**Advertising Sales Channel Prediction**

Introduction

We have media as an important source of communication. The advertising industry is large and have lager contribution in sales for majority of industries. 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.

We have data set of advertising channels TV Radio, Newspaper and sales if our target.

Analysis

| **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 |
| **...** | ... | ... | ... | ... | ... |
| **195** | 196 | 38.2 | 3.7 | 13.8 | 7.6 |
| **196** | 197 | 94.2 | 4.9 | 8.1 | 9.7 |
| **197** | 198 | 177.0 | 9.3 | 6.4 | 12.8 |
| **198** | 199 | 283.6 | 42.0 | 66.2 | 25.5 |
| **199** | 200 | 232.1 | 8.6 | 8.7 | 13.4 |

200 rows × 5 columns

We have data set of dimension 200 rows and 5 columns. Column names are Unnamed, TV, Radio, Newspaper and sales.

Unnamed: 0 int64

TV float64

radio float64

newspaper float64

sales float64

dtype: object

The data set is of float data type.

| **Unnamed: 0** | **TV** | **radio** | **newspaper** | **sales** |
| --- | --- | --- | --- | --- |
| **count** | 200.000000 | 200.000000 | 200.000000 | 200.000000 | 200.000000 |
| **mean** | 100.500000 | 147.042500 | 23.264000 | 30.554000 | 14.022500 |
| **std** | 57.879185 | 85.854236 | 14.846809 | 21.778621 | 5.217457 |
| **min** | 1.000000 | 0.700000 | 0.000000 | 0.300000 | 1.600000 |
| **25%** | 50.750000 | 74.375000 | 9.975000 | 12.750000 | 10.375000 |
| **50%** | 100.500000 | 149.750000 | 22.900000 | 25.750000 | 12.900000 |
| **75%** | 150.250000 | 218.825000 | 36.525000 | 45.100000 | 17.400000 |
| **max** | 200.000000 | 296.400000 | 49.600000 | 114.000000 | 27.000000 |

Data description

From the above data we could see the data is showing the count of people using the media for entertainment and Companies are using these means to advertise there various product to bring in notice of people. Many companies are trying to use social media platform as well.

Unnamed: 0 0

TV 0

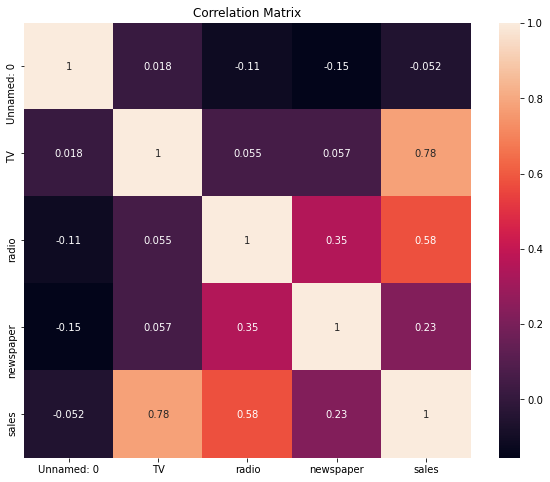
radio 0

newspaper 0

sales 0

dtype: int64

We observe that we do not have any null values. Hence we will not be treating any null values further.



We observe that TV is having positive direct relationship with the sales, we understand that TV contribute significantly in sales of product followed by Radio and Newspaper. The unnamed column is of no use hence we will drop it further.

Unnamed: 0 AxesSubplot(0.125,0.536818;0.110714x0.343182)

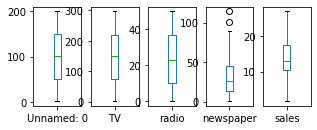
TV AxesSubplot(0.257857,0.536818;0.110714x0.343182)

radio AxesSubplot(0.390714,0.536818;0.110714x0.343182)

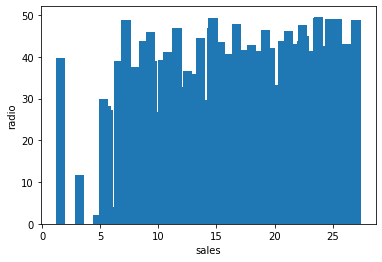
newspaper AxesSubplot(0.523571,0.536818;0.110714x0.343182)

sales AxesSubplot(0.656429,0.536818;0.110714x0.343182)

