### **Case Details : Linear Regression**

|  |  |
| --- | --- |
| Problem Statement | It is a dataset containing the impact of three advertising medias (youtube, facebook and newspaper) on sales. The first three columns are the advertising budget in thousands of dollars along with the fourth column as sales. The advertising experiment has been repeated 200 times. Hence, it has 200 rows. |
| Objectives | Perform Multiple Linear Regression |
| Research Qs | Is sales related to other variables - youtube, facebook, newspaper  Does the model exist ? |
| Tasks | * Name the dataframe : marketing * Perform Linear regression model that can be used to predict sales by establishing a statistically significant linear relationship with other variable * #also explain the model output, draw boxplot, density and histogram to understand the data. predict sales for any unknown value of combination of other variables * R2, Adjt R2, RMSE |

### **DataSet**

|  |
| --- |
| Data Description |
| Marketing Data   * A data frame containing the impact of three advertising medias (youtube, facebook and newspaper) on sales. Data are the advertising budget in thousands of dollars along with the sales. The advertising experiment has been repeated 200 times. |

|  |  |
| --- | --- |
| R | data('marketing', package='datarium') |
| Python | Use this link <https://raw.githubusercontent.com/DUanalytics/datasets/master/R/marketing.csv> |

### 

### **Results**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Description** | **Value** |
| Name of IV selected | Based on p-values of P(t-stats), each variable to be selected |  |
| MAE (Mean absolute error) | represents the difference between the original and predicted values. We get this number by averaging the absolute difference over the data set. |  |
| MSE (Mean Squared Error) | represents the difference between the original and predicted values extracted by averaging the squared difference over the data set. |  |
| RMSE (Root Mean Squared Error) | is the square root of the the arithmetic mean of the squares of difference over the set. |  |
| R-squared (Coefficient of determination)  Adjusted R-squared for Multiple Linear Regression | represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.  measures the strength of the relationship between your model and the dependent variable |  |
| F Statistics (p-value) | If the P value for the F-test of overall significance test is less than your significance level, you can reject the null-hypothesis and conclude that your model provides a better fit than the intercept-only model. |  |
| AIC (Akaike information criterion) | AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth |  |

### **Solution in Python**

|  |
| --- |
| #%% Linear Regression -1 Marketing Data - Sales - YT, FB, print  #libraries  import numpy as np  import pandas as pd  from sklearn.model\_selection import train\_test\_split  from sklearn.metrics import mean\_squared\_error, r2\_score  from sklearn import linear\_model #1st method  import statsmodels.api as sm #2nd method  import matplotlib.pyplot as plt  import seaborn as sns  url ='https://raw.githubusercontent.com/DUanalytics/datasets/master/R/marketing.csv'  marketing = pd.read\_csv(url)  marketing.head()  #describe data  #visualise few plots to check correlation  #split data into train and test  #build the model  #predict on test values  #find metrics - R2, Adjt R2, RMSE, MAPE etc  #predict on new value  newdata = pd.DataFrame({'youtube':[50,60,70], 'facebook':[20, 30, 40], 'newspaper':[70,75,80]})  newdata  #your ans should be close to [ 9.51, 11.85, 14.18]  #conclude by few lines |