**HR Analytics Project**



**By**

**PUNEET JAWA**

Today I am going to write about HR Analytics Project to understanding the Attrition in HR which should serve as a guiding path for many Data Science aspirants. When I began my journey into studying Data Science, I doubted myself , whether I have made correct decision by going through the content that is available online and most of them scattered into chunks emphasizing on a deeper knowledge that a newbie like me would be able to understand the concepts of this new technology named machine learning. I’ve done multiple projects now on this technology and with every project got new learning of data Science field and still going on.  
  
So, now without any further delay, let me explain the agenda for this blog post. In this article, I have jotted 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.

**1.      Problem Definition**

The project I am using for this article is for HR Analytics which is basically a fictional dataset that was created by the Data Scientists at IBM. the dataset is in the [link](https://github.com/dsrscientist/IBM_HR_Attrition_Rate_Analytics) provided.Basically**,** 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?

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.



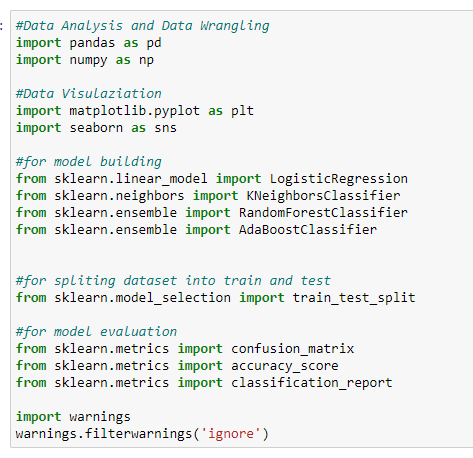
HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.

Attrition in 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 is a major problem since 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.

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 we are going to import all the necessary libraries that will be used in our project and obtain the rest as and when required.



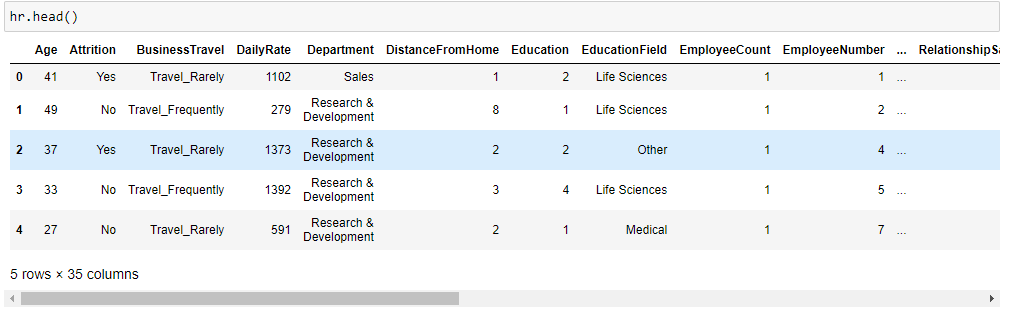
Before we begin with any process, we need to get the dataset in our Jupyter Notebook that can be achieved by a single step.



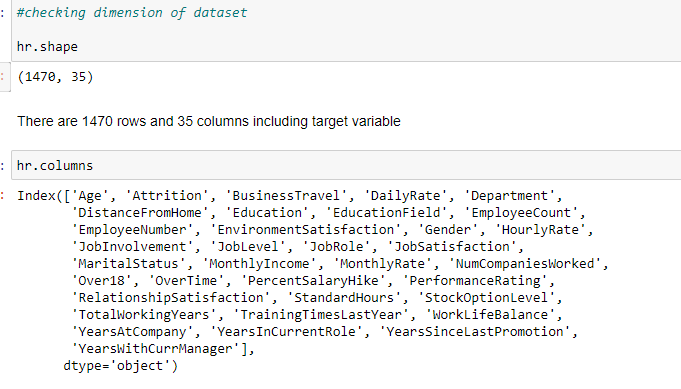
This gives us our entire dataset stored in the variable name “hr” for our dataframe.

**2.      Data Analysis**

For data analysis part, we will use different commands, that are as follows:



head() function is used to display the columns heading in the dataset.



