**Name: MD Akif Perwez**

**B.Tech (Mechanical Engineering)**

**2+ years of experience as Technical Sales Engineer**

**PGP in Data Science, Machine Learning and Neural Networks from Data Trained in collaboration with IBM.**

**Blog**

**Advertising Sales Prediction**

**Problem Definition**

**Problem Statement:**

* I got the data set from the GitHub and the data set has some unwanted data.

**Impact:**

Lack of outliers and skewness are presents, which may not give us the best r2 score.

In this project, I have built and evaluated multiple linear regression models using python. I have used sciket-learn to calculate the regression, while using pandas for data management and matplotlib and seaborn as visualization. The data set for this project consists of the very popular Advertising spending through media such as TV, radio, and newspaper.

**Project Overview**

Objective: To predict sales for given budget spend on TV, Radio and Newspaper in dollars. Regression Problem Data cleaning and Data preprocessing Exploratory Data Analysis Multiple Linear Regression model training andprediction

**Atrributes are:**

* **TV**
* **Radio**
* **Newspaper**
* **Sales**

**Sales is my Target Variable.**

**Data Analysis:** I applied the concepts of Machine learning and Data Science to predict the Advertising Sales Channel Prediction.

The target variable is of continuous type (Sales), which simply define that the problem is of linear regression type.

**Libraries which are used for the Analysis are:**

* **Pandas**
* **Numpy**
* **Matplotlib**
* **Seaborn**
* **Scikit learn**
* **Scipy**
* **joblib**

While doing the Analysis part I have to check the shape of the column.

There are 200 rows and 4 columns.

In this data set all the columns are of float types. So we don’t have to apply any method to change the types.

**EDA (Exploratory Data Analysis)**

In statistics exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods**.**

While doing the Analysis, firstly we have to check is there any null values are present in the rows?

In this case, no null values are present in the data set.

I checked the null values by Visualization and simply by writing code df.isnull().sum()

**Checking Statistic Description.**

Statistic Description gives us the statistic over view the numeric columns.

As in this case all the columns has numeric values only, so we found that Skewness and Outliers are present in the data set.

I do visualization of both skewness and outliers in next step.

**Uni-Variate Analysis**

Checking with the bar plot with indexing of each columns, so that to optimize the maximum sales as per given data.

I found that “TV” has the maximum number of sales every time.

**Bi-Variate Analysis**

* By using Scatter plot, I found that advertising dollars spent on TV, Sales also increases.
* By using Scatter plot, I found that advertising dollarsspent on Radio, Sales also increases but less incomparison with TV.
* By using Scatter plot, I found that advertising dollars spent on Newspaper, Sales decreases, it means sales is inversely proportional to Advertising**.**

**Conclusion:**

From the above visualization, I got a point that,

We can spend advertising amount in TV and Radio but not in Newspaper.

Or it’s better to spend more on TV rather than Radio.

**Multi-Variate Analysis**

Target Sales is directly correlated with TV, little in Radio but there is no correlation with Newspaper. There is no correlation seems between all features.

**Correlation**

In statistics, correlation or dependence is any statistical relationship, whether casual or not, between two random variables or bivariate data.

As visible from the pairplot and the heatmap, the variable TV seems to be most correlated with sales

**Data Cleansing**

Outliers Detection by using boxplot: Outliers and nothing but some unnecessary data present in the data set.

As far as execute the code, I found that outliers are present in the “Newspaper” column only.

Firstly I used IQR method to remove the outliers. I found that data loss is 1% from the data set, which is obviously a good sign from data cleansing.

* Skewness: Skewness is measure by the symmetry of distribution. The highest Point of a distribution is its mode. The mode is on the x-axis.

Skewness are of two types Right Skewness and Left Skewness.

In the data set, skewness is present only in “Newspaper” column.

**Conclusion:**

Once I removed the outliers with IQR method, Automatically Skewness is removed and hence the data is ready for next step that is our model building with multiple method.

**Train Test Split**

Before building the model, we have to do train-test-split, that our target variables are in y group and remaining columns are in x group.

X=df.iloc [:,:-1]

Y=df.iloc[:,-1]

**Finding Best Random State**

After splitting the model, will find the best random state to put those x and y values and build our model.

In my case I got best random state in 69 with 93% accuracy score.

Now we do train\_test\_split and give the test size that this much data go into test and will do prediction on the model building.

I put 33% data into test and remaining 67% into training.

As from the Analysis I already mentioned that my problem is our Linear Regression type and we will use only Regressor model to build.

**Here I used five method for model building**

1. Linear Regression.
2. RandomForestRegressor.
3. SVR.
4. AdaboostRegressor.
5. KNeighborsRegressor.

**While making a model we need to follow few steps:**

Suppose x is any model, we will assign this x with any variable suppose Y is the variable.

