1. 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.

RnD admin marketing profit

count 50.000000 50.000000 50.000000 50.000000

mean 73721.615600 121344.639600 211025.097800 112012.639200

std 45902.256482 28017.802755 122290.310726 40306.180338

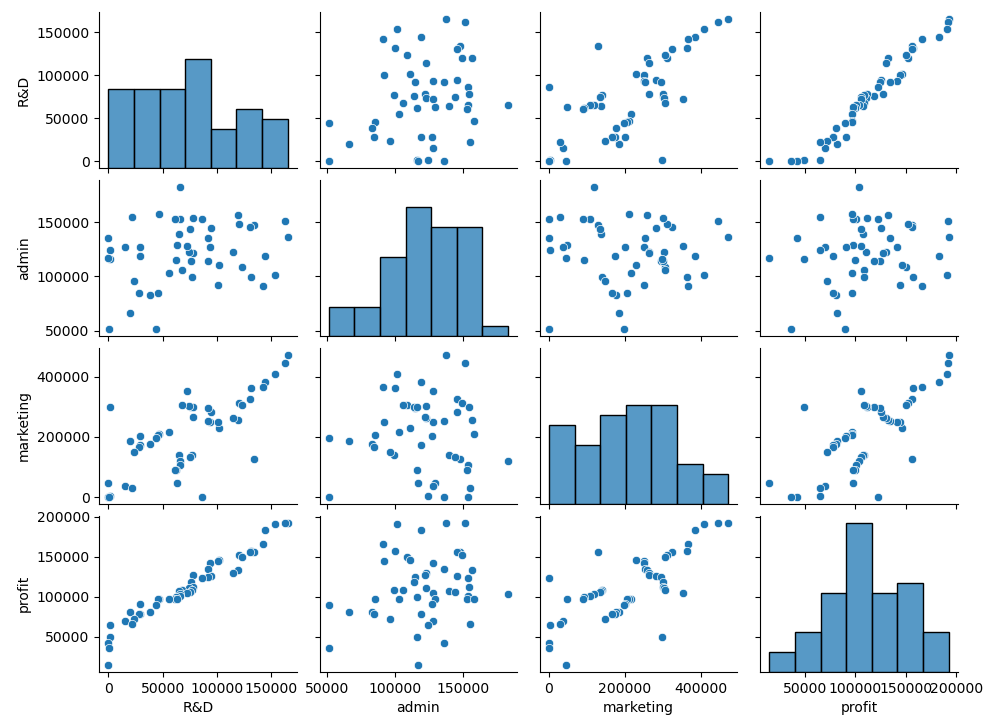
min 0.000000 51283.140000 0.000000 4681.400000

25% 39936.370000 103730.875000 129300.132500 90138.902500

50% 73051.080000 122699.795000 12716.240000 107978.190000

75% 101602.800000 144842.180000 299469.085000 139765.977500

max 165349.200000 182645.560000 471784.100000 192261.830000



From the graph we can see that R&D is the most important variable to predict profit. This can be shown that in corelation matrix below 0.9729

RnD admin marketing profit

RnD 1.000000 0.241955 0.724248 0.972900

admin 0.241955 1.000000 -0.032154 0.200717

marketing 0.724248 -0.032154 1.000000 0.747766

profit 0.972900 0.200717 0.747766 1.000000

to drop variable. We must calculate VIF, If VIF value more than 10 ; remove it from analysis

No Variables VIF

0 RnD 2.468903

1 admin 1.175091

2 marketing 2.326773

From the table above the are no need to remove any variable. Below is the R2 if we remove at least one variables

|  |  |  |  |
| --- | --- | --- | --- |
| No | Model | R square | Adjusted R square |
| 1 | Profit = RnD + Marketing + admin | 0.9510 | 0.9480 |
| 2 | Profit = RnD + admin | 0.9480 | 0.9460 |
| 3 | Profit = Marketing + admin | 0.6100 | 0.5930 |
| 4 | Profit = Marketing + RnD | 0.9500 | 0.9480 |

Model 1 is have higher R2, so this model will be choose as model to train

================================================================================

OLS Regression Results

==============================================================================

Dep. Variable: profit R-squared: 0.951

Model: OLS Adj. R-squared: 0.948

Method: Least Squares F-statistic: 296.0

Date: Fri, 22 Jan 2021 Prob (F-statistic): 4.53e-30

Time: 11:14:32 Log-Likelihood: -525.39

No. Observations: 50 AIC: 1059.

Df Residuals: 46 BIC: 1066.

Df Model: 3

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

----------------------------------------------------------------------------------------------------------------------------------

Intercept 5.012e+04 6572.353 7.626 0.000 3.69e+04 6.34e+04

RnD 0.8057 0.045 17.846 0.000 0.715 0.897

admin -0.0268 0.051 -0.526 0.602 -0.130 0.076

marketing 0.0272 0.016 1.655 0.105 -0.006 0.060

==============================================================================

Omnibus: 14.838 Durbin-Watson: 1.282

Prob(Omnibus): 0.001 Jarque-Bera (JB): 21.442

Skew: -0.949 Prob(JB): 2.21e-05

Kurtosis: 5.586 Cond. No. 1.40e+06

The model is little bit over fitting , as a shown below

test\_rmse: 9979.960150070354, train\_rmse : 8593.129189075678

Predict sales of the computer

Scatter plot show that no collinearity but we must check coefficient matrix

Matrix dia besar sangat tak boleh copy semua

OLS Regression Results

==============================================================================

Dep. Variable: price R-squared: 0.776

Model: OLS Adj. R-squared: 0.775

Method: Least Squares F-statistic: 2399.

Date: Sat, 23 Jan 2021 Prob (F-statistic): 0.00

Time: 21:05:37 Log-Likelihood: -44039.

No. Observations: 6259 AIC: 8.810e+04

Df Residuals: 6249 BIC: 8.817e+04

Df Model: 9

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------------------------------------------------------

Intercept 307.9880 60.353 5.103 0.000 189.675 426.301

trend -51.8496 0.629 -82.470 0.000 -53.082 -50.617

speed 9.3203 0.185 50.364 0.000 8.958 9.683

hd 0.7818 0.028 28.311 0.000 0.728 0.836

ram 48.2560 1.066 45.265 0.000 46.166 50.346

screen 123.0890 3.999 30.776 0.000 115.249 130.929

ads 0.6573 0.051 12.809 0.000 0.557 0.758

cdyes 60.9167 9.516 6.402 0.000 42.263 79.571

multiyes 104.3238 11.413 9.141 0.000 81.951 126.697

premiumyes -509.2247 12.342 -41.259 0.000 -533.420 -485.030

==============================================================================

Omnibus: 1014.821 Durbin-Watson: 1.939

Prob(Omnibus): 0.000 Jarque-Bera (JB): 3190.887

Skew: 0.832 Prob(JB): 0.00

Kurtosis: 6.077 Cond. No. 9.16e+03

Calculate VIF all variable . any variable have VIF > 10 , remove it from equation

No Variables VIF

0 speed 1.265364

1 hd 4.207395

2 ram 2.974628

3 screen 1.081644

4 ads 1.217218

5 trend 2.022790

6 cdyes 1.859370

7 multiyes 1.290568

8 premiumyes 1.109388

There are no variable can be thrown out. Using model 1 for training and testing

test\_rmse: 261.55800023374746

train\_rmse : 280.99978666341826

this model is quite good