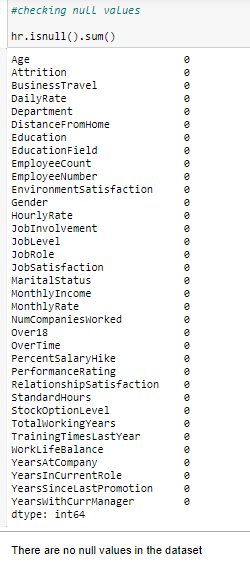
shape function, we can interpret how much rows and columns are there in our dataset.

columns function, we can interpret what all columns are there in our dataset.

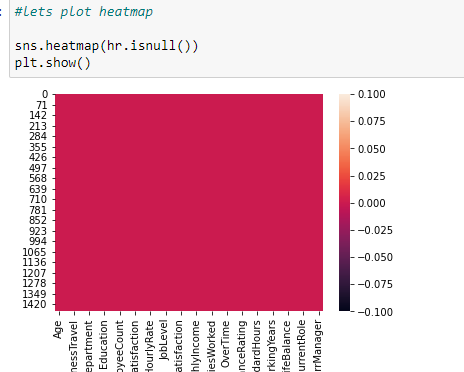
**3.      EDA**

EDA also known as Exploratory Data Analysis and is considered the most important aspect in Data Science industry by most Data Scientist. After following a huge number of expert Data Scientist on various platforms I can confirm one thing that it boils down to a single important thing of conveying a story on how we are able to achieve each and every step via code showing the provided problem statement, the observation, the challenges faced and what was done to tackle or rectify those issues.  
  
Building a good model comes only when we understand clearly what we are doing and why we are doing it. Making sure that we have a clean data in proper processed format to feed into our model and get appropriate result. Because no amount of Machine Learning model usage and hyper parameter tuning is going to help if we have not invested time to sort out and fix our data that’s the only input we have in hand.

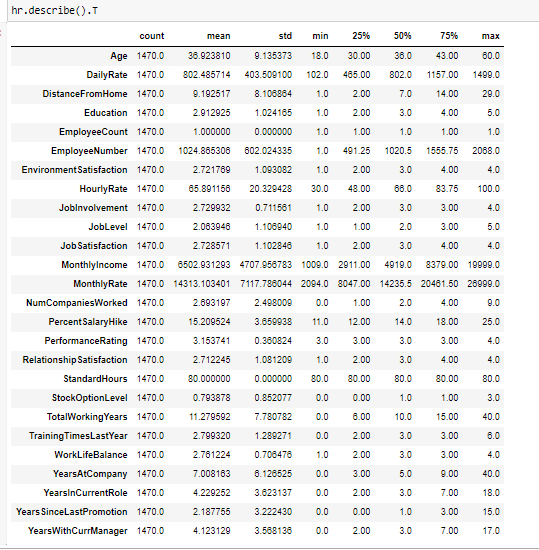
The first thing I am going to take a look at is the missing data information in our dataset by using the codes below, also including the output after executing the code.



Let’s check once heatmap also to be double sure of null values, here is the code:

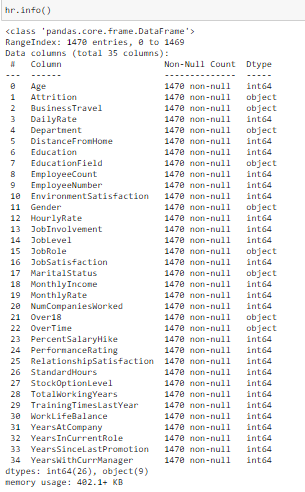


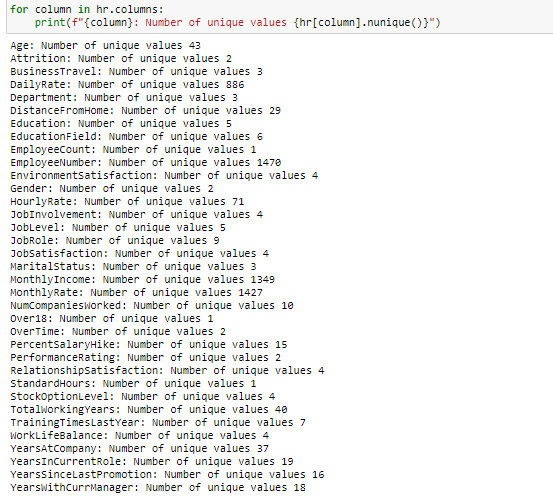
Next, let’s move on to using the describe method to take a look at the count value, mean data, standard deviation information and the minimum, maximum, 25% quartile, 50% quartile and 75% quartile details. As the describe method works best for numeric data all the object (text) type data gets ignored. Take a look at the below code and we will get an idea on how to use it.



