

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

I am going to write end-to-end Advertising Sales Prediction Project. When I was a beginner I did not have any idea of machine learning I had to face much difficulty but as soon as I found sort of blog on Google my all problem get resolved . My concern to write the entire blog to help and inspired other to learn and forward it those who is true aspirant. Since there are so many blog already present there in internet then what was the need to write down the one more blog . Yes , no doubt the all blog of Advertising Sales prediction project mind-blowing . I went through many of them and found that some steps could have more easier to write and understand. I want that step to fetch you easy to understand and Machine learning model in better way.

Let it be stopped this discussion overhere and proceed ahead with the project .I have put down my all techniques in the form of subtopics or points and those all point s are as follows..

1. Problem definition

2. Data Analysis

3.EDA

4 .Pre-processing

5. Building Machine Learning Model

6. Conclusion Remarks

Let’s start from Problem definition to know why we kept this point on the first step.

**Problem Defination**

Sales Channel Prediction Case Study

When a company enters a market, the distribution strategy and channel it uses are keys to its success in the market, as well as market know-how and customer knowledge and understanding. Because an effective distribution strategy under efficient supply-chain management opens doors for attaining competitive advantage and strong brand equity in the market, it is a component of the marketing mix that cannot be ignored.

The distribution strategy and the channel design have to be right the first time. The case study of Sales channel includes the detailed study of TV, radio and newspaper channel. The predict the total sales generated from all the sales channel.

**import** **warnings**

warnings.simplefilter("ignore")

**import** **joblib**

**import** **pandas** **as** **pd**

**import** **numpy** **as** **np**

**import** **seaborn** **as** **sns**

**import** **matplotlib.pyplot** **as** **plt**

%**matplotlib** inline

**from** **math** **import** sqrt

**from** **sklearn.preprocessing** **import** StandardScaler

**from** **sklearn.model\_selection** **import** train\_test\_split

**from** **sklearn.linear\_model** **import** LinearRegression

**from** **sklearn.svm** **import** SVR

**from** **sklearn.tree** **import** DecisionTreeRegressor

**from** **sklearn.ensemble** **import** RandomForestRegressor

**from** **sklearn.neighbors** **import** KNeighborsRegressor

**from** **sklearn.ensemble** **import** AdaBoostRegressor

**from** **sklearn.ensemble** **import** ExtraTreesRegressor

**from** **sklearn.ensemble** **import** GradientBoostingRegressor

**from** **sklearn.metrics** **import** r2\_score

**from** **sklearn.metrics** **import** mean\_squared\_error

**from** **sklearn.model\_selection** **import** cross\_val\_score

**from** **sklearn.model\_selection** **import** GridSearchCV

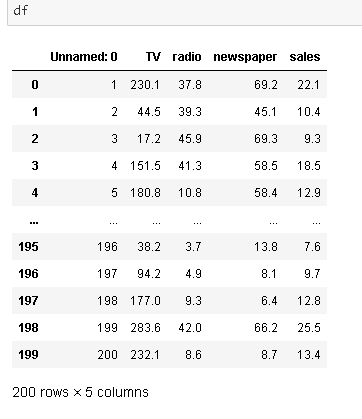
First of all I have imported all the necessary libraries or tools that is sure going to come into help of building best machine learning model. I have to import whole the dataset into single jupytor notebook that we can apply upon various method and machine learning Algorithm.

df = pd.read\_csv("https://raw.githubusercontent.com/dsrscientist/DSData/master/Advertising.csv")

From above pd.read\_csv method we have imported into entire the dataset and stored into single variable 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 .



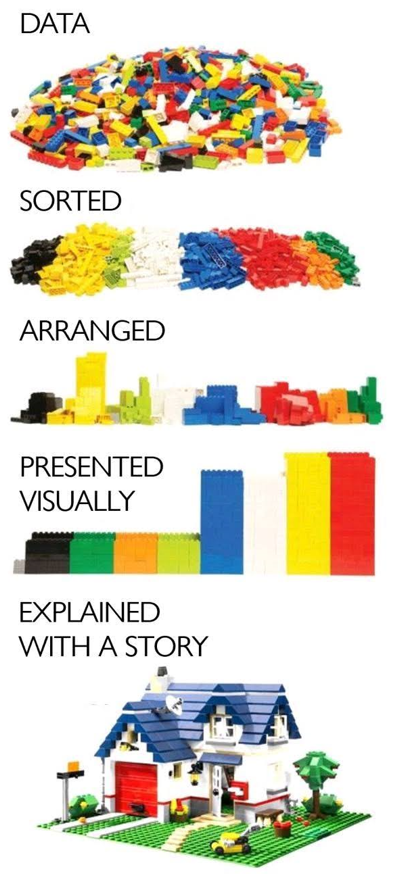
From the above we can say our data set is having 200 rows and 5 columns.

