EAS 508 Homework – 1

Name : Venkata Satya Surya Sai Vineet Atyam

UB Person number : 50419767

UBIT Name : vatyam

My suggested method for predicting Property1 is by using Principal Component Analysis and choosing the first 4 Principal Components that give us around 95 percent variance followed by Support Vector Regression on the chosen parameters

Parameters for the modelled SVR with 11 Support Vectors are (tuned best model) :

1.) W

```
     [,1] [,2] [,3]    [,4]        [,5] [,6] [,7] [,8]     [,9]          [,10]         [,11]
[1,]  1   -1    1    0.1007381   -1    1    1     -1   -0.3332735  -0.4909832   -0.2764814
```

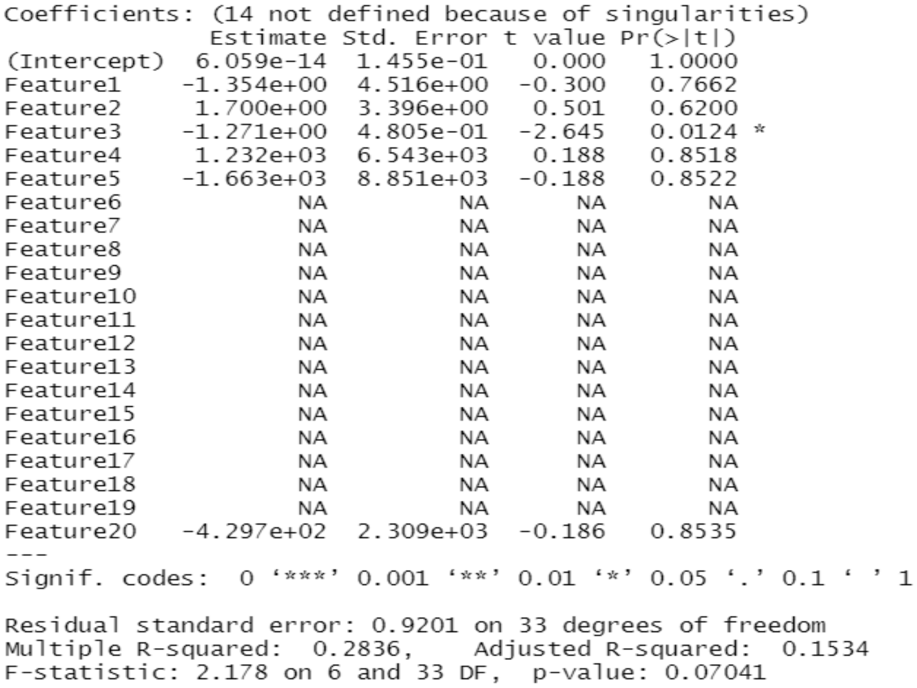
2.) b: -0.2559929

3.) Epsilon = 1

4.) Cost = 1

A.) This model was chosen because, it could be seen that the features had a lot of correlation between them and hence during the multiple linear regression, 14 of the features became singularities due to high relation between them.

Hence, there was a need to reduce the number of features that were to be considered.

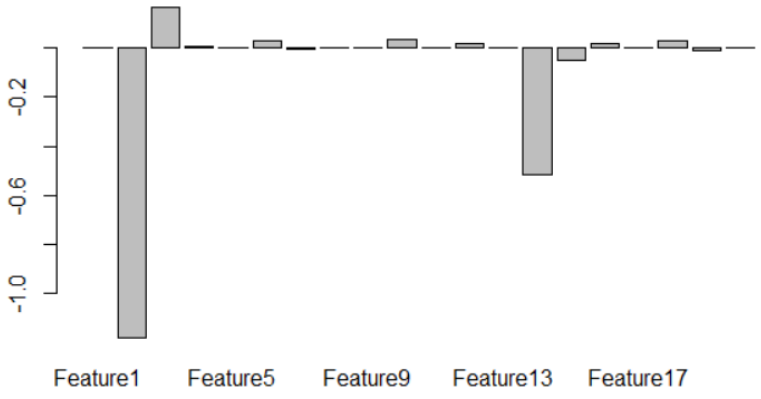


```
Coefficients: (14 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.059e-14  1.455e-01   0.000   1.0000
Feature1    -1.354e+00  4.516e+00  -0.300   0.7662
Feature2     1.700e+00  3.396e+00   0.501   0.6200
Feature3    -1.271e+00  4.805e-01  -2.645   0.0124 *
Feature4     1.232e+03  6.543e+03   0.188   0.8518
Feature5    -1.663e+03  8.851e+03  -0.188   0.8522
Feature6            NA         NA      NA       NA
Feature7            NA         NA      NA       NA
Feature8            NA         NA      NA       NA
Feature9            NA         NA      NA       NA
Feature10           NA         NA      NA       NA
Feature11           NA         NA      NA       NA
Feature12           NA         NA      NA       NA
Feature13           NA         NA      NA       NA
Feature14           NA         NA      NA       NA
Feature15           NA         NA      NA       NA
Feature16           NA         NA      NA       NA
Feature17           NA         NA      NA       NA
Feature18           NA         NA      NA       NA
Feature19           NA         NA      NA       NA
Feature20   -4.297e+02  2.309e+03  -0.186   0.8535
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9201 on 33 degrees of freedom
Multiple R-squared:  0.2836,	Adjusted R-squared:  0.1534 
F-statistic: 2.178 on 6 and 33 DF,  p-value: 0.07041
```

B.) By making the barplot of the weights matrix constructed using the eigen values, features 2, 3 ,14 ,15 we selected as they had more importance compared to the other features. These features can then be used for linear regression directly to produce acceptable results by reducing the dimesnsions.