When we are able to draw insights from the describe method we can take a look at the datatype information using the code below and that shall give us the list of all the columns marking them to be either integer, float or object datatype depending on the values present inside the columns.

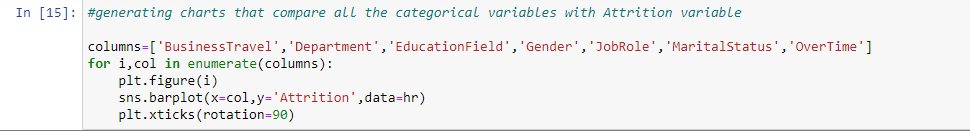
Below is the code with output.

  
  
This is explaining the datatypes of all the columns present in our data frame. We also get an opportunity to drop or remove any unwanted columns from the data frame here.  
  
Next we will check the unique values and will drop the unnecessary columns also which is of no use for us.

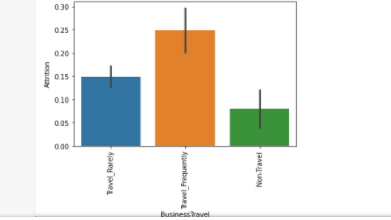
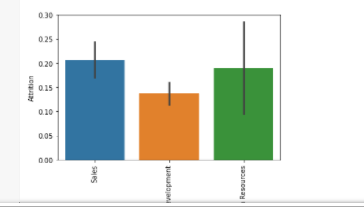


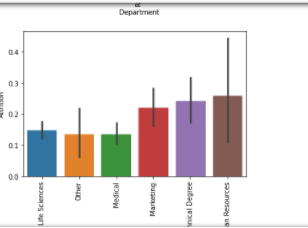
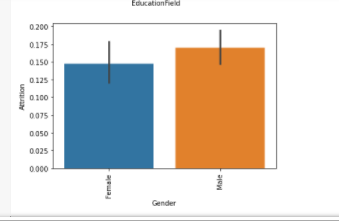
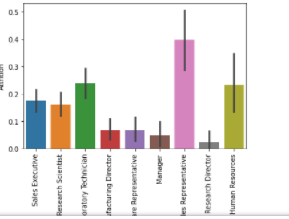
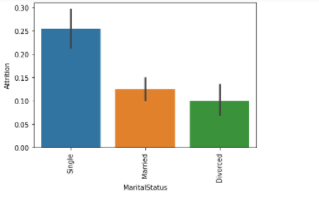
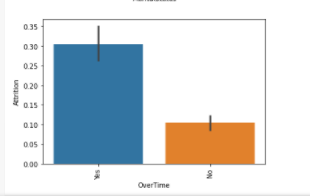


Next step is to list down all the visualization codes with their output.



Output for above code, contains multiple snips.

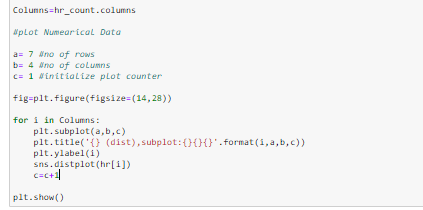
 

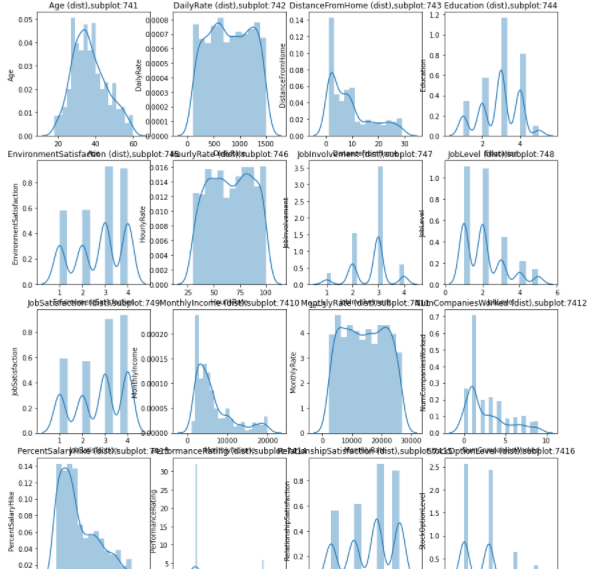
next code:



next code:



Output of above code:



We can see that with the help of above codes and the outputs I was able to take a look at all the column values/counts, bar/histplot 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 dealt with before I even think of building my Classification Machine Learning models.

