But no problem we can afford max 10% loss our datset if dataset is large.



As we it is very clear to all of us that our skewness acceptable range is +/-.5 , the skewness lying between this range is acceptable other wise we need to treat As we have treated above.



We had to split our dataset columns into two parts namely X and Y. X variable containing all the feature columns and Y variable containing our label/target variable.

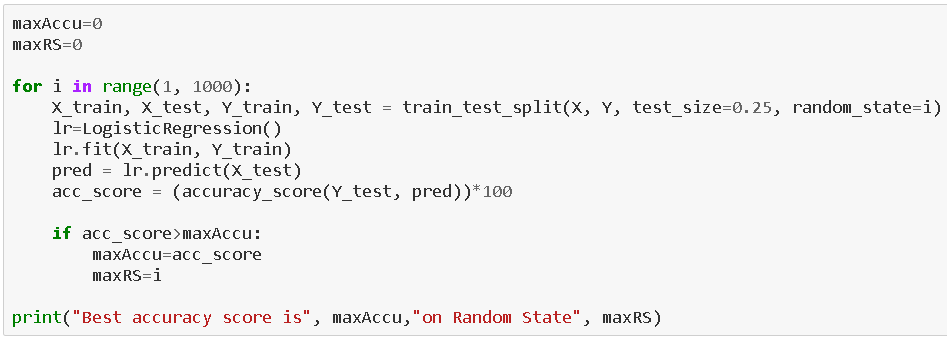
Also we have to take care of target imbalance issue .



By the help of this Smote method we have resolved the issue of imbalance .So, that our label don’t get biased .



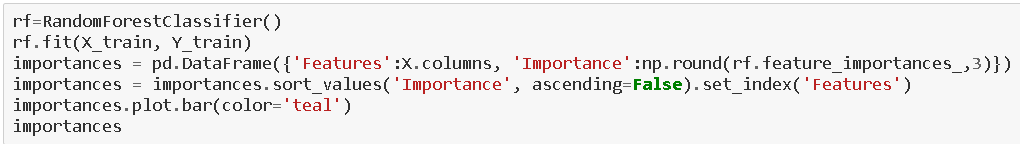
From the above standardScaler method we can scale that all the columns of our dataset so our dataset refrain from biasing for any particular column.

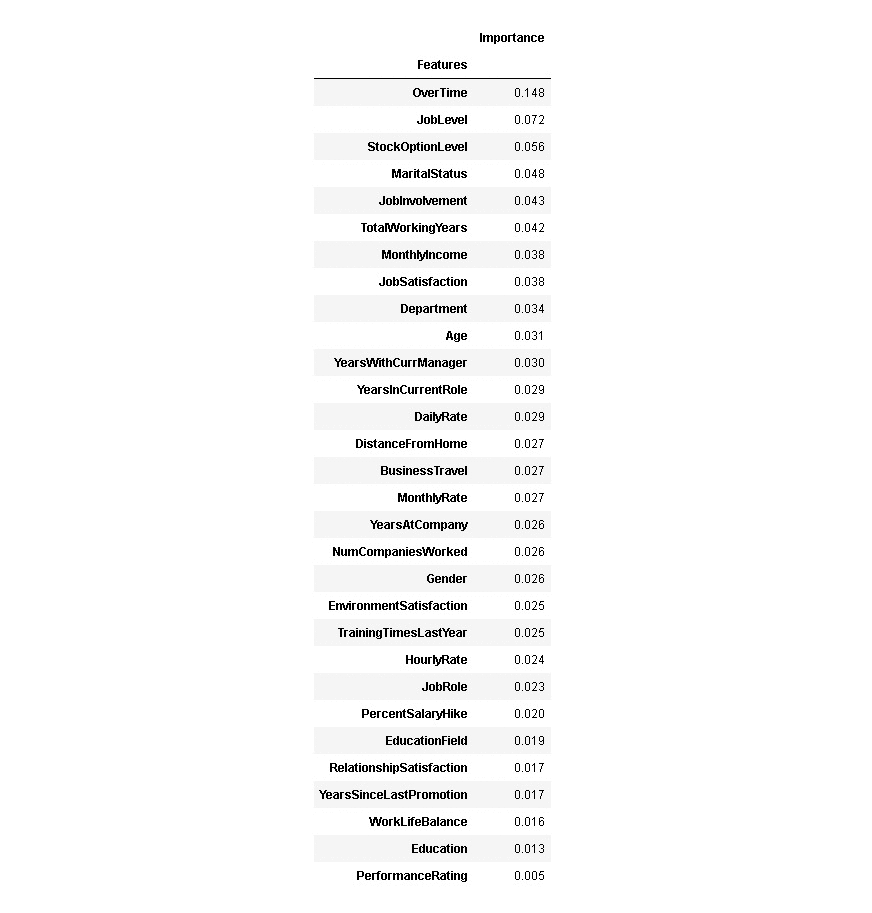


To find the best Random State for our Machine Learning Model .from the above I have sent 25% data for testing and 75% data for training. Actually, It’s up to you how precisely you want to make your model is.

