Prepare a prediction model for profit of 50\_startups data.

Do transformations for getting better predictions of profit and

make a table containing R^2 value for each prepared model.

BUSSINESS PROBLEM: To predict a model for 50 start-up data.

|  |  |  |
| --- | --- | --- |
| R & D spent:  HISTOGRAM | Data is right skewed | |
| BOX-PLOT | No outliers ,and more data is present in upper whisker | |
| QQPLOT | Shows data is normally distributed | |
| skewness(start\_up$R.D.Spend)  [1] 0.1590405 right skewed because value is gre  > kurtosis(start\_up$R.D.Spend)  [1] 2.194932 | | right skewed because value is greater than 0  leptokurtic, due to its peakedness and data is more distributed in  tails |

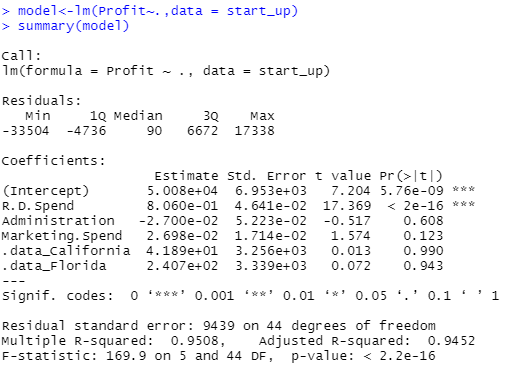
|  |  |  |
| --- | --- | --- |
| ADMINISTRATION SPENT:  HISTOGRAM | Data is left skewed | |
| BOX-PLOT | No outliers ,and more data is present in lower whisker, and its left skewed | |
| QQPLOT | Data is normally distributed | |
| skewness(start\_up$Administration)  [1] -0.4742301  > kurtosis(start\_up$Administration)  [1] 3.085538 | | left skewed because value is less than 0  leptokurtic, due to its peakedness and data is more distributed in  tails |

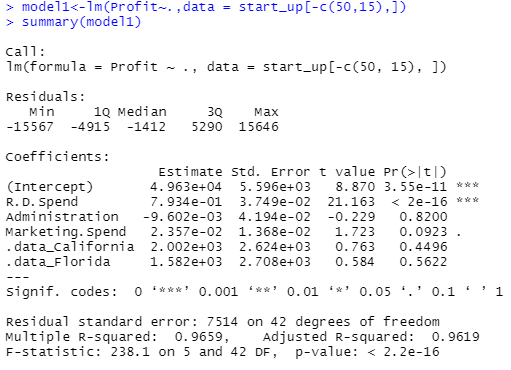
|  |  |  |
| --- | --- | --- |
| MARKETING SPENT:  HISTOGRAM | Data is Normally distributed | |
| BOX-PLOT | No outliers ,and it has more variance in upper whisker | |
| QQPLOT | Data is normally distributed | |
| skewness(start\_up$Administration)  [1] -0.4742301  > kurtosis(start\_up$Administration)  [1] 3.085538 | | left skewed because value is less than 0  leptokurtic, due to its peakedness and data is more distributed in  tails |

|  |  |  |
| --- | --- | --- |
| PROFITS:  HISTOGRAM | Data is left skewed | |
| BOX-PLOT | No outliers ,and it has more variance in lower whisker and data is left skewed | |
| QQPLOT | Data is normally distributed | |
| skewness(start\_up$Profit)  [1] 0.02258638  > kurtosis(start\_up$Profit)  [1] 2.824704 | | left skewed because value is less than 0  leptokurtic and data is more distributed in tails |

MULTI-VARIET ANALYSIS

|  |
| --- |
|  |
| cor(start\_up)  R.D.Spend Administration Marketing.Spend  R.D.Spend 1.0000000 0.24195525 0.72424813  Administration 0.2419552 1.00000000 -0.03215388  Marketing.Spend 0.7242481 -0.03215388 1.00000000  Profit 0.9729005 0.20071657 0.74776572  .data\_California -0.1431652 -0.01547811 -0.16887523  .data\_Florida 0.1057111 0.01049309 0.20568545  Profit .data\_California .data\_Florida  R.D.Spend 0.9729005 -0.14316522 0.10571106  Administration 0.2007166 -0.01547811 0.01049309  Marketing.Spend 0.7477657 -0.16887523 0.20568545  Profit 1.0000000 -0.14583704 0.11624426  .data\_California -0.1458370 1.00000000 -0.49236596  .data\_Florida 0.1162443 -0.49236596 1.00000000 |
| cor2pcor(cor(start\_up))  [,1] [,2] [,3] [,4]  [1,] 1.00000000 0.21017626 0.0451801936 0.934189530  [2,] 0.21017626 1.00000000 -0.2853240281 -0.077706761  [3,] 0.04518019 -0.28532403 1.0000000000 0.230863131  [4,] 0.93418953 -0.07770676 0.2308631311 1.000000000  [5,] -0.02652716 0.01339520 -0.0009306794 0.001939379  [6,] -0.04595612 0.04725638 0.1622276251 0.010866323  [,5] [,6]  [1,] -0.0265271585 -0.04595612  [2,] 0.0133951994 0.04725638  [3,] -0.0009306794 0.16222763  [4,] 0.0019393790 0.01086632  [5,] 1.0000000000 -0.47766837  [6,] -0.4776683674 1.00000000 |
| INFERENCE: R & D and profit are strongly correlated positively, profit & administration are weakly correlated positively, profit & marketing spent are weakly correlated after influence of other input is paralysed by using partial cor-coef,  MODEL-BUILDING |





