STATISTICAL DATA MINING – I
HOMEWORK V

NAME: SIDDIQ SYED

UB Person # 50291566

Class # 50

1) (10 points ~ Exercise 15.6) Fit a series of random-forest classifiers to the SPAM data, to explore the sensitivity to m (the number of randomly selected inputs for each tree). Plot both the OOB error as well as the test error against a suitably chosen range of values for m.

Exploratory analysis of the data

The dataset spam has 4601 observations with 58 variables, in which type i.e. spam or nonspam is the classifier. Below is the summary for the given dataset for few variables.

```
      make              address              all               num3d              our          
 Min.   :0.0000000   Min.   : 0.0000000   Min.   :0.0000000   Min.   : 0.00000000   Min.   : 0.0000000  
 1st Qu.:0.0000000   1st Qu.: 0.0000000   1st Qu.:0.0000000   1st Qu.: 0.00000000   1st Qu.: 0.0000000  
 Median :0.0000000   Median : 0.0000000   Median :0.0000000   Median : 0.00000000   Median : 0.0000000  
 Mean   :0.1045534   Mean   : 0.2130146   Mean   :0.2806564   Mean   : 0.06542491   Mean   : 0.3122234  
 3rd Qu.:0.0000000   3rd Qu.: 0.0000000   3rd Qu.:0.4200000   3rd Qu.: 0.00000000   3rd Qu.: 0.3800000  
 Max.   :4.5400000   Max.   :14.2800000   Max.   :5.1000000   Max.   :42.81000000   Max.   :10.0000000  
      over              remove            internet          order              mail        
 Min.   :0.00000000   Min.   :0.0000000   Min.   : 0.0000000   Min.   :0.00000000   Min.   : 0.0000000  
 1st Qu.:0.00000000   1st Qu.:0.0000000   1st Qu.: 0.0000000   1st Qu.:0.00000000   1st Qu.: 0.0000000  
 Median :0.00000000   Median :0.0000000   Median : 0.0000000   Median :0.00000000   Median : 0.0000000  
 Mean   :0.09590089   Mean   :0.1142078   Mean   : 0.1052945   Mean   :0.09006738   Mean   : 0.2394132  
 3rd Qu.:0.00000000   3rd Qu.:0.0000000   3rd Qu.: 0.0000000   3rd Qu.:0.00000000   3rd Qu.: 0.1600000  
 Max.   :5.88000000   Max.   :7.2700000   Max.   :11.1100000   Max.   :5.26000000   Max.   :18.1800000  
    receive              will              people             report          addresses       
 Min.   :0.00000000   Min.   :0.0000000   Min.   :0.00000000   Min.   : 0.00000000   Min.   :0.00000000  
 1st Qu.:0.00000000   1st Qu.:0.0000000   1st Qu.:0.00000000   1st Qu.: 0.00000000   1st Qu.:0.00000000  
 Median :0.00000000   Median :0.1000000   Median :0.00000000   Median : 0.00000000   Median :0.00000000  
 Mean   :0.05982395   Mean   :0.5417018   Mean   :0.09392958   Mean   : 0.05862639   Mean   :0.04920452  
 3rd Qu.:0.00000000   3rd Qu.:0.8000000   3rd Qu.:0.00000000   3rd Qu.: 0.00000000   3rd Qu.:0.00000000  
 Max.   :2.61000000   Max.   :9.6700000   Max.   :5.55000000   Max.   :10.00000000   Max.   :4.41000000  
      free            business           email                you               credit              your        
 Min.   : 0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   : 0.0000   Min.   : 0.00000000   Min.   : 0.0000000  
 1st Qu.: 0.0000000   1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.: 0.0000   1st Qu.: 0.00000000   1st Qu.: 0.0000000  
 Median : 0.0000000   Median :0.0000000   Median :0.0000000   Median : 1.3100   Median : 0.00000000   Median : 0.2200000  
 Mean   : 0.2488481   Mean   :0.1425864   Mean   :0.1847446   Mean   : 1.6621   Mean   : 0.08557705   Mean   : 0.8097609  
 3rd Qu.: 0.1000000   3rd Qu.:0.0000000   3rd Qu.:0.0000000   3rd Qu.: 2.6400   3rd Qu.: 0.00000000   3rd Qu.: 1.2700000  
 Max.   :20.0000000   Max.   :7.1400000   Max.   :9.0900000   Max.   :18.7500   Max.   :18.18000000   Max.   :11.1100000  
      font              num000             money                hp                hpl          
 Min.   : 0.0000000   Min.   :0.0000000   Min.   : 0.00000000   Min.   : 0.0000000   Min.   : 0.0000000  
 1st Qu.: 0.0000000   1st Qu.:0.0000000   1st Qu.: 0.00000000   1st Qu.: 0.0000000   1st Qu.: 0.0000000  
 Median : 0.0000000   Median :0.0000000   Median : 0.00000000   Median : 0.0000000   Median : 0.0000000  
 Mean   : 0.1212019   Mean   :0.1016453   Mean   : 0.09426864   Mean   : 0.5495045   Mean   : 0.2653836  
 3rd Qu.: 0.0000000   3rd Qu.:0.0000000   3rd Qu.: 0.00000000   3rd Qu.: 0.0000000   3rd Qu.: 0.0000000  
 Max.   :17.1000000   Max.   :5.4500000   Max.   :12.50000000   Max.   :20.8300000   Max.   :16.6600000  
     george              num650              lab                labs              telnet       
```

