

```

> #import train data
> train_data <- read.csv("blogData_train.csv",1)
> #a. Read the dataset and identify the right features.
> #Columns 51,52,55,56,57,60,61,62,263 to 269, 270 to 276, 277, 278, 279,280,
281
> #have been identified pdf attached describing attributes.
> train_data1<- train_data[c(51,52,55,56,57,60,61,62,263:269, 270:276, 277, 2
78, 279,280, 281)]
> #51,52,55,61,
> #test data collated from 60 data test files.
> test_data1<-read.csv('test.csv',1)
> test_data1<- test_data1[c(51,52,55,56,57,60,61,62,263:269, 270:276, 277, 27
8, 279,280, 282)]
> #Model Building
> model1 = lm(formula = x1.0.2 ~ .,data = train_data1)
> summary(model1)

```

Call:

```
lm(formula = x1.0.2 ~ ., data = train_data1)
```

Residuals:

Min	1Q	Median	3Q	Max
-308.68	-5.16	-1.61	1.59	1416.78

Coefficients: (3 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	8.799e+00	8.237e-01	10.683	< 2e-16	***
x2.0.1	1.431e-02	3.501e-03	4.087	4.37e-05	***
x2.0.2	2.634e-01	8.578e-03	30.700	< 2e-16	***
x2.0.4	1.595e-02	5.231e-03	3.049	0.002298	**
x0.0.15	7.767e-01	2.251e-01	3.450	0.000560	***
x0.0.16	-6.122e-01	5.394e-01	-1.135	0.256404	
x0.0.19	4.714e-01	3.149e-01	1.497	0.134348	
x10.0.1	-1.694e-01	7.560e-03	-22.410	< 2e-16	***
x0.0.20	2.421e-04	3.806e-05	6.360	2.03e-10	***
x0.0.221	-4.238e-01	6.254e-01	-0.678	0.498021	
x0.0.222	-1.586e+00	7.203e-01	-2.202	0.027672	*
x0.0.223	-2.716e+00	7.391e-01	-3.676	0.000238	***
x0.0.224	-2.280e+00	6.965e-01	-3.274	0.001061	**
x1.0	-2.417e+00	6.311e-01	-3.829	0.000129	***
x0.0.225	-1.496e+00	5.695e-01	-2.627	0.008614	**
x0.0.226	NA	NA	NA	NA	
x0.0.227	1.690e-01	6.495e-01	0.260	0.794647	
x0.0.228	8.820e-01	6.966e-01	1.266	0.205460	
x0.0.229	-3.511e-02	7.373e-01	-0.048	0.962018	
x1.0.1	-8.999e-01	7.524e-01	-1.196	0.231694	
x0.0.230	-3.314e-01	7.178e-01	-0.462	0.644300	
x0.0.231	-3.206e-01	7.140e-01	-0.449	0.653387	
x0.0.232	NA	NA	NA	NA	
x0.0.233	-6.177e-02	1.015e-01	-0.609	0.542669	
x0.0.234	NA	NA	NA	NA	
x0.0.235	-3.870e-03	1.157e-02	-0.335	0.737943	
x0.0.236	3.575e-03	1.558e-02	0.230	0.818471	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.06 on 52372 degrees of freedom

Multiple R-squared: 0.2315, Adjusted R-squared: 0.2312

F-statistic: 685.9 on 23 and 52372 DF, p-value: < 2.2e-16

```

> #based on significance codes we remove all non star attributes.
> model2 = lm(formula = x1.0.2~x2.0.1+x2.0.2+x2.0.4+x2.0.15+x10.0.1+x0.0.20+x
0.0.222+x0.0.223+x0.0.224+x1.0+x0.0.225,data = train_data1)

