**Advertising Sales Channel Prediction using Machine Learning**

Introduction:

Advertisement is the most important way by which Companies are promoting their businesses at a large scale since a very long time. Advertising Sales Channels helps a Company to determine from which all promotional ways the company is getting most revenue from. Mostly used sales promotion methods are T.V, Radio and Newspaper. By analyzing the sales from different channels, the companies get to know more about their customers and their understandings. So we will predict the total sales done by company using different sales channel with the help of Machine Learning.

For prediction using Machine Learning we need to download the dataset which we can get from the link: **https://www.kaggle.com/ashydv/sales-prediction-simple-linear-regression/data**

**1. Problem Definition:**

In the dataset there are Feature columnsand Target column. By seeing dataset we can figure out it is a **Regression model** as the target column has discrete numeric data.

The columns in the dataset are:

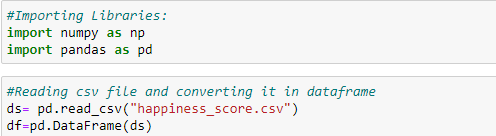
a) T.V

b) Newspaper

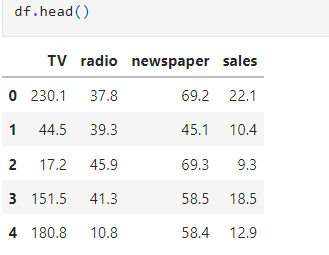
c) Radio

d) Sales

i) Importing Dataset:

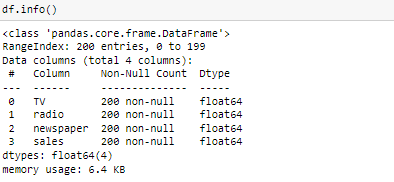


By Importing Pandas and Numpylibraries we will read the CSV(Comma Separated Variable) file and convert it into dataframe. We can see the first 5 data rows using df.head()



ii) Study Dataset:

Now we will check the data in details for proper analysis. Firstly we start with dataset information.

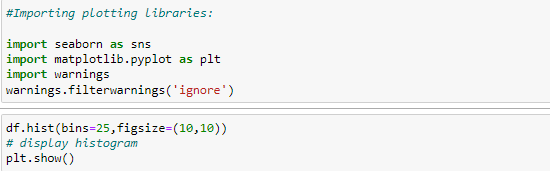


From the information, we get to know that in the dataset there are no null values. There are 4 columns and 200 rows and the datatype are 4 Numeric types.

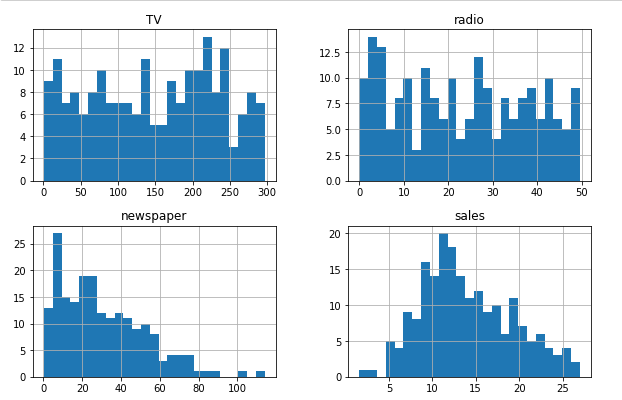
In machine learning data should be in numeric datatype and if object datatype is present then encoding technique is use to convert object into numeric datatype.

**2. Data Visualization:**

Importing Visualization libraries like seaborn and matplotlib. Plotting helps to see the data in graphical form for better clarification.

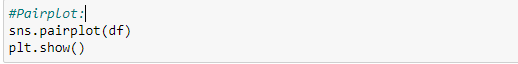


Output:

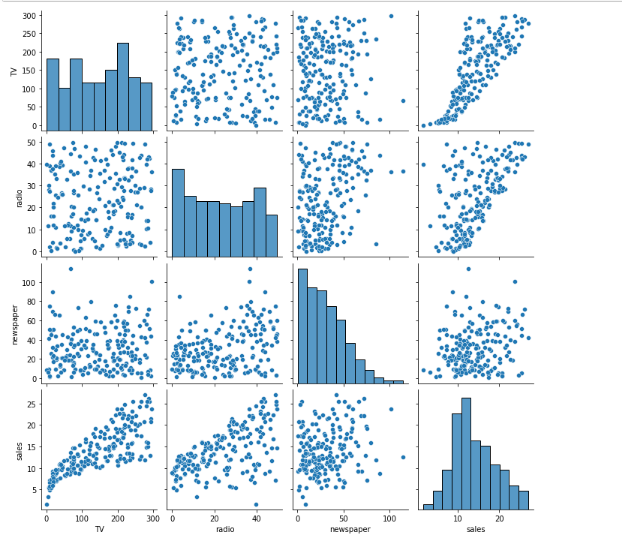


The above histogram plotting shows features are easily plotted. We can see the each data is not equally distributed and also have skewness.

