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.

R&D Spend -- Research and develop spend in the past few years

Administration -- spend on administration in the past few years

Marketing Spend -- spend on Marketing in the past few years

State -- states from which data is collected

Profit -- profit of each state in the past few years

> dataset<-read.csv("E:/Data Science Asignments/Multilinear Regression/50\_Startups.csv")

> # Encoding categorical data

> dataset$State = factor(dataset$State,

+ levels = c('New York', 'California', 'Florida'),

+ labels = c(1, 2, 3))

> # Splitting the dataset into the Training set and Test set

> install.packages("caTools")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/

Installing package into ‘C:/Users/A M Computer/Documents/R/win-library/4.0’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/caTools\_1.18.0.zip'

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/caTools\_1.18.0.zip'

Content type 'application/zip' length 317214 bytes (309 KB)

downloaded 309 KB

package ‘caTools’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\A M Computer\AppData\Local\Temp\RtmponLp27\downloaded\_packages

> library(caTools)

> set.seed(123)

> split = sample.split(dataset$Profit, SplitRatio = 0.8)

> training\_set = subset(dataset, split == TRUE)

> test\_set = subset(dataset, split == FALSE)

> # Feature Scaling

> # training\_set = scale(training\_set)

> # test\_set = scale(test\_set)

>

> # Fitting Multiple Linear Regression to the Training set

> regressor = lm(formula = Profit ~ .,

+ data = training\_set)

> summary(regressor)

Call:

lm(formula = Profit ~ ., data = training\_set)

Residuals:

Min 1Q Median 3Q Max

-33128 -4865 5 6098 18065

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.965e+04 7.637e+03 6.501 1.94e-07 \*\*\*

R.D.Spend 7.986e-01 5.604e-02 14.251 6.70e-16 \*\*\*

Administration -2.942e-02 5.828e-02 -0.505 0.617

Marketing.Spend 3.268e-02 2.127e-02 1.537 0.134

State2 1.213e+02 3.751e+03 0.032 0.974

State3 2.376e+02 4.127e+03 0.058 0.954

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9908 on 34 degrees of freedom

Multiple R-squared: 0.9499, Adjusted R-squared: 0.9425

F-statistic: 129 on 5 and 34 DF, p-value: < 2.2e-16

>

>

> # Predicting the Test set results

> y\_pred = predict(regressor, newdata = test\_set)

> y\_pred

4 5 8 11 16 20 21 24

173981.09 172655.64 160250.02 135513.90 146059.36 114151.03 117081.62 110671.31

31 32

98975.29 96867.03

>

> #Backward Elimination

> regressor = lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend + State,

+ data = dataset)

> summary(regressor)

Call:

lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +

State, data = dataset)

Residuals:

Min 1Q Median 3Q Max

-33504 -4736 90 6672 17338

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.008e+04 6.953e+03 7.204 5.76e-09 \*\*\*

R.D.Spend 8.060e-01 4.641e-02 17.369 < 2e-16 \*\*\*

Administration -2.700e-02 5.223e-02 -0.517 0.608

Marketing.Spend 2.698e-02 1.714e-02 1.574 0.123

State2 4.189e+01 3.256e+03 0.013 0.990

State3 2.407e+02 3.339e+03 0.072 0.943

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9439 on 44 degrees of freedom

Multiple R-squared: 0.9508, Adjusted R-squared: 0.9452

F-statistic: 169.9 on 5 and 44 DF, p-value: < 2.2e-16

>

> regressor = lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend,

+ data = dataset)

> summary(regressor)

Call:

lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend,

data = dataset)

Residuals:

Min 1Q Median 3Q Max

-33534 -4795 63 6606 17275

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.012e+04 6.572e+03 7.626 1.06e-09 \*\*\*

R.D.Spend 8.057e-01 4.515e-02 17.846 < 2e-16 \*\*\*

Administration -2.682e-02 5.103e-02 -0.526 0.602

Marketing.Spend 2.723e-02 1.645e-02 1.655 0.105

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9232 on 46 degrees of freedom

Multiple R-squared: 0.9507, Adjusted R-squared: 0.9475

F-statistic: 296 on 3 and 46 DF, p-value: < 2.2e-16

>

> regressor = lm(formula = Profit ~ R.D.Spend + Marketing.Spend,

+ data = dataset)

> summary(regressor)

Call:

lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:

Min 1Q Median 3Q Max

-33645 -4632 -414 6484 17097

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.698e+04 2.690e+03 17.464 <2e-16 \*\*\*

R.D.Spend 7.966e-01 4.135e-02 19.266 <2e-16 \*\*\*

Marketing.Spend 2.991e-02 1.552e-02 1.927 0.06 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9161 on 47 degrees of freedom

Multiple R-squared: 0.9505, Adjusted R-squared: 0.9483

F-statistic: 450.8 on 2 and 47 DF, p-value: < 2.2e-16

>

> regressor = lm(formula = Profit ~ R.D.Spend,

+ data = dataset)

> summary(regressor)

Call:

lm(formula = Profit ~ R.D.Spend, data = dataset)

Residuals:

Min 1Q Median 3Q Max

-34351 -4626 -375 6249 17188

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.903e+04 2.538e+03 19.32 <2e-16 \*\*\*

R.D.Spend 8.543e-01 2.931e-02 29.15 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9416 on 48 degrees of freedom

Multiple R-squared: 0.9465, Adjusted R-squared: 0.9454

F-statistic: 849.8 on 1 and 48 DF, p-value: < 2.2e-16

>