**Predicting Advertising vs Sales using Python**

**By: ROSHAN KUMAR**

Problem Statement

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.

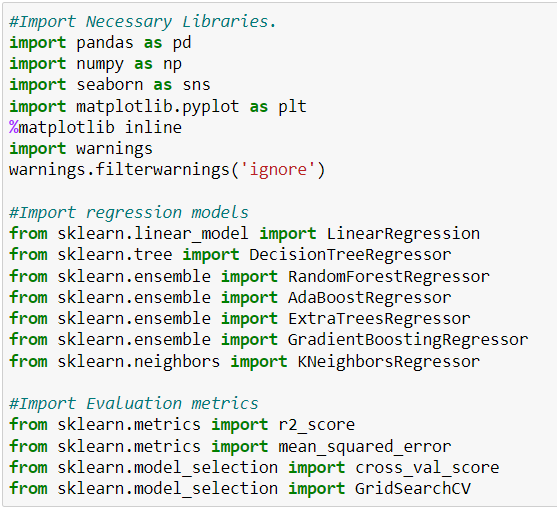
Problem Solving Description

As per the problem statement I have to predict the sales hence 'Sales' is my target variable and rest all the features are my independent variable. In sales column I observed that the numbers are in continuous so this is basically a regression problem. I will be building some regression models and will compare them according to their metrics, then I choose the best model using evaluation metrics to predict Sales count.

TARGET VARIABLE = SALES

https://github.com/ROKR7381/data\_trained-project\_only/blob/main/advertisement\_channel\_prediction.ipynb

# Import Necessary Libraries



# Reading Dataset

*#import datasets*

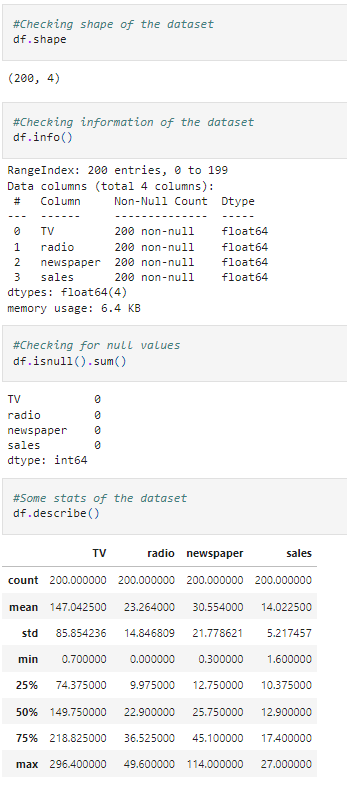
df **=** pd.read\_csv("https://raw.githubusercontent.com/dsrscientist/DSData/marising. df.head()

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Unnamed: 0** | **TV** | **radio** | **newspaper** | **sales** |
| **0** 1 | 230.1 | 37.8 | 69.2 | 22.1 |
| **1** 2 | 44.5 | 39.3 | 45.1 | 10.4 |
| **2** 3 | 17.2 | 45.9 | 69.3 | 9.3 |
| **3** 4 | 151.5 | 41.3 | 58.5 | 18.5 |
| **4** 5 | 180.8 | 10.8 | 58.4 | 12.9 |

# Feature Description

* Unnamed:0 - does not contribute in sales will remove it.
* TV - It refers to the sales count generated from TV advertisement.
* Radio - Here radio refers to the sale count generated from radio advertisement.
* Newspaper - Newspaper column shows the sales count generated from newspaper sales.
* Sales - Sales column refers to the total sales of TV, radio and newspaper.