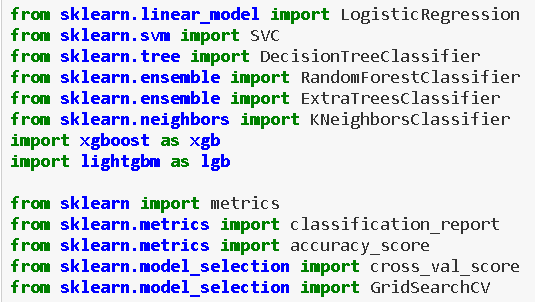


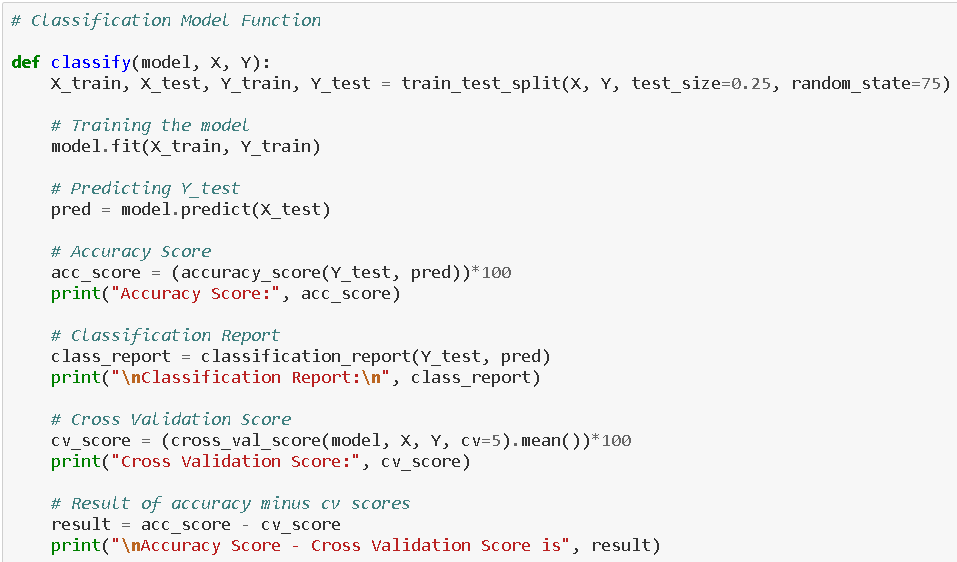
Before building the proper Machine learning model we have to decide how much data do you want to send as test and how much data do you want for training . I’ts totally up to you.

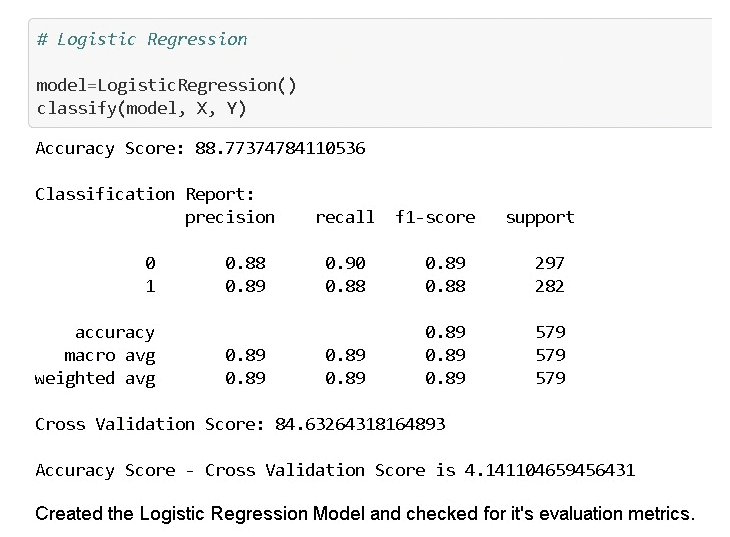




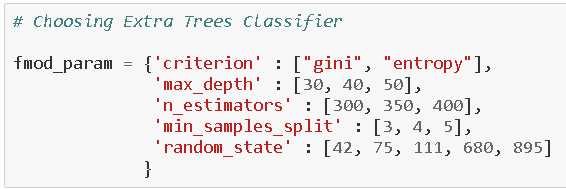
**5.      Building Machine Learning Models**  
  
In order to build a classification method I have imported the necessary libraries and created a function that contains all our machine learning model creation and its evaluation metrics steps. This makes our job easier since later on we just need to feed the model’s name and get the result without repeating/rewriting the same code again and again.