Now we will use pair plot for plotting all the columns.



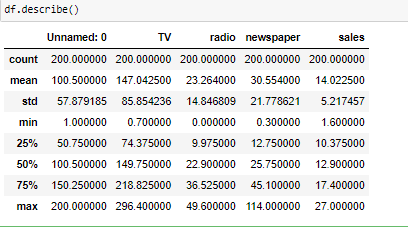
Output:



From the above plot we can see that the sales are increasing in TV and the also with the radio promotions. Newspaper channel sale values are constant.

i)Data Description:

Using describe ( ) we can see the presence of skewness, outliers and also can see the spread of data in the dataset.



From the dataset we can figure out:

a) Skewness: Difference between mean and 50% is small so skewness is not present much.

b) Outliers: 75% and max has larger difference in Newspaper so it has outliers.

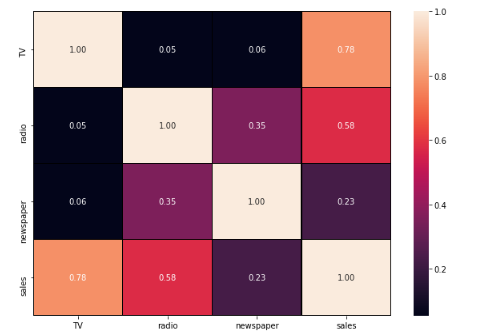
c) Data figure: Std( standard deviation) is less than mean so high peak data

ii) Correlation of the columns with the target columns:

A correlation shows the relation of each column with other columns in the dataset. Correlation will be positive and negative one. Positive correlation shows that the close link between the columns and negative correlation shows the not so close link between the columns. We use dataframe\_name.corr ( ) to check the correlation of the dataframe. Using heatmap for plotting correlations helps to analyse data in better way.



Output:



The lighter shade shows positive correlation and the darker color shows negative correlation.

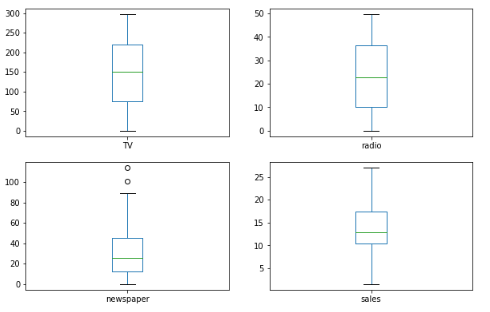
The target column Sales has positive correlation with TV and negative correlation with Newspaper. It helps for figure out which column is having positive and negative correlation with the target column.

iii) Outliers:

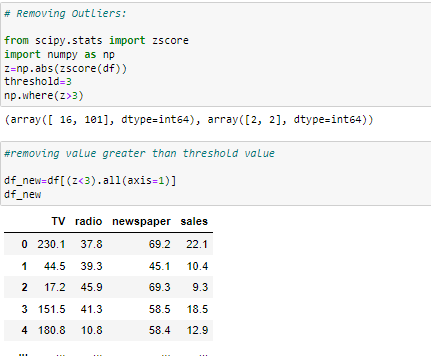
Now we will check for outliers in the dataset using boxplot. As outliers are value which are out of range from the other observation in the data.



Output:



Only Newspaper column has outliers as two circles lies on top of the boxplot. The outliers need to be removed for accurate predictions so for that we will use z-score method.



Threshold value in z score is 3 so value above 3 are considered outliers and removed from the data.

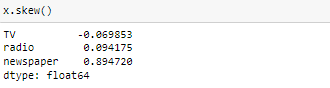
iv) Skewness: Skewness shows how much a sample of data is distorted in comparison with the normal distribution. We use skewness only on feature column but not on target column. So firstly we need to separate the feature and target columns.



x= the feature columns

y= target column

Now we have to check for skewness in ‘x’ by x.skew ( ) and it will show the column having maximum skewness. Skewness range is between +/-0.5



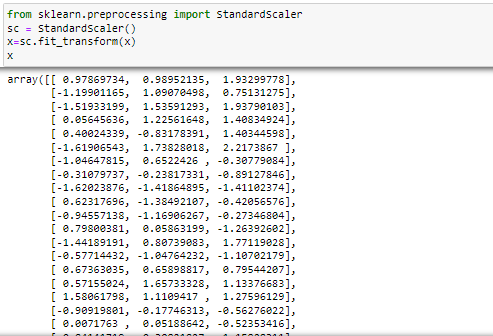
As per the skewness range of +/-0.5 Newspaper has value of 0.89 but it can be neglected as no other column have skewness.