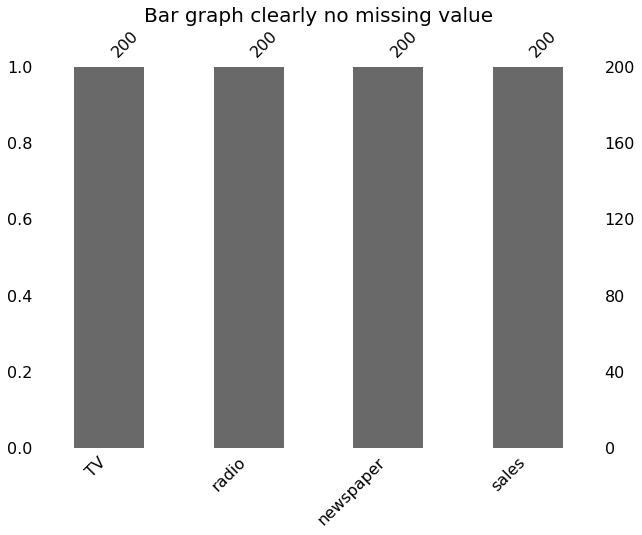
# Stats of the Dataset



In this dataset we have 200 rows and 4 columns, no object data only float. No Null values are present in the dataset.

ms.bar(df,figsize**=**(10,7))

plt.title("Bar graph clearly no missing value",size**=**20);



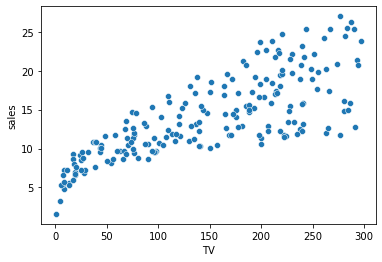
# Data Visualization and Pre-processing

checking correlation amomg features ans sales

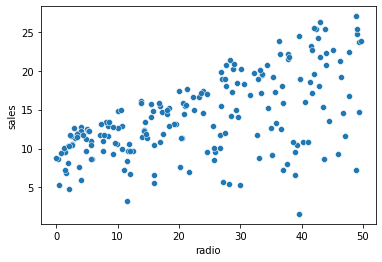
df.corr()['sales'].sort\_values()

|  |  |
| --- | --- |
| newspaper | 0.228299 |
| radio | 0.576223 |
| TV | 0.782224 |
| sales | 1.000000 |
| Name: sales, | dtype: float64 |

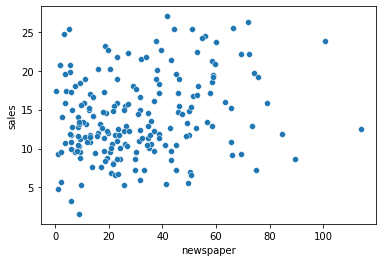
sns.scatterplot(x**=**'TV',y**=**"sales",data**=**df);



sns.scatterplot(x**=**'radio',y**=**"sales",data**=**df);

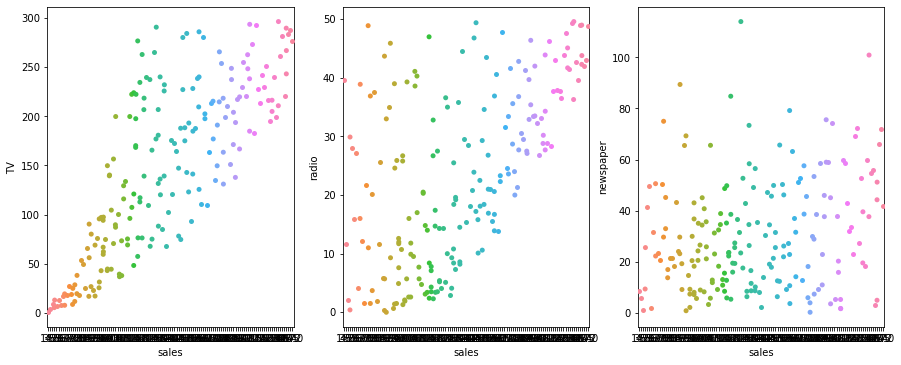


sns.scatterplot(x**=**'newspaper',y**=**"sales",data**=**df);