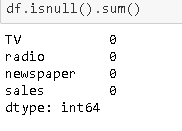
**3.EDA**

Without any hesitation we can say that EDA, ( 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 pysthon libraries.

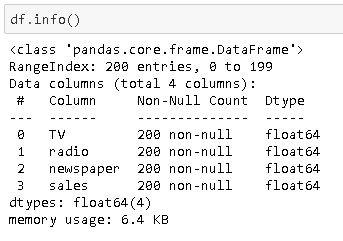
From the below picture i can explain :

1. 1. For any sort of Machine learning Model data is the most important thing that you must have .Threre are so many way to collect the data . I will not go into depth here but I can say primary and secondry there two are the main sources of data collection
2. Then I sorted the entire data as per their feauture 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 you doing the work upon.

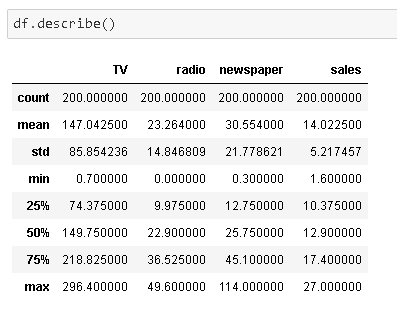




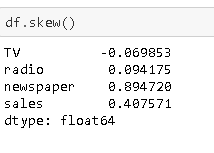
From the above isnull(). Method and sum function we can say that our dataset is free from the issue of missing value or data.



From the df.info method it is clearly displaying that our entire dataset is in the form of numeric . that is better to perform operation ahead. Since not a single column is belonging from object data type group so we are free from using any Encoder to convert into numerical datatype.

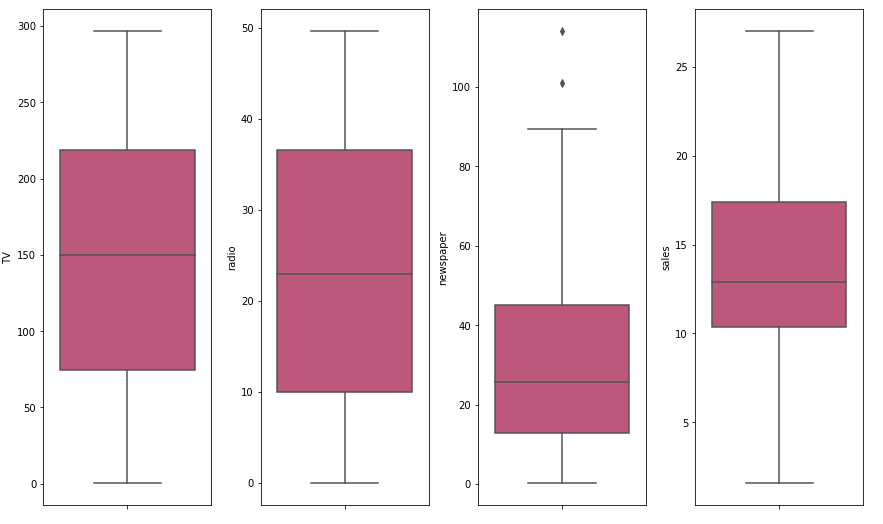


We have used df.describe method to look some insight information of our dataset. Describe method is very important for our dataset .by using this method we can get outlier present in our dataset and also missing numerical data present in our dataset.