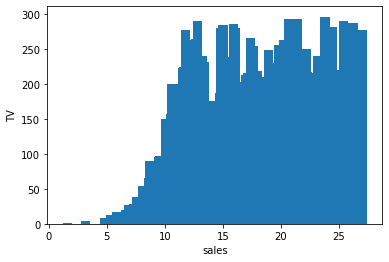
dtype: object



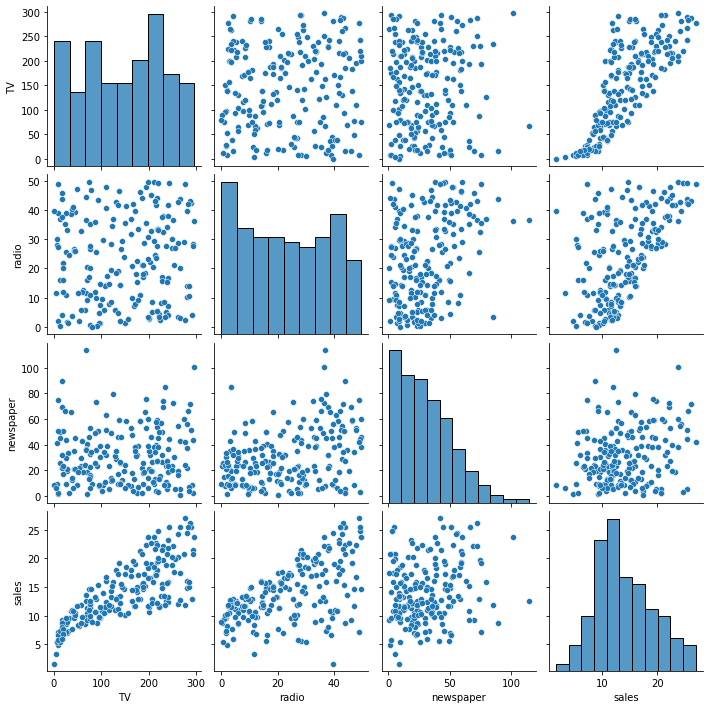
We have outliers in Newspaper which we will treated with Z score.



Sales vs radio



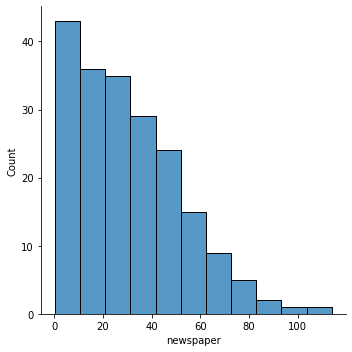
Sales vs TV

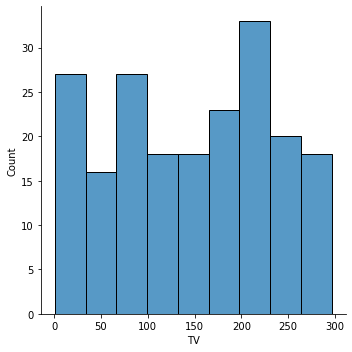


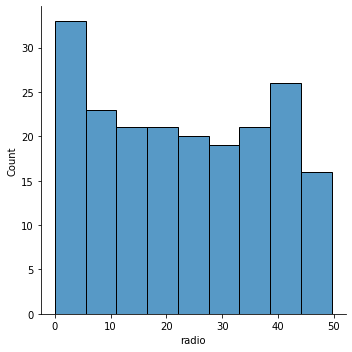
Pair plot results

The pair plot have the information of the data and we observe that there is positive relationship, skewness information.

Depending upon the information we can perform various other operation to clean or treat the data.







Above is the bar graph that states the count of respective feature.

We observe that there is an skewness and this will be manged by using Yeo-Johnson method.

Observations - from skewness and outliers we can see that the Newspaper contains both.

Here we will be losing 25% of data if we remove the newspaper data or column which will impact our model and prediction further, hence we will not drop the column.

Libraries

As we are getting the continuous values we are using the Regression model for this dataset.

import numpy as np

import pandas as pd

import sklearn

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler

from sklearn.preprocessing import power\_transform

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

from sklearn.linear\_model import LinearRegression

from sklearn.model\_selection import train\_test\_split, cross\_val\_score

from sklearn.metrics import r2\_score, mean\_squared\_error

from sklearn.metrics import mean\_absolute\_error

from sklearn.linear\_model import Lasso,Ridge

from sklearn.linear\_model import ElasticNet

from sklearn.svm import SVR

import pickle

Above are the libraries that we have used in this project.

Model

We are using the regression techniques to build the model. The regression model allows us to evaluate various other parameters like root mean squared error, cross validation etc. However we will be evaluating the efficiency of the model based on r2 score.

We have achieved the linear regression model as best model and below are the predicted result of the same.

Here our target variable was sales.

Below are the predicted results:

Prediction [12.26547423 19.87256169 9.26992787 22.34448115 9.68454469 12.3232653

13.2190952 18.70120637 15.33483148 15.16183262 20.6621318 15.42167888

2.13852152 21.15941233 17.57667882 15.56201986 17.10886845 15.14178274

10.80056928 10.24177051 12.91556785 19.42556005 20.45641538 22.3147705

15.13065066 12.24833378 15.92951482 17.07300743 11.75441489 5.01229823

16.16283949 22.44075966 13.99300042 15.83751172 16.56560457 11.2115412

20.42837974 11.64899827 19.77417125 10.18856043]

Actual value 31 11.9

176 20.2

143 10.4

61 24.2

158 7.3

9 10.6

135 11.6

184 17.6

123 15.2

178 11.8

69 22.3

26 15.0

8 4.8

47 23.2

74 17.0

162 14.9

99 17.2

110 13.4

151 11.6

38 10.1

179 12.6

28 18.9

84 21.7

36 25.4

146 13.2

1 10.4

87 16.0

62 15.7

148 10.9

195 7.6

118 15.9

147 25.4

161 13.3

45 14.9

194 17.3

75 8.7

52 22.6

164 11.9

53 21.2

90 11.2

Name: sales, dtype: float64

We are considering the r2 score.

r2 score -> coefficient of determination is the proportion of the variance in the dependent variable that is predictable from the independent variable (s) i.e change coming in y whenever x is being changed

Our linear regression model is 90.05% precise and we will go with this model. Make sure to save the model using joblib. Saving the model is very important.

Conclusion

Overall observation is that all the factors are critical and influence the fees of consultation. The Linear regression is the best model after comparing the r2 score and cross-validation difference with Random forest and lasso.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***