*#relationship between independent variables and their categories*



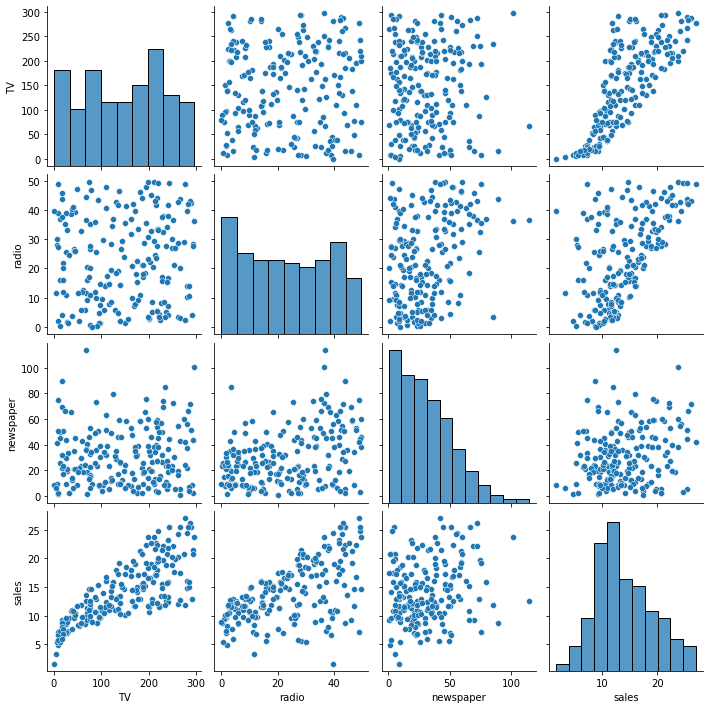


We observe that as the advertisement in radio and tv increases sales also shows a positive growth but

# **advertisement in the newspaper has not perform much growth**

plt.figure(figsize **=** (15,10)) sns.pairplot(df)

plt.show();

****

From the above bar plot, I can see strong relationship between TV and Sales, Sales are surely increasing because of TV advertisement.

plt.figure(figsize **=** (10,7))

sns.heatmap(df.corr(), cmap **=** 'RdBu',linewidth**=**2,annot **= True**) plt.show()

# 

observation;

1. TV and radio has a strong relationship with sales
2. newspaper has not so much srongly correlated with sales
3. features has no strong relationship among themselves which is good for model building.