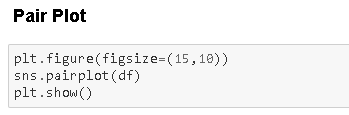


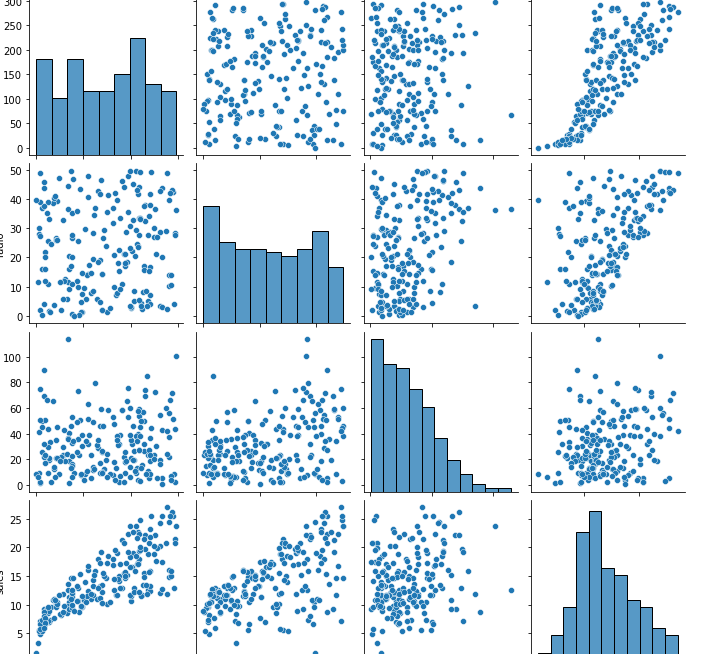
If we want to know how data of our dataset is skewed we use this df.skew method . Acceptable range for the columns is +/-.5. The skewness of data lies between this range we need to accept that column otherwise we need to use some technique to reduce skewness.

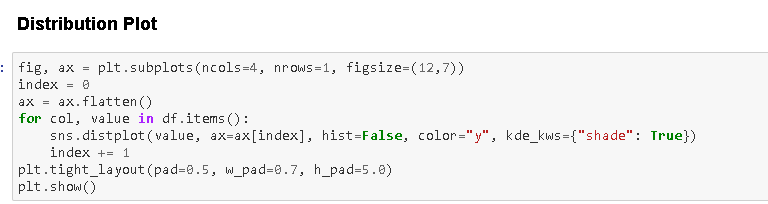


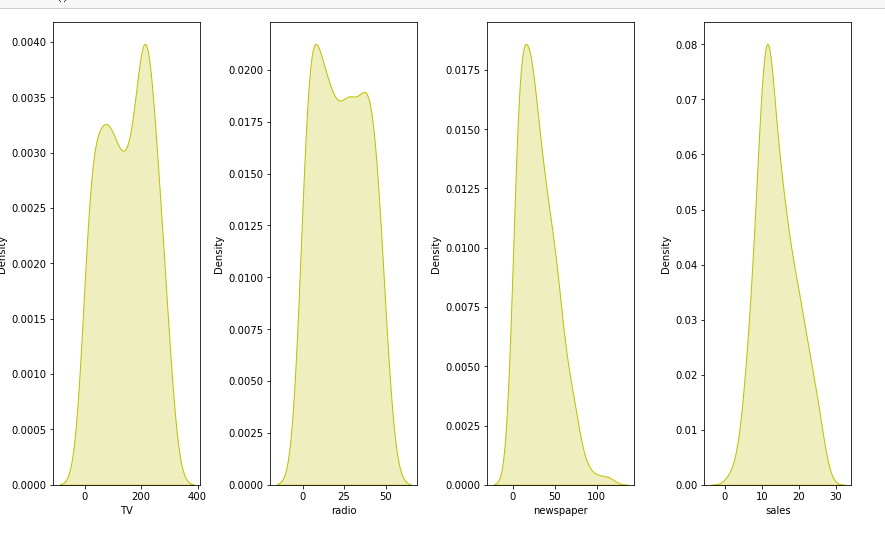


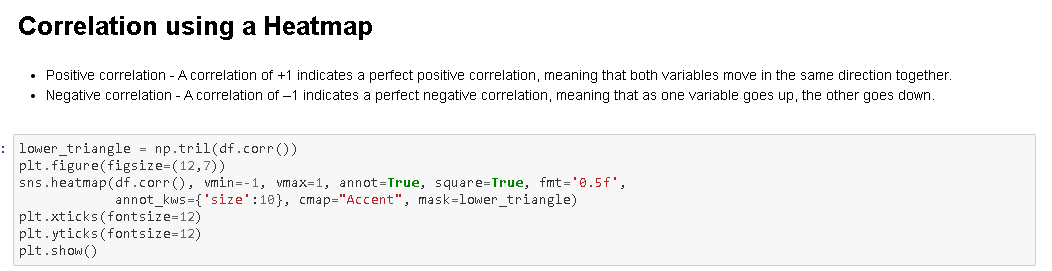
From the above box plot we can see that there some outliers present in newspaper column in our dataset. We need to treat them before heading for building machine learning model.