From which we can observe that most of the values are zeros for the data. The data is telling us about the words that were being used in emails with their counts and few other parameters.

To fit the series of random forest classification model to the SPAM data we first divided the data into training and testing sets.

With m values 1,5,10,15,20,25 and setting the number of trees to 500 we have developed models to which the test data has been predicted. The error rates for each of the m value has been shown in the below plot.

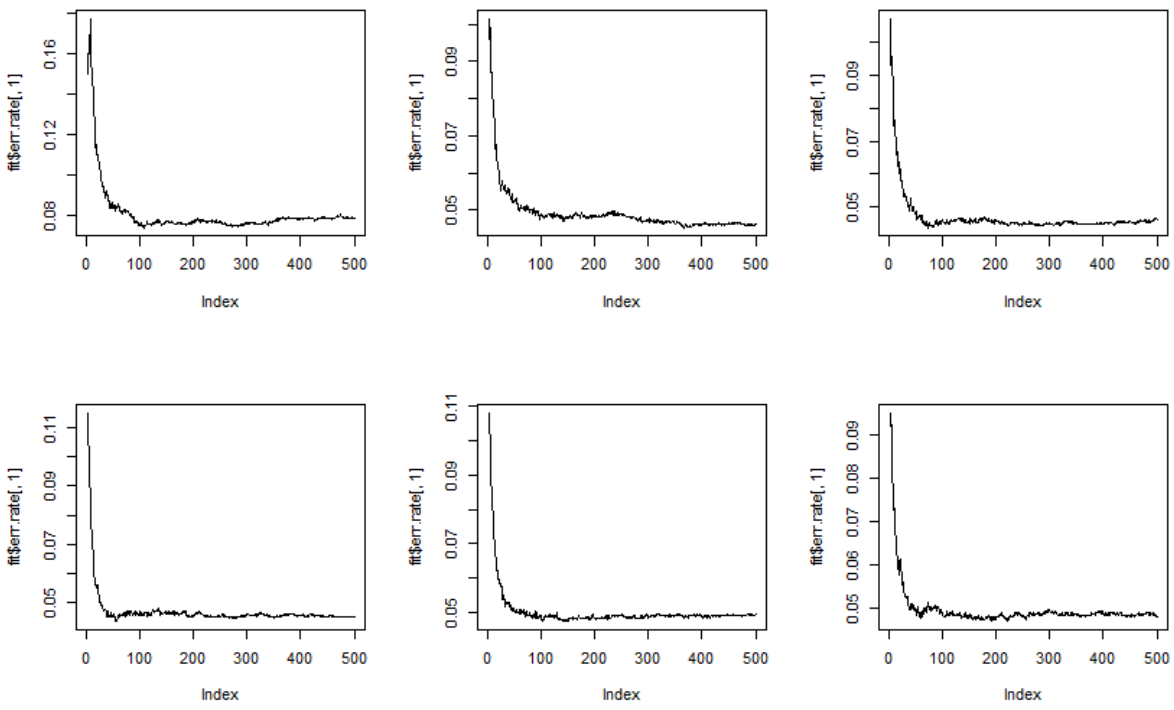


Figure with error rates for m = 1,5,10,15,20,25 staring for left top to right bottom.

From the above graphs we can observe that the error rate pretty much remains constant from 100 trees for all the values of m.

Also to plot the miss classification rate for each value of m