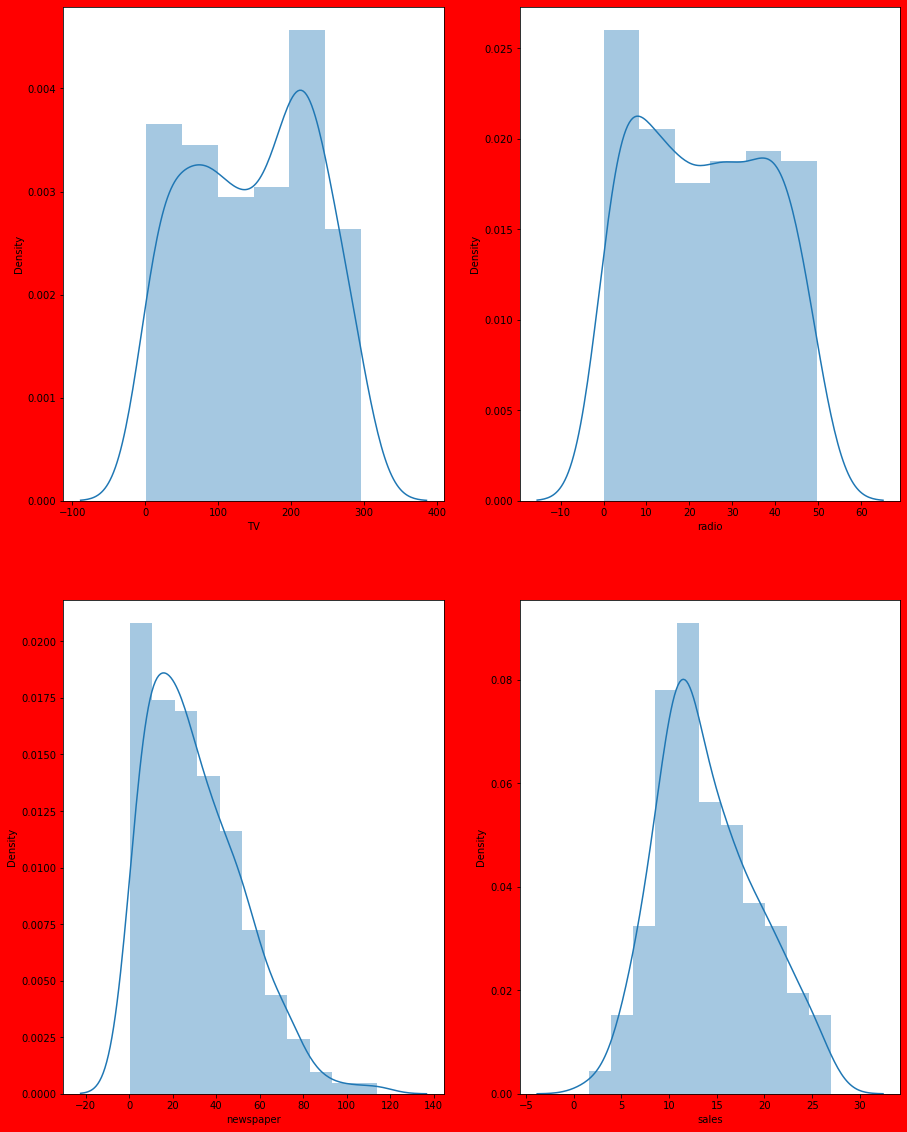
checking skewness of the data

df.skew()

|  |  |
| --- | --- |
| TV | -0.069853 |
| radio | 0.094175 |
| newspaper | 0.894720 |
| sales | 0.407571 |