**Now let’s follow the step:**

* Y= x,where x is the model and y is any variable.
* Y=x.fit(x\_train,y\_train), here we are fitting those values of x and y. As already mentioned above that training part should 67% data. So 67% data will go in it.
* Pred = Y.predict(x\_test), here we are predicting the remaining 33% data.

After all these process, now we can get the score from ‘y\_test’.

* Print(r2\_score(y\_test,pred)

**Output:**

After following the above steps, I got the score of every model.

* In Linear Regression I got the r2 score of **92%**
* In RandomForestRegressor I got the r2 score of **98%**
* In SVR I got the r2 score of **79%**
* In AdaBoostRegressor I got the r2 score of **96%**
* In KNeighborsRegressor I got the r2 score of **89%**

By seeing the above accuracy score still we are not sure that RandomForestregressor has maximum Accuracy Score and is suitable for the final model.

**We will to do Cross Validation for all the above model.**

Cross validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a simple parameters called ‘k’ that refers to the number of groups that a give data sample is to be split into. As such, the procedure is often called k-fold cross validation.

* **I used k=5 here**

**After Cross Validation I found that**

* **LR=88% r2 score**
* **RF=97% r2 score**
* **SVR=84% r2 score**
* **AdaBoost=95% r2 score**
* **KN=90% r2 score**

From the above r2 score in both of them, now we have to check the minimum difference.

So I got minimum difference i.e. 1 in RandomForestRegresssor and AdaBoostRegressor.

Now will do the Hyper Parameter Tuning with both of them.

**Hyper Parameter Tuning:**

In machine learning hyper parameter is a parameter whose value is set before the learning process.

It help us to increase the Accuracy score for any model we are making.

For this i used “GridSearchCV”.

Now in hyper parameter tuning we have to define the parameter first in order to check which parameter is the best one.

**Hyper Parameter Tuning With RandomForestRegressor**

I used “estimator’s” and “loss” parameters.

No doubt there are many more parameter to use. But we did not use each and every parameter as the execution will take time more than 7-8hrs or may be more. Which is obviously! Not a good idea to complete a project.

Here I got the best para in n\_estimators:10, loss=exponential.

We used this parameter to check the accuracy score and build our model.

**Steps do to for hyper parameter tuning**

Creating parameter list to pass the in GridSearchCv

* parameters= {"n\_estimators":(x), where x is any integer values. "criterion":["mse","mae"]}
* GCV=GridSearchCV(RandomForestRegressor(),parametrs,cv=5), here GCV is any variable we can use anything as per our ease.
* GCV.fit(x\_train,y\_train), passing the values of x and y.
* GCV.best\_params\_ (This code will give the best parameter from the above)
* rf\_mod=RandomForestRegressor(criterion="mae", n\_estimators=12), so here I got ‘mae’ and ‘12’ as best parameter which will give me the maximum r2 score. **rf\_mod** is the model name.
* rf\_mod.fit(x\_train,y\_train)
* pred=rf\_mod.predict(x\_test)
* print(r2\_score(y\_test,pred)\*100)

**By doing this I got 95% r2\_score.**

**Hyper Parameter With AdaBoostRegressor**

**Same step has to be followed for AdaboostRegressor**

I used “n\_neighbors” and “algorithm” parameters.

Here I got the best para in n\_neighbors:2, algorithm=auto.

Here my model name is **ad\_mod**

**By doing this I got 88% r2\_score.**

Sometimes, the r2 score will as same before the hyper parameter tuning, it doesn’t mean that your model is bad.

We can use this model for prediction.

**Conclusion:**

It was a wonderful and learning experience for me while working on this project. This project took me through various phases of project development and gave me real insight into the world of Machine Learning and Data Scientist. The joy of work and the thrill involved while tackling the various problems and challenges gave me a feel of Data Scientist.

After Hyper Parameter Tuning I got two r2 score which is 95% and 88% for RandomForestRegressor and AdaBoostRegressor respectively. Now I have to choose one of them.

As the r2 score already define that 95% is best here to make the model.

**Saving The Model**

Joblib.dump(rf\_mod,”Advertising.pkl”)

Note: here rf\_mod is my model name

My purpose of choosing this project is to observe how we can spend money on ads. Because ads plays most vital role in order to increase our business sales.

By doing all the analysis I come to a point that we are moving forward technologies and we may see our future in it.

So why not we spend our money to airing our ads on YouTube, Facebook and Instagram and so on.

As social media is the biggest platform for each and everything.

**https://github.com/akifperwez/Advertising-Sales-Report/blob/main/Advertising%20Sales%20Project.ipynb**