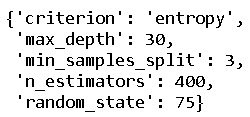
It is always advised to all of us that atleast we need to use 5 Algorithm in order to figure out which one is performing best among them and we choose that one and we send that for hyper parameter tuning to know that best parameter .



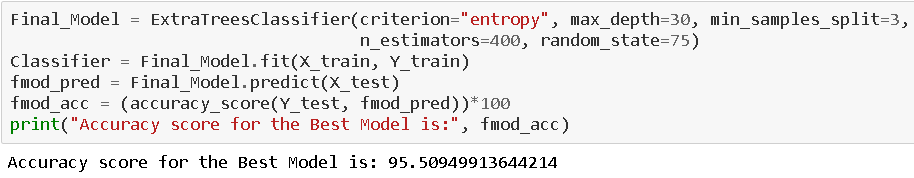


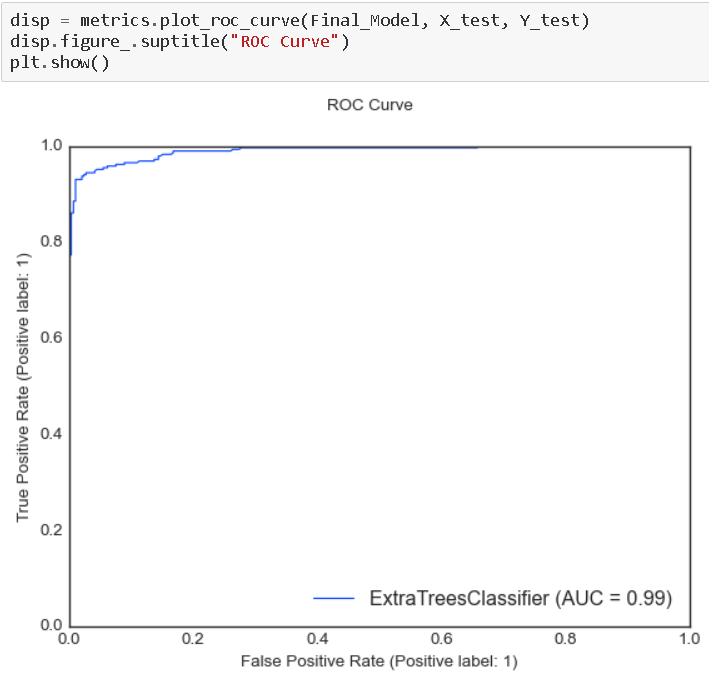






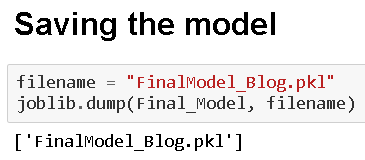
After applying the above steps to get the best parameters list, I simply have to plug it into my final model and receive the output of it. I have created an ROC curve plot and Confusion matrix for the final model.

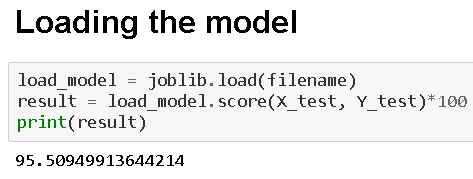






From the above confusion matrix we can see that our dataset is having 286 as True positive and 267 as True negative where as 6 as True negative and 20 as True negative value present in our dataset .





**6.     Concluding Remarks**

Let me go back to each steps I have taken in this project. from beginning we have

Understanding the problem definition to go through EDA Processes . we went to some certain pre-processing steps and finally build Machine learning Model.

In this entire project I have given my best and put all possible potential that I have. Most of time I used my own code but when I found on internet some better code I opted that code.

I don’t take other good work on ego rather I do admire. Obviously , I do believe in learning buy other project but yes it doesnot means that I copied the code.

For any of machine learning project my suggestion is first you have to understand the problem on ground level .if you don’t allow yourself to work with diligence .if you don’ t work harder anything that you are doing or will do , not only in case of machine learning but also in life cycle would be futile. Maybe, my endeavour assist you when ever you will get stuck

**Disclaimer**

I am new comer here in data science domain with some knowledge of 9 month. I have shared my effort in this blog to someone who is stepping in this field and can take some advantage from it . but to be honest it’s definitely inspired by others I saw many blog of this project on internet who have worked on this projects before me I just went through each of project and concluded my best way to make this blog.