dtype: float64

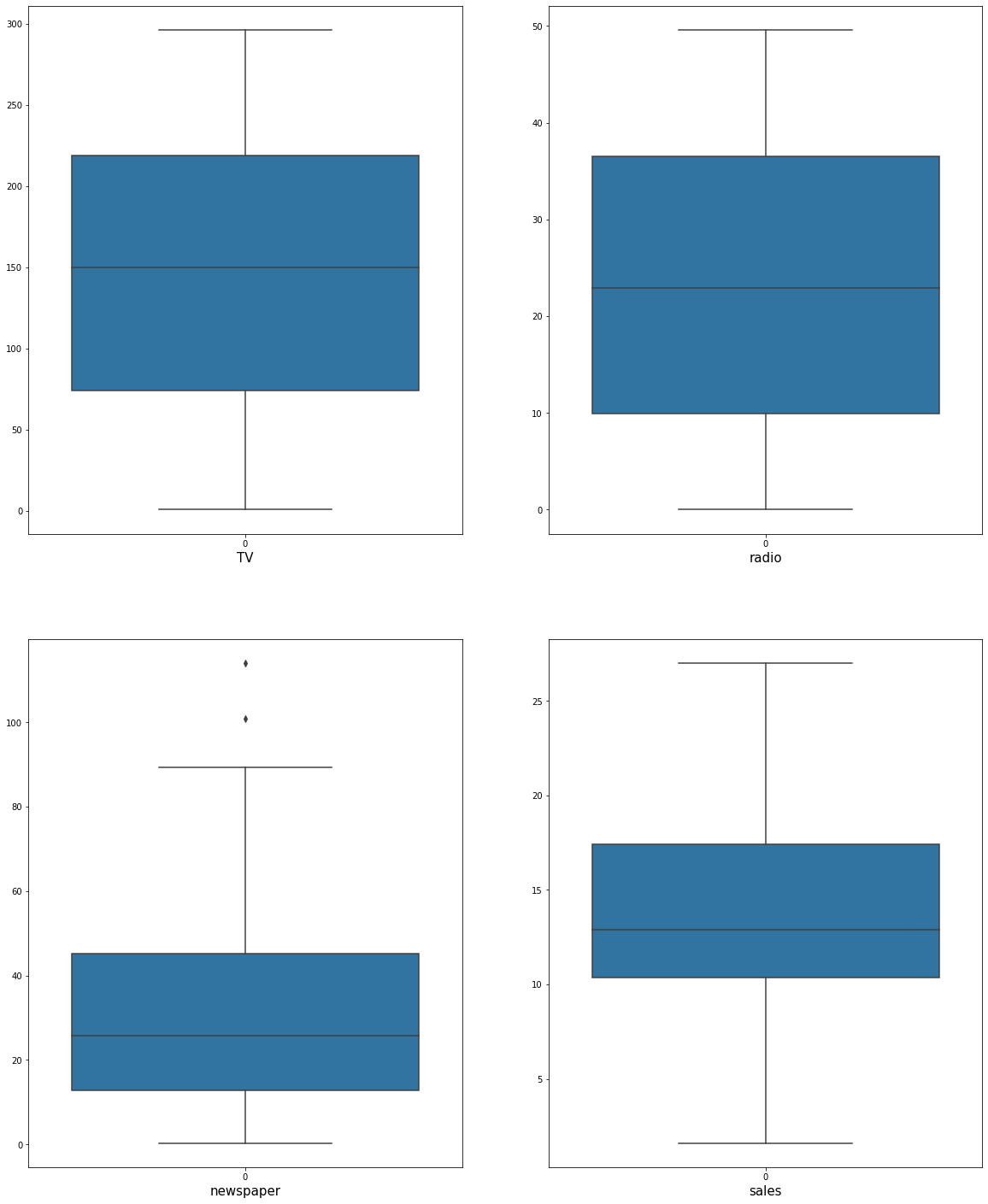




skewness looks good and hence strong evidence not been concluded so look for other plot to search skweness if any.

**Applying boxplot to check outliers**





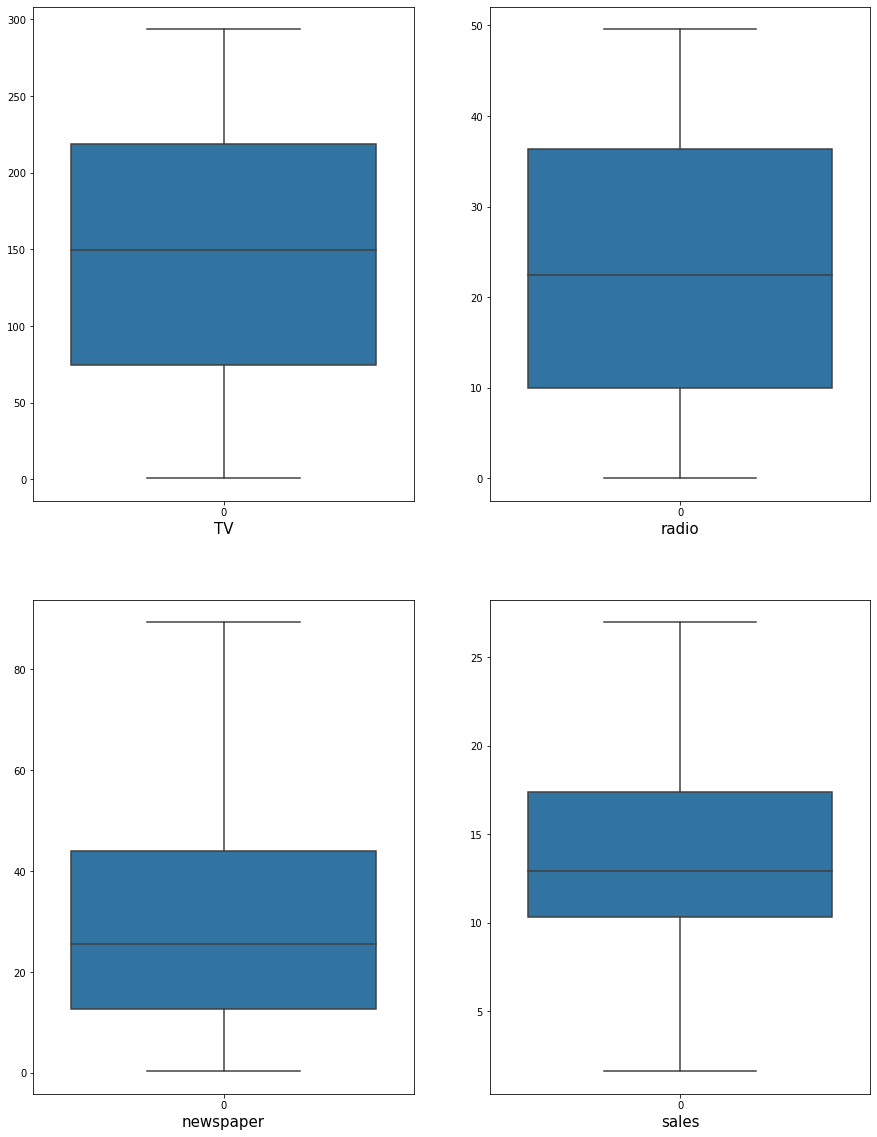
only news paper column has outliers which can be removed easily