**4.  Pre-processing Data**

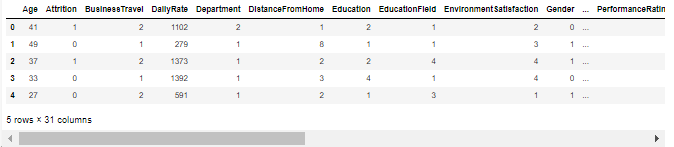
In the pre-processing step I am going to tackle all the miss fits and fix them one by one starting with the problem that out dataset has object datatype values where as our Machine Learning models can only understand numeric values. I am making use of the encoding methods to convert all the object datatype values. For our label I am using Label Encoder.

Code:





Output:

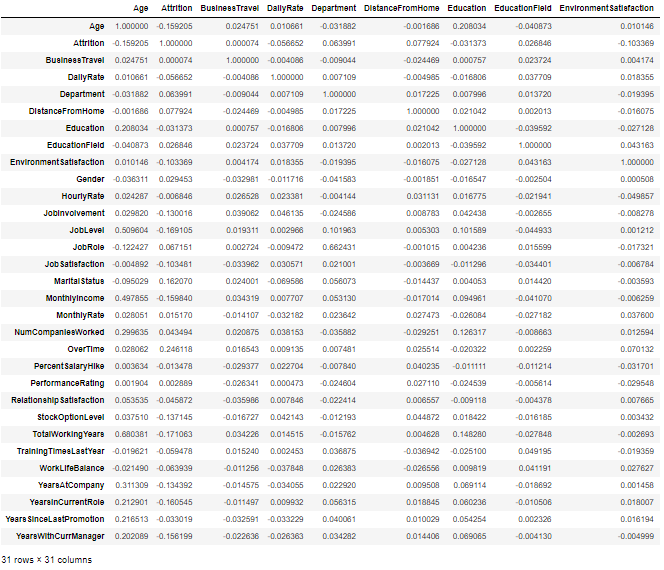


After I have encoded all the columns in our dataset, next I’ll check for correlation.

Here is the code:



Output :

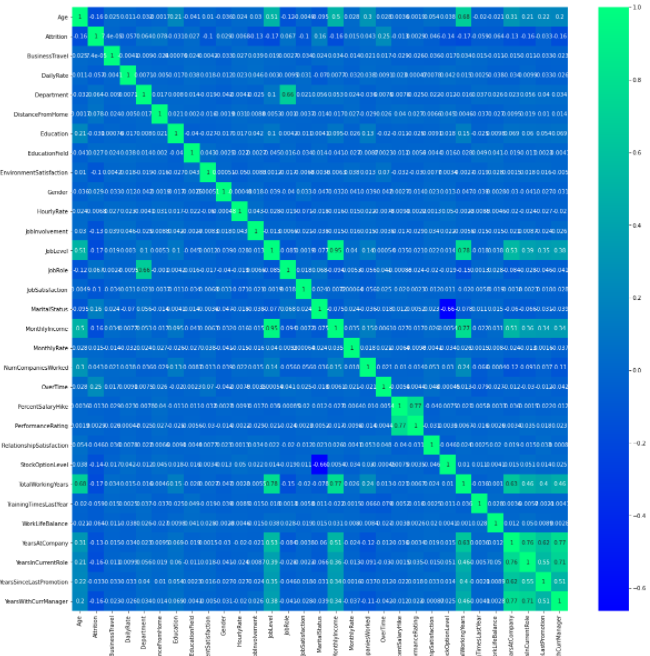


Next is Heatmap for the correlation:

Code:



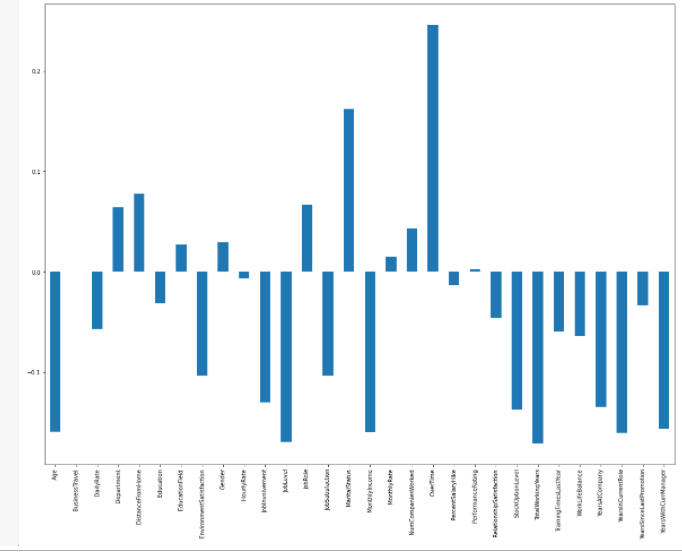
Output:



correlation between our label and feature columns I use a Bar Plot comparison and code is below:

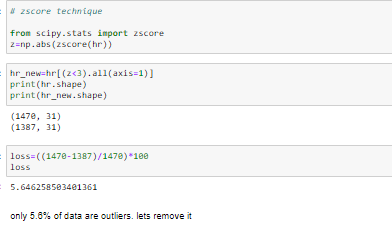


Output:



In the above Bar Plot we are able to clearly define the feature columns that are positively correlated with our label and the feature columns that are negatively correlated with our label. Now coming back to the outlier and skewness concern in our dataset I will be using the Z score method.

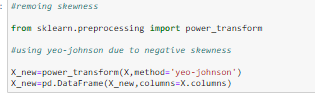
Code:



As for the usage of Z score, I was able to lose only about 5% of data.

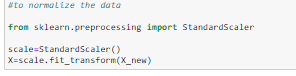
To Remove Skewness, I will use yeo johnson method because of negative skewness.

Here is the code :



After dealing with the data concerns I will then split our columns into feature and label. I am storing the feature columns in X and the target label column in the Y variable.

  
Now I will scale the feature columns that is stored in the X variable to avoid any kind of biasness over column values. Some integers cover thousands place and some cover hundreds or tens place then it can make the machine learning model assume the column with thousands place has a higher importance when in real that won’t be true due to difference in unit range.  
Code:



Next step is to check the best random state for the machine learning models.  
Code:



Now I will use the train test split to bifurcate our entire data set into training data and testing data. Here I am using 6% data for training purpose and 30% data for testing purpose. Some people provide training and test data separately as well and hence it completely depends on us how we want to use this step, code is below:

**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.

There are 4 models that I’ve used here to evaluate the best from those 4.

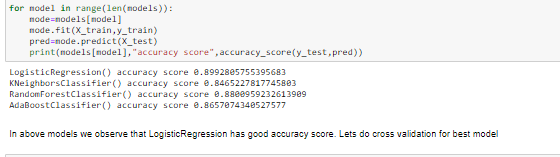
1. KNeighbors Classifier
2. Random Forest Classifier
3. AdaBoost Classifier
4. Logistic Regression.

Code:

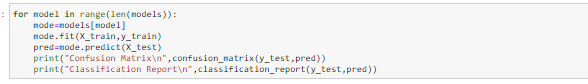


Code:

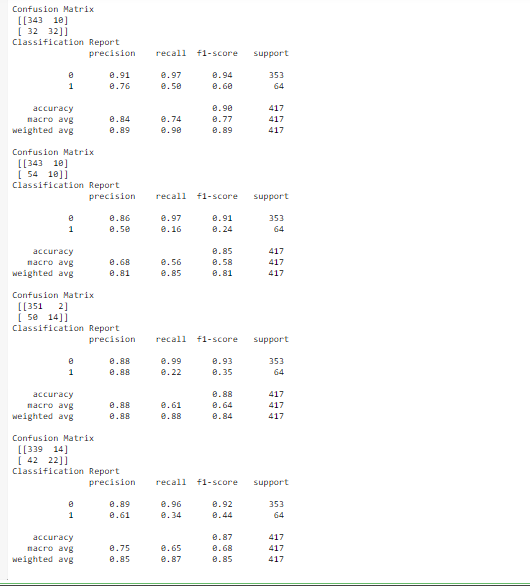




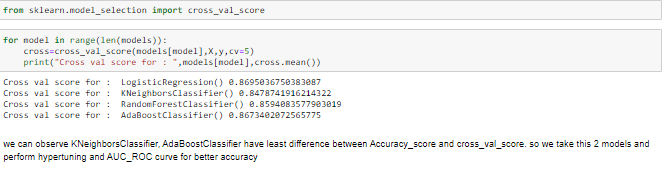
Code:



Output:



Next step is checking cross val score:



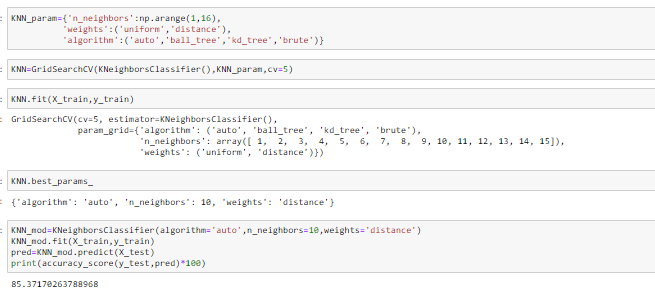
Next step is Hyper tunning the model:

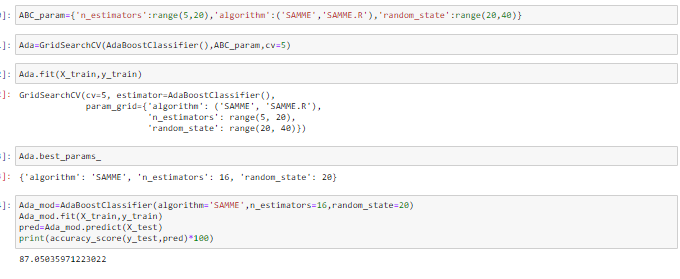
Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning.



We have applied hyper parameter tunning for 2 models i.e KNN and Adaboost to find out our best model.

For KNN, here we got accuracy score of 85%

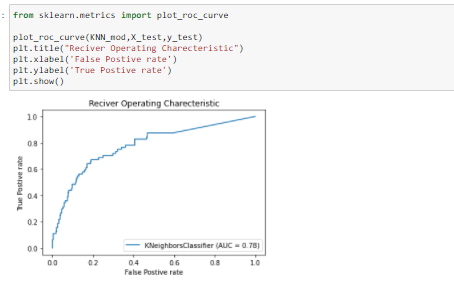
For ADABoost, here we got accuracy score of 87%



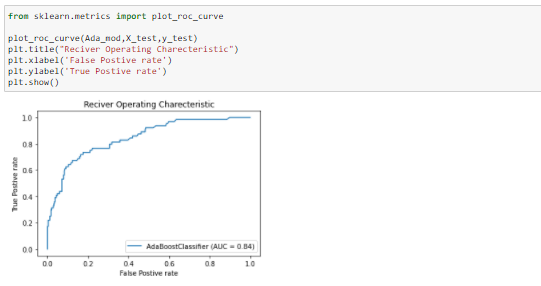
**AUC/ROC Cure**:

In Machine Learning, performance measurement is an essential task. When it comes to a classification problem, we can count on an AUC - ROC Curve. When we need to check or visualize the performance of the multi-class classification problem, we use the AUC (**Area Under The Curve**) ROC (**Receiver Operating Characteristics**) curve. It is one of the most important evaluation metrics for checking any classification model’s performance. It is also written as AUROC (**Area Under the Receiver Operating Characteristics**).

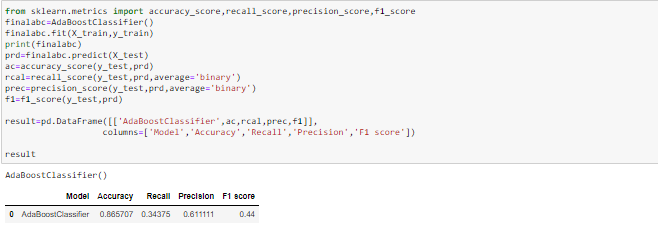
**AUC score for KNN is 78%.**



**AUC score for ADABOOST is 84%.**



Best model is AdaBoostClassifier.









6.     **Concluding Remarks**

Let me provide a quick recap on all the steps that we went through starting from understanding the Problem Definition then going through the Data Analysis and EDA processes. We went through the necessary Pre-processing Data steps before the final Building Machine Learning Models step came into picture.  
  
In this entire project I took help from internet to check further and improvise on accuracy or beautify the visuals. However I have observed that other people doing complete copy paste and don’t even check it, what is mentioned in that code.

Before wrapping up my only advise to everyone is to try the code on your own and if u are not able to get any proper content or solution then refer to internet for the code.