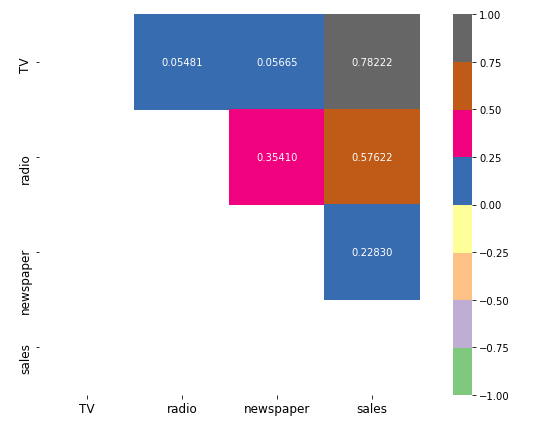




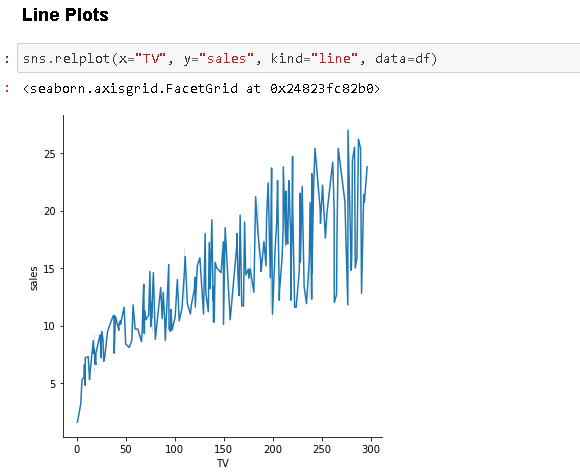


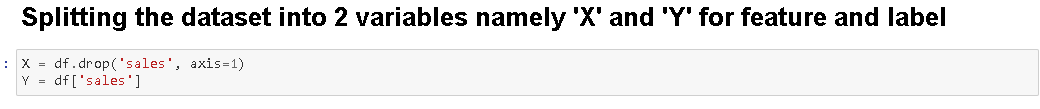




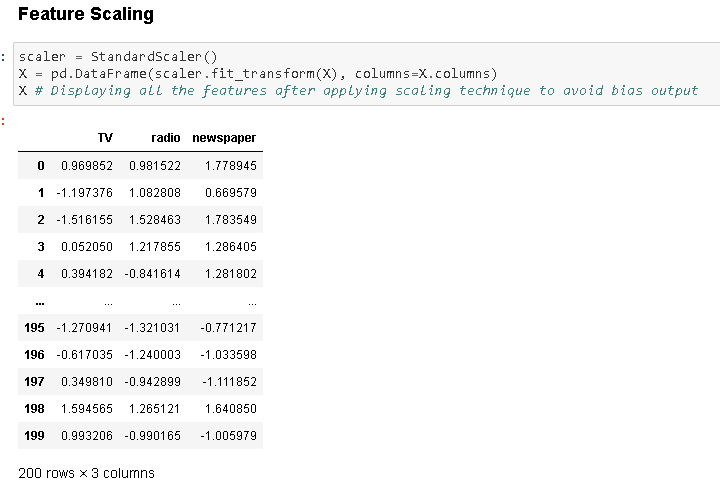


From the above sns.heat map method we found how our data in dataset is correlated to each other . Here we can see that there is no any columns showing us multicollearnity.