```

```
Error in eval(predvars, data, env) : object 'x2.0.15' not found
> #based on significance codes we remove all non star attributes.
> model2 = lm(formula = x1.0.2~x2.0.1+x2.0.2+x2.0.4+x0.0.15+x10.0.1+x0.0.20+x
0.0.222+x0.0.223+x0.0.224+x1.0+x0.0.225,data = train_data1)
> summary(model2)
```

```
Call:
lm(formula = x1.0.2 ~ x2.0.1 + x2.0.2 + x2.0.4 + x0.0.15 + x10.0.1 +
    x0.0.20 + x0.0.222 + x0.0.223 + x0.0.224 + x1.0 + x0.0.225,
    data = train_data1)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-308.33   -5.18   -1.63    1.57  1416.93
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.110e+00  4.193e-01  19.339 < 2e-16 ***
x2.0.1       1.643e-02  3.091e-03   5.316 1.07e-07 ***
x2.0.2       2.577e-01  7.366e-03  34.990 < 2e-16 ***
x2.0.4       2.017e-02  4.479e-03   4.502 6.74e-06 ***
x0.0.15      5.512e-01  1.098e-01   5.020 5.18e-07 ***
x10.0.1     -1.672e-01  7.434e-03 -22.487 < 2e-16 ***
x0.0.20      2.394e-04  3.802e-05   6.297 3.07e-10 ***
x0.0.222     -9.608e-01  5.168e-01  -1.859 0.063042 .
x0.0.223     -1.689e+00  4.878e-01  -3.463 0.000535 ***
x0.0.224     -1.344e+00  4.585e-01  -2.932 0.003371 **
x1.0         -1.824e+00  4.562e-01  -3.998 6.38e-05 ***
x0.0.225     -1.297e+00  4.610e-01  -2.814 0.004899 **
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 33.06 on 52384 degrees of freedom
Multiple R-squared:  0.2313, Adjusted R-squared:  0.2312
F-statistic: 1433 on 11 and 52384 DF, p-value: < 2.2e-16
```

```
> #removing x0.0.222
> model3 = lm(formula = x1.0.2~x2.0.1+x2.0.2+x2.0.4+x0.0.15+x10.0.1+x0.0.20+x
0.0.223+x0.0.224+x1.0+x0.0.225,data = train_data1)
> summary(model3)
```

```
Call:
lm(formula = x1.0.2 ~ x2.0.1 + x2.0.2 + x2.0.4 + x0.0.15 + x10.0.1 +
    x0.0.20 + x0.0.223 + x0.0.224 + x1.0 + x0.0.225, data = train_data1)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-308.26   -5.19   -1.63    1.57  1416.98
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.753e+00  3.729e-01  20.791 < 2e-16 ***
x2.0.1       1.644e-02  3.091e-03   5.317 1.06e-07 ***
x2.0.2       2.577e-01  7.366e-03  34.983 < 2e-16 ***
x2.0.4       2.012e-02  4.479e-03   4.491 7.09e-06 ***
x0.0.15      5.513e-01  1.098e-01   5.021 5.16e-07 ***
x10.0.1     -1.656e-01  7.389e-03 -22.418 < 2e-16 ***
x0.0.20      2.390e-04  3.802e-05   6.286 3.28e-10 ***
x0.0.223     -1.377e+00  4.579e-01  -3.007 0.002643 **
x0.0.224     -1.039e+00  4.281e-01  -2.427 0.015242 *
x1.0         -1.520e+00  4.258e-01  -3.569 0.000359 ***
x0.0.225     -9.940e-01  4.312e-01  -2.305 0.021152 *
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.06 on 52385 degrees of freedom
Multiple R-squared: 0.2313, Adjusted R-squared: 0.2311
F-statistic: 1576 on 10 and 52385 DF, p-value: < 2.2e-16