**3. EDA Concluding Remarks:**

From the EDA we can conclude that the target column sale is mostly correlated with the TV column. The Newspaper column had some outliers so they are removed. Data is separated into feature and target column and skewness is also checked. Skewness range in the dataset is low so no need to remove skewness. Now the data is properly analyzed and cleaned so we will move towards Preprocessing of data.

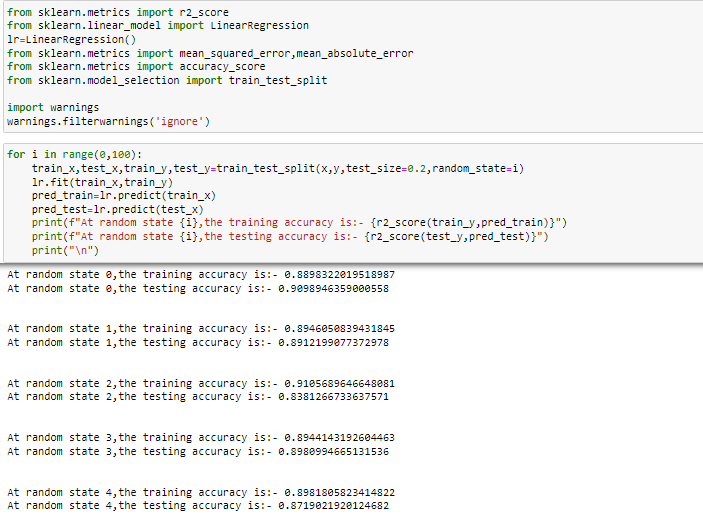
**4. Pre-Processing Pipeline:**

Pre-Processing of data is done by scaling the feature columns in the dataset. There are various methods for it. We will be using one the method which is mostly used that is Standard scalar for scaling.



i) Train-test split:

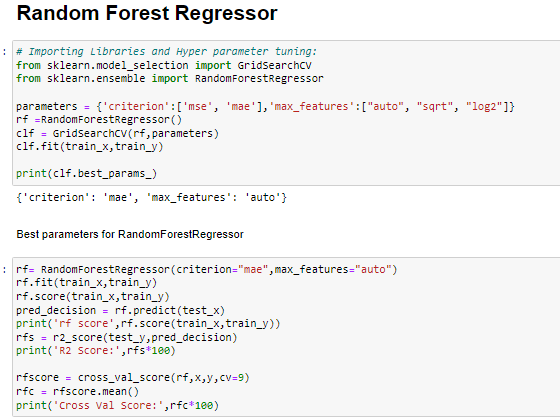
Splitting the data into train and test as it helps for train the data first and then test its outcome. It helps in model selection in machine learning.



**5. Building Machine Learning Model:**

For model building we use Algorithms like Decision Tree Regressor, Random forest Regressor and many others to find the best fit model for our dataset.

For this dataset the best algorithm is **Random Forest Regressor** as its r2score and cross validation score has smaller difference.

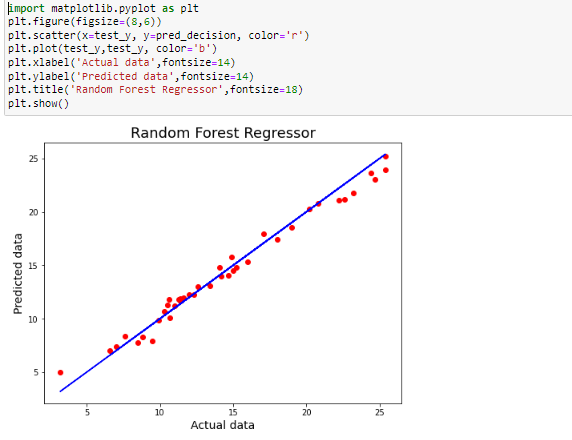


Output:



As the r2 score is 98% and Cross Val Score is 97.7% so difference is less.

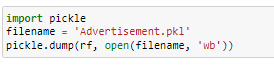
We can plot the model outcome using matplotlib



All the data are touching the best fit line so model is performing well.

i) Saving model:

Pickle library is used to save the best fit model for the dataset.



After this step our prediction is over for this dataset.

**6. Concluding Remarks:**

The topic of the project is really interesting as it helped me to know more insights about how company manages his brand value by promoting its product/services via various advertisement sales channels. This project also helped me to know about the importance of distribution strategy and channel design for company’s growth.

For this project I have used simple language and have shown step wise process for using Machine Learning for Advertising sales channel prediction.

* Author: Anita Thapa