R AND D SHOWS MORE SIGFICIANCE

LINE ASSUMPTIONS of model2

|  |  |
| --- | --- |
|  | Residuals are dependent of each other,  (no auto-correlation )  And residuals are symmetrically distributed, hence homoscedasticity in nature |
|  | Residuals are normally distributed |
|  | VIF values are <10,hence the input variables are independent of each other  ( No multi-collinearity) |
|  | Residuals have equal variance |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MODEL | MULTIPL R2 | ADJUSTED R2 | TRAIN ERROR | TEST ERROR | INFERENCE |
| Model2 | 0.96 | 0.96 | 6838.876 | 7984.531 | Over-fitted/variance |

PYTHON CODE

import pandas as pd

import numpy as np

import statsmodels.formula.api as smf

import statsmodels.api as sm

import statsmodels as sm

import seaborn as sns

import matplotlib.pylab as plt

import pylab

import scipy.stats as stats

dataset=pd.read\_csv("C:/Users/USER/Desktop/50\_Startups.csv")

dataset.columns="R\_D\_SPEND","ADMINISTRATION","MARKETING\_SPEND","STATE","PROFIT"

dummies=pd.get\_dummies(dataset.STATE)

merge=pd.concat([dummies,dataset],axis='columns')

final=merge.drop(['STATE','Florida'],axis='columns')

final.columns="CALIFORNIA","NEWYORK","R\_D\_SPEND","ADMINISTRATION","MARKETING\_SPEND","PROFIT"

######univarient analysis#########

#####r&d

plt.hist(final.R\_D\_SPEND)

plt.boxplot(final.R\_D\_SPEND)

measurements = np.random.normal(loc = 20, scale = 5, size=100)

stats.probplot(final.R\_D\_SPEND, dist="norm", plot=pylab)

pylab.show()

######administration######

plt.hist(final.ADMINISTRATION)

plt.boxplot(final.ADMINISTRATION)

measurements = np.random.normal(loc = 20, scale = 5, size=100)

stats.probplot(final.ADMINISTRATION, dist="norm", plot=pylab)

pylab.show()

#######marketing spent#######

plt.hist(final.MARKETING\_SPEND)

plt.boxplot(final.MARKETING\_SPEND)

measurements = np.random.normal(loc = 20, scale = 5, size=100)

stats.probplot(final.MARKETING\_SPEND, dist="norm", plot=pylab)

pylab.show()

########profits#############

plt.hist(final.PROFIT)

plt.boxplot(final.PROFIT)

measurements = np.random.normal(loc = 20, scale = 5, size=100)

stats.probplot(final.PROFIT, dist="norm", plot=pylab)

pylab.show()

########multi-variet analysis########

sns.pairplot(final)

np.corrcoef(final)

corr = final.dropna().corr()

#######model1########

#multi-collinearity check########

model1=smf.ols('PROFIT~CALIFORNIA+NEWYORK+R\_D\_SPEND+ADMINISTRATION+MARKETING\_SPEND',data=final).fit()

model1.summary()

#####vif model#######

vif\_model=smf.ols('MARKETING\_SPEND~CALIFORNIA+NEWYORK+R\_D\_SPEND+ADMINISTRATION',data=final).fit()

vif\_model.summary()

vif\_MARKETING\_SPEND=1/(1-(0.59\*0.59))##vif=1.53

vif\_model=smf.ols('R\_D\_SPEND~CALIFORNIA+NEWYORK+MARKETING\_SPEND+ADMINISTRATION',data=final).fit()

vif\_model.summary()

vif\_R\_D\_SPEND=1/(1-(0.59\*0.59))###1.53

vif\_model=smf.ols('ADMINISTRATION~CALIFORNIA+NEWYORK+MARKETING\_SPEND+R\_D\_SPEND',data=final).fit()

vif\_model.summary()

vif\_ADMINISTRATION=1/(1-(0.15\*0.15))###1.02

############influentail factor check

sm.graphics.influence\_plot(model1)

plt.show()

#######model2###########

final1=final.drop(final.index[49])###record-49 is more influential

model2=smf.ols('PROFIT~CALIFORNIA+NEWYORK+R\_D\_SPEND+ADMINISTRATION+MARKETING\_SPEND',data=final1).fit()

model2.summary()

#######checking for train test error

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

final1\_train,final1\_test=train\_test\_split(final1,test\_size=0.2)

pred=model2.predict(final1\_test)

err=pred-final1\_test.PROFIT

mse=np.mean(err\*err)

rmse=np.sqrt(mse)

rmse

pred=model2.predict(final1\_train)

err=pred-final1\_train.PROFIT

mse=np.mean(err\*err)

rmse=np.sqrt(mse)

rmse