

COTM AUGUST

By Krish Yadav

Video link : <https://www.youtube.com/watch?v=XOAlp92hgo>

Code link: https://colab.research.google.com/drive/luNdH_ydOeNKE7qPsJRCmgLwnoeblv9Om

Dataset

Finding the most suitable dataset for the purpose, was a very daunting task, after weeks of searching with various keywords I was able to find the appropriate dataset.

RangeIndex: 4572 entries, 0 to 4571

Data columns (total 5 columns):

#	Column	Non-Null Count	Dtype
0	TV	4562 non-null	float64
1	Radio	4568 non-null	float64
2	Social Media	4566 non-null	float64
3	Influencer	4572 non-null	object
4	Sales	4566 non-null	float64

dtypes: float64(4), object(1)

memory usage: 178.7+ KB

All columns are regarding promotion, and data figures are in millions.

Data preparation

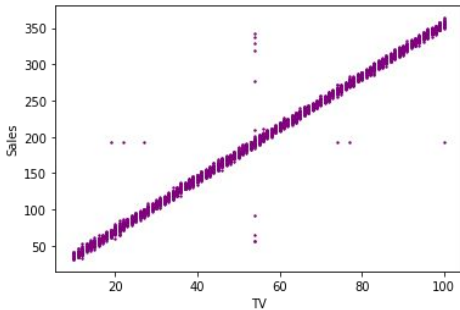
The data also contained missing values, which were filled using mean of other values in the column. The categorical column was dropped, since we are doing linear regression and which requires data of type continuous.

Linear regression

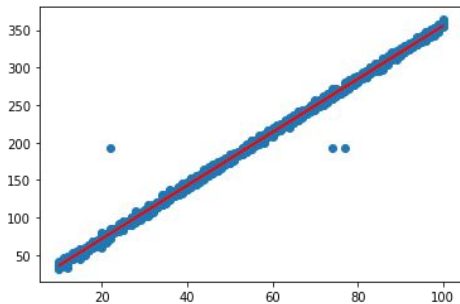
After all the processing, I did the Linear regression pairwise between remaining col and sales as visible in figures below

TV

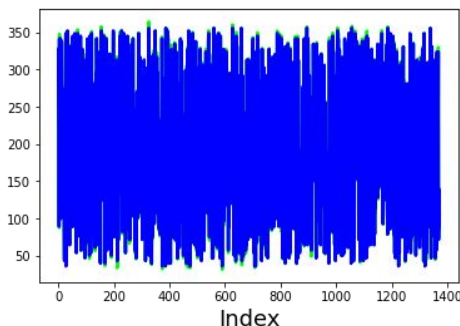
Scatter plot of TV in respect to Sales



TV to Sales linear regression
MSE Score: 26.91850181656196

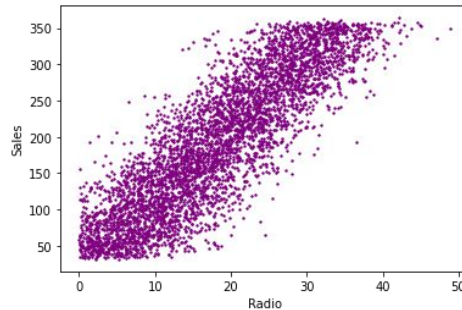


Actual and Predicted

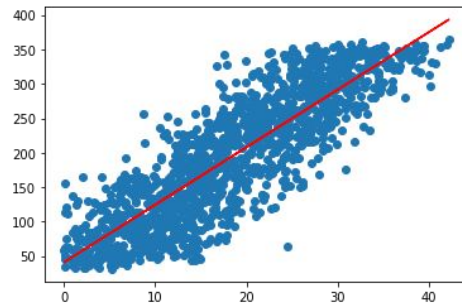


RADIO

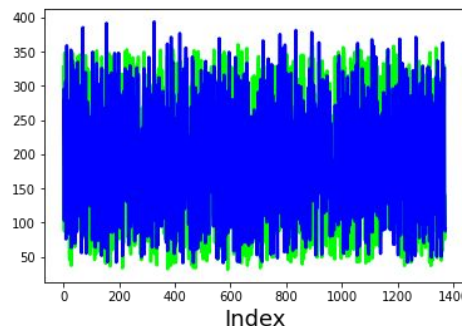
Scatter plot of Radio in respect to Sales



Radio to Sales linear regression
MSE Score: 2164.776450231296

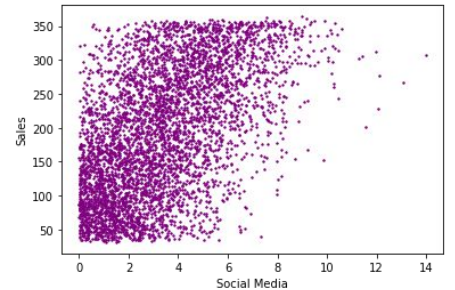


Actual and Predicted

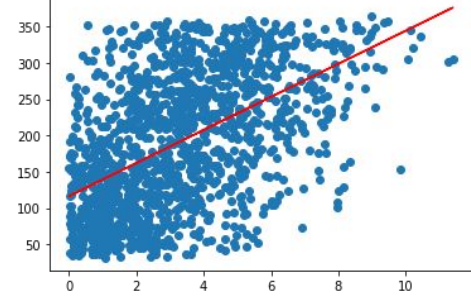


SOCIAL MEDIA

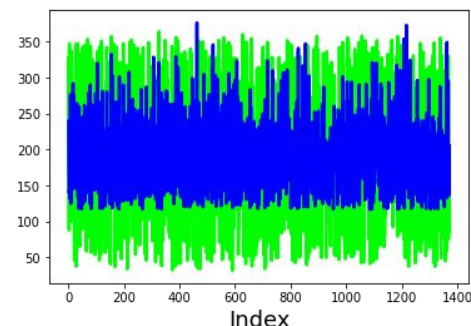
Scatter plot of TV in respect to Sales



Social Media to Sales linear regression
MSE Score: 6375.375397912058



Actual and Predicted



XGBoost

I also tried using one of the best models in AI for tabular data called XGBoost. Unlike other models, the XGBoost was trained on all the three cols. It gave a MSE of **31.271749**.

Trends



TV

In the data we can clearly see that TV is extremely linearly related to sales, although there are these crosshair type outliers present in the TV column. As the points were at a close distance to each other, the linear regression model performed very well with a MAE of **27**

Since 2-14 age group (As per statista.com) sees the TV most, it reflects we need Kids our using our product



Radio

The impact of radio on sales was dense, but as of TV. The distance of points from any central line was larger, yielding to larger MAE score: **2164**.


Radio is more used in villages and by senior citizens, which indicates, they too influence the sales



Social Media

Social media in low budgets can definitely give proportional sales, as there is chunk of values in bottom left corner of graph. Also the values become sparse when we go towards the top right corner. Due this regions the regression model gave a MAE of **6375**.

The main factors can be too many ads occur on social media, which users ignore, second our product might not required by teenagers and working class



Choose your model

Model:

TV:

Radio:

Social_Media:

[Show code](#)

Probable sales is : 875.5022907111027

Predictor

I also created a predictor where you can choose the model, input the values and then get a probable value of sales.