```
> #removing x0.0.224
> model4 = lm(formula = x1.0.2~x2.0.1+x2.0.2+x2.0.4+x0.0.15+x10.0.1+x0.0.20+x
0.0.223+x1.0+x0.0.225,data = train_data1)
> summary(model4)
```

Call:
lm(formula = x1.0.2 ~ x2.0.1 + x2.0.2 + x2.0.4 + x0.0.15 + x10.0.1 +
x0.0.20 + x0.0.223 + x1.0 + x0.0.225, data = train_data1)

Residuals:

Min	1Q	Median	3Q	Max
-308.14	-5.22	-1.64	1.59	1417.00

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.414e+00	3.457e-01	21.445	< 2e-16 ***
x2.0.1	1.641e-02	3.091e-03	5.309	1.11e-07 ***
x2.0.2	2.576e-01	7.366e-03	34.971	< 2e-16 ***
x2.0.4	2.018e-02	4.479e-03	4.505	6.66e-06 ***
x0.0.15	5.513e-01	1.098e-01	5.021	5.17e-07 ***
x10.0.1	-1.652e-01	7.387e-03	-22.361	< 2e-16 ***
x0.0.20	2.398e-04	3.802e-05	6.307	2.87e-10 ***
x0.0.223	-1.052e+00	4.379e-01	-2.402	0.01633 *
x1.0	-1.197e+00	4.045e-01	-2.959	0.00309 **
x0.0.225	-6.717e-01	4.102e-01	-1.637	0.10158

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.06 on 52386 degrees of freedom
Multiple R-squared: 0.2312, Adjusted R-squared: 0.2311
F-statistic: 1750 on 9 and 52386 DF, p-value: < 2.2e-16

```
> #removing x0.0.225
> model5 = lm(formula = x1.0.2~x2.0.1+x2.0.2+x2.0.4+x0.0.15+x10.0.1+x0.0.20+x
0.0.223+x1.0,data = train_data1)
> summary(model5)
```

Call:
lm(formula = x1.0.2 ~ x2.0.1 + x2.0.2 + x2.0.4 + x0.0.15 + x10.0.1 +
x0.0.20 + x0.0.223 + x1.0, data = train_data1)

Residuals:

Min	1Q	Median	3Q	Max
-308.67	-5.24	-1.63	1.60	1416.48

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.259e+00	3.325e-01	21.828	< 2e-16 ***
x2.0.1	1.641e-02	3.091e-03	5.309	1.10e-07 ***
x2.0.2	2.576e-01	7.366e-03	34.970	< 2e-16 ***
x2.0.4	2.020e-02	4.479e-03	4.509	6.52e-06 ***
x0.0.15	5.503e-01	1.098e-01	5.011	5.42e-07 ***
x10.0.1	-1.652e-01	7.387e-03	-22.371	< 2e-16 ***
x0.0.20	2.402e-04	3.802e-05	6.319	2.66e-10 ***
x0.0.223	-8.957e-01	4.274e-01	-2.096	0.03612 *
x1.0	-1.040e+00	3.931e-01	-2.647	0.00813 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.07 on 52387 degrees of freedom
Multiple R-squared: 0.2312, Adjusted R-squared: 0.231
F-statistic: 1969 on 8 and 52387 DF, p-value: < 2.2e-16

```
> #removing x0.0.223
> model6 = lm(formula = x1.0.2~x2.0.1+x2.0.2+x2.0.4+x0.0.15+x10.0.1+x0.0.20+x
1.0,data = train_data1)
> summary(model6)
```

Call:
lm(formula = x1.0.2 ~ x2.0.1 + x2.0.2 + x2.0.4 + x0.0.15 + x10.0.1 +
x0.0.20 + x1.0, data = train_data1)

Residuals:

	Min	1Q	Median	3Q	Max
	-308.39	-5.25	-1.64	1.60	1416.67

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.068e+00	3.198e-01	22.102	< 2e-16 ***
x2.0.1	1.651e-02	3.091e-03	5.341	9.28e-08 ***
x2.0.2	2.573e-01	7.365e-03	34.932	< 2e-16 ***
x2.0.4	2.039e-02	4.479e-03	4.553	5.29e-06 ***
x0.0.15	5.510e-01	1.098e-01	5.017	5.25e-07 ***
x10.0.1	-1.640e-01	7.362e-03	-22.273	< 2e-16 ***
x0.0.20	2.402e-04	3.802e-05	6.317	2.69e-10 ***
x1.0	-8.923e-01	3.867e-01	-2.308	0.021 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.07 on 52388 degrees of freedom
Multiple R-squared: 0.2311, Adjusted R-squared: 0.231
F-statistic: 2249 on 7 and 52388 DF, p-value: < 2.2e-16