The less the number of descriptors used, the model will not be over fit and hence would work similarly when tested on a new dataset.



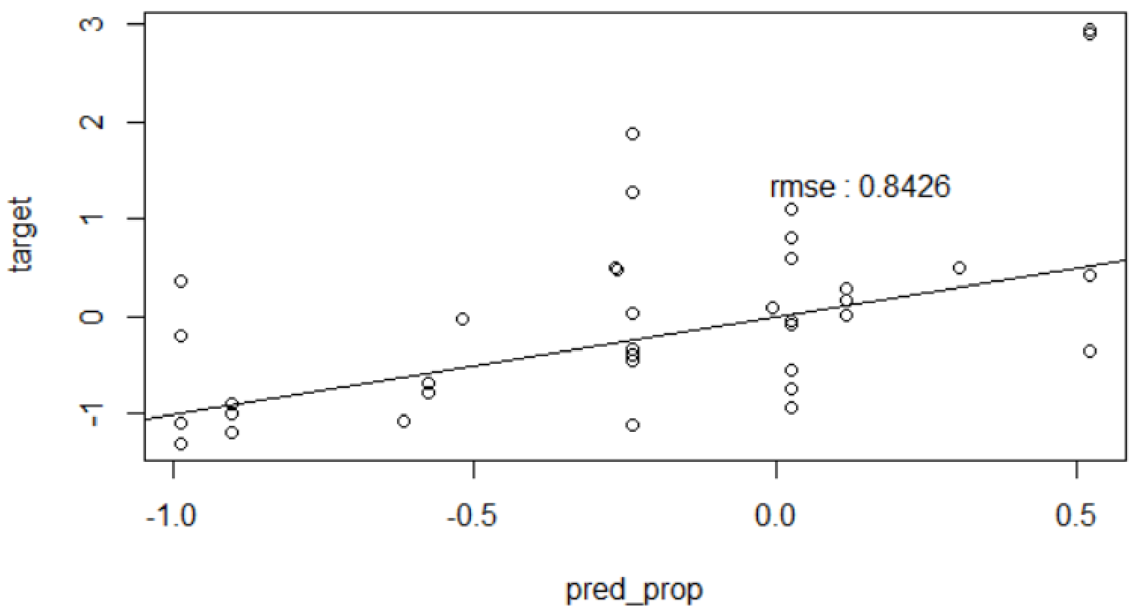
C.) By only using the required amount of Principal Components we are collecting most of the information that can be provided. Though this reduces the accuracy by a little, it is most important to increase the interpretability as with less number of features , we can easily understand the features being responsible for the outcome.

Reducing the features also helps in preventing the model to overfit it and would in turn help increase the robustness.

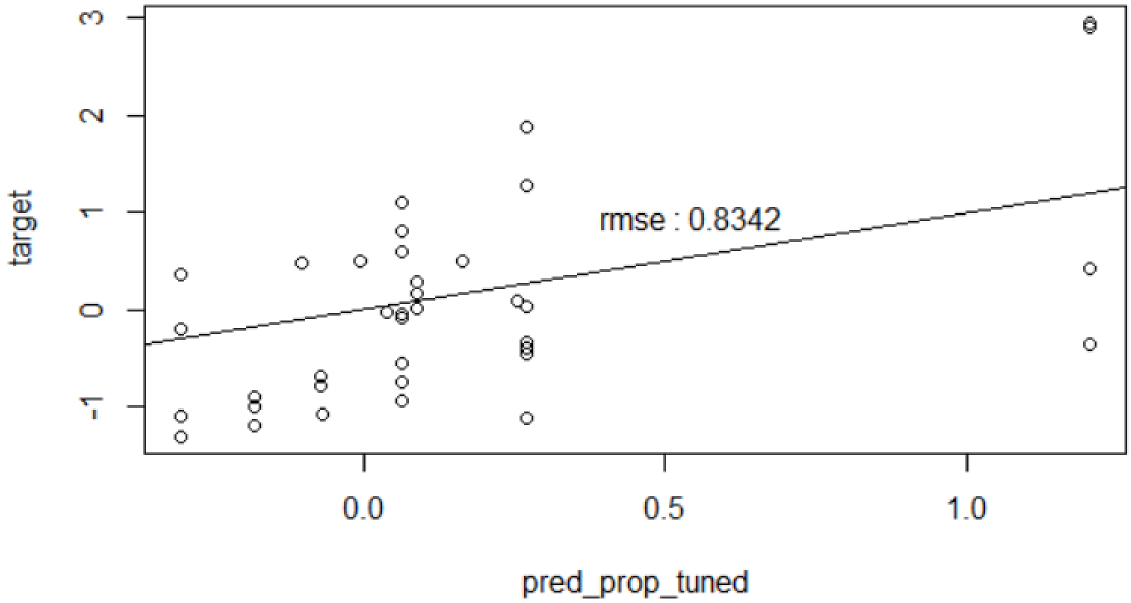
This is observed by cross validation of the model over different splits of data into test and training data, where the training and test rmse is nearly equal.

D.) Figures to understand the SVM plot.

Model Scatter Plot



TUNED SVM Model


target
rmse : 0.8342
pred_prop_tuned