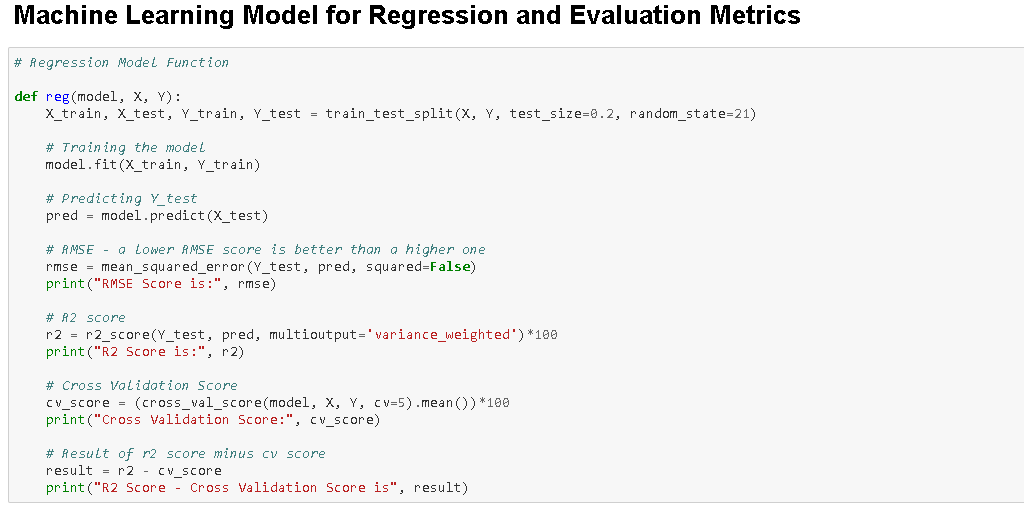
We have splitted our entire dataset into two parts namely X and Y.



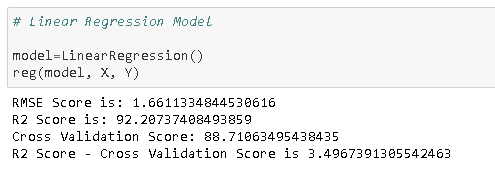
From the help of this feature scaling we get free from the issue of biasness for any particular column .



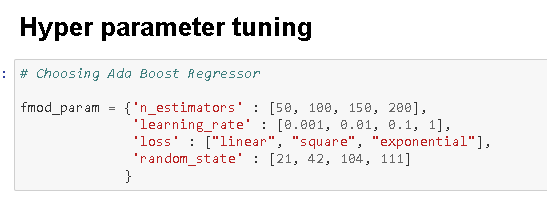
From the above train test split method we have sent 20% of our data for testing and 80% of data I kept as training purpose .



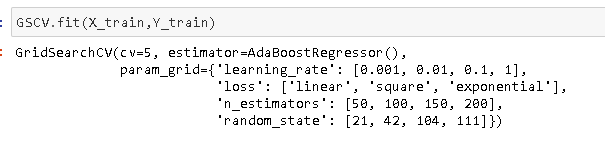
We have created a Machine Learning Model for Regression function and I have built a function that splits the training and testing features and labels, then trains the model, predicts the label, calculates the RMSE score, generates the R2 score, calculates the Cross Validation score and finally finds the difference between the R2 score and Cross Validation score.



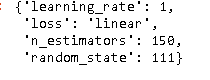
I am always told that atleasst we have to use five sort of Algorithm to find best in among them and then we just need to go with that Algorithm for Hyper parameter tuning .

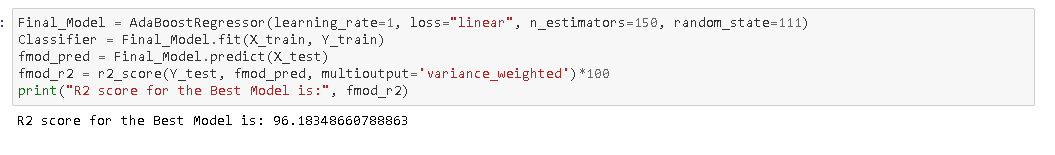












Here is the final model we have with r2 score of 96% .



We have finally saved our final data by the help of joblib. Dump method.

**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 accumulated knowledge of 1year’s Since I felt the more you will share your knowledge by any means , the more your concept gets cleared and it is mathematical rule that if u give someone something you will always be on positive side. So ,I have shared my all effort in this blog to the some1 who is stepping in this field and can take some advantage from it .To be honest I saw many awesome blog of this project on internet I just went through each of project and concluded my best way to satisfy all them in once.