```
> #removing x1.0
> model7= lm(formula = x1.0.2~x2.0.1+x2.0.2+x2.0.4+x0.0.15+x10.0.1+x0.0.20,da
ta = train_data1)
> summary(model7)
```

Call:
lm(formula = x1.0.2 ~ x2.0.1 + x2.0.2 + x2.0.4 + x0.0.15 + x10.0.1 +
x0.0.20, data = train_data1)

Residuals:

	Min	1Q	Median	3Q	Max
	-308.21	-5.25	-1.64	1.60	1416.82

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.920e+00	3.133e-01	22.086	< 2e-16 ***
x2.0.1	1.646e-02	3.091e-03	5.326	1.01e-07 ***
x2.0.2	2.573e-01	7.365e-03	34.938	< 2e-16 ***
x2.0.4	2.039e-02	4.479e-03	4.552	5.32e-06 ***
x0.0.15	5.513e-01	1.098e-01	5.020	5.18e-07 ***
x10.0.1	-1.641e-01	7.363e-03	-22.285	< 2e-16 ***
x0.0.20	2.409e-04	3.802e-05	6.335	2.39e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.07 on 52389 degrees of freedom
Multiple R-squared: 0.231, Adjusted R-squared: 0.2309

F-statistic: 2623 on 6 and 52389 DF, p-value: < 2.2e-16

```
> #we have therefore fine tuned the model as all attributes have high
> #significance level of ***
> #Visualising
> pairs(~x1.0.2+x2.0.1+x2.0.2+x2.0.4+x0.0.15+x10.0.1+x0.0.20,data = train_data1)
> #we have therefore fine tuned the model as all attributes have high
> #significance level of ***
> #Visualising
> pairs(~x1.0.2+x2.0.1+x2.0.2+x2.0.4+x0.0.15+x10.0.1+x0.0.20,data = train_data1)

> #predicting with model7 and test.
> predict_1<-predict(model7,test_data1[, -27])
Error in eval(predvars, data, env) : object 'x2.0.1' not found
> #predicting with model7 and test.
> predict_1<-predict(model7,test_data1[, -27])
Error in eval(predvars, data, env) : object 'x2.0.1' not found
> #predicting with model7 and test.
> test_data2<- test_data1[, -27]
> predict_1<-predict(model7,test_data2)
Error in eval(predvars, data, env) : object 'x2.0.1' not found
> colnames(test_data1) <- colnames(train_data1)
> predict_1<-predict(model7,test_data1[, -27])
> test_data1$Prediction<-predict_1
> #accuracy of prediction
> r<-cor(test_data1$x1.0.2,test_data1$Prediction)
> r
[1] 0.4277208
> rsquared <-cor(test_data1$x1.0.2,test_data1$Prediction)^2
> rsquared
[1] 0.1829451
> #accuracy of prediction for training
> predict_2 <- predict(model7,train_data1[, -27])
> train_data1$Prediction<-predict_2
> r <- cor(train_data1$x1.0.2,train_data1$Prediction)
> r1 <- cor(train_data1$x1.0.2,train_data1$Prediction)
> r1
[1] 0.480635
> rsquared<- cor(train_data1$x1.0.2,train_data1$Prediction)^2
> rsquared1
Error: object 'rsquared1' not found
> rsquared1<- cor(train_data1$x1.0.2,train_data1$Prediction)^2
> rsquared1
[1] 0.23101
```