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **index** | **TV** | **radio** | **newspaper** | **sales** |
| **0** 0 | 230.1 | 37.8 | 69.2 | 22.1 |
| **1** 1 | 44.5 | 39.3 | 45.1 | 10.4 |
| **2** 2 | 17.2 | 45.9 | 69.3 | 9.3 |
| **3** 3 | 151.5 | 41.3 | 58.5 | 18.5 |
| **4** 4 | 180.8 | 10.8 | 58.4 | 12.9 |



****

**outliers has been removed completely**

df.describe()

|  |  |  |  |
| --- | --- | --- | --- |
| **TV** | **radio** | **newspaper** | **sales** |
| **count** 198.000000 | 198.000000 | 198.000000 | 198.000000 |
| **mean** 146.688384 | 23.130808 | 29.777273 | 13.980808 |
| **std** 85.443221 | 14.862111 | 20.446303 | 5.196097 |
| **min** 0.700000 | 0.000000 | 0.300000 | 1.600000 |
| **25%** 74.800000 | 9.925000 | 12.650000 | 10.325000 |
| **50%** 149.750000 | 22.400000 | 25.600000 | 12.900000 |
| **75%** 218.475000 | 36.325000 | 44.050000 | 17.375000 |
| **max** 293.600000 | 49.600000 | 89.400000 | 27.000000 |



x.head ()

|  |  |  |
| --- | --- | --- |
| **TV** | **radio** | **newspaper** |
| **0** 230.1 | 37.8 | 69.2 |
| **1** 44.5 | 39.3 | 45.1 |
| **2** 17.2 | 45.9 | 69.3 |
| **3** 151.5 | 41.3 | 58.5 |
| **4** 180.8 | 10.8 | 58.4 |

y.head()

|  |  |
| --- | --- |
| 0 | 22.1 |
| 1 | 10.4 |
| 2 | 9.3 |
| 3 | 18.5 |
| 4 | 12.9 |

Name: sales, dtype: float64

. **scaling the dataset**

In [36]:

x\_scaled.shape

Out[36]:

(198, 3)

y.shape

Out[37]:

(198,)

**checking multicollinearity**



. **vif Features**

**0** 1.003863 TV

**1** 1.137887 radio

**2** 1.137631 newspaper

As we observe that vif factor for each columns is less than 5 so we can say that multicollinearity may not be present in our features



now our model is ready for training





RMSE Score is: 1.711290080827877 R2 score is: 90.72259879826814

Cross Validation Score is: 88.53605756683912

R2 - Cross Validation score is: 2.1865412314290182



. RMSE Score is: 0.7427964727972246 R2 score is: 98.25209132582104

Cross Validation Score is: 97.50064025197311

R2 - Cross Validation score is: 0.7514510738479316



RMSE Score is: 1.3627178724886526 R2 score is: 94.11710664288589

Cross Validation Score is: 95.15274905338408

R2 - Cross Validation score is: -1.0356424104981983

*#*

RMSE Score is: 1.4864588793505187 R2 score is: 93.00021225301826

Cross Validation Score is: 90.85884467177155

R2 - Cross Validation score is: 2.141367581246712

*#XGBOOST METHOD*

**import** xgboost **as** xgb



RMSE Score is: 0.7753030012610452 R2 score is: 98.09575860253759

Cross Validation Score is: 96.90907247774507

R2 - Cross Validation score is: 1.1866861247925158



RMSE Score is: 0.6768900168988874 R2 score is: 98.5485058497015

Cross Validation Score is: 97.80829328019823

R2 - Cross Validation score is: 0.7402125695032709



RMSE Score is: 1.0072505075582863 R2 score is: 96.78593939391807

Cross Validation Score is: 95.43424820049168

# AS WE OBSERVE THAT :

1. tbest result obtained from random forest REGRESSOR and Gradientboostingregressor as as there cross validation report is better than other models and also R2-cross validation score is also lower which is healthy sign for the model

FEATURES SELECTION FOR THE BEST MODEL

we have observe above that Tv and radio has a good and strong positive relationship with the target variables and newspaper has not so much strong correlation but since features are very less and newspaper also show good correlation although so we decide to keep all the features for model building

# HYPERTUNING BY USING GRIDSEARCHCV METHOD

**Applying on adaboost method for best result**



GridSearchCV(cv=5, estimator=AdaBoostRegressor(),

param\_grid={'learning\_rate': [0.001, 0.01, 0.1, 1],

'loss': ['linear', 'square', 'exponential'], 'n\_estimators': [50, 100, 150, 200],

'random\_state': [21, 42, 104, 111]})

grid.best\_params\_

{'learning\_rate': 1, 'loss': 'linear', 'n\_estimators': 200, 'random\_state': 21}



AdaBoostRegressor()



RMSE Score is: 0.8802900668735797 R2 score is: 97.54511769956919

Cross Validation Score is: 95.81770933126768

R2 - Cross Validation score is: 1.7274083683015107

Slight improvement in the results as compared to previous results

**Lets apply hyper tuning model in the Random forest**



{'criterion': 'mae', 'max\_features': 'auto'}



R2 score: 98.47256314923288

cross val score: 97.73480291636088

# random forest model provides the best result as evident from the score .lets save it.

**SAVING THE MODEL**



# conclusion



98.47256314923288

conclusion = pd.DataFrame([loaded\_model.predict(x\_test)[:],pred\_decision[:]],index = ['predicted','original'])

conclusion

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **...** |
| **predicted** 13.16 | 13.05 | 5.101 | 15.123 | 11.133 | 8.016 | 15.2975 | 11.422 | 12.163 | 12.449 | ... 14.39 |
| **original** 13.16 | 13.05 | 5.101 | 15.123 | 11.133 | 8.016 | 15.2975 | 11.422 | 12.163 | 12.449 | ... 14.39 |

2 rows × 40 columns

# Conclusion:

To predict which advertising platform is giving the best sales performance I have followed some EDA steps and visualizations.we have Checked correlation among features and removed outliers using Inter Quantile Range Method. After this removing skewness from our dataset as much as possible or else our prediction can will not be perfect. Next, I divided the dataset into X and Y variable which is X = features which will help in prediction and Y = target column which is our target variable. Next, I have scaled the x dataset using Standard Scaler. Standard scaler helps to scale the values so that it’s easy for our model to understand. I have checked for Multicollinearity problem using Variance Inflation Factor which helps us to understand if any feature is correlated to other feature, we should always check multicollinearity issue before building our model or else our model can overfit. However, in this dataset we do not have multicollinearity problem hence, we can go ahead and start building our model. I have built 7 models and RandomForest Regressor has given the best performance where RMSE Score is: 74%, R2 Score is 98% and Cross Validation score is 75%. Then I did feature selection for my best model and removed newspaper. Then Hypertuned the parameters using GridSearchCV however my R2 score for RandomForest model was 96.78% hence saved the original model using pickle and later I have loaded the saved model and made prediction. My model did a good job in terms of prediction as test R2 score was 98% which indicates no overfitting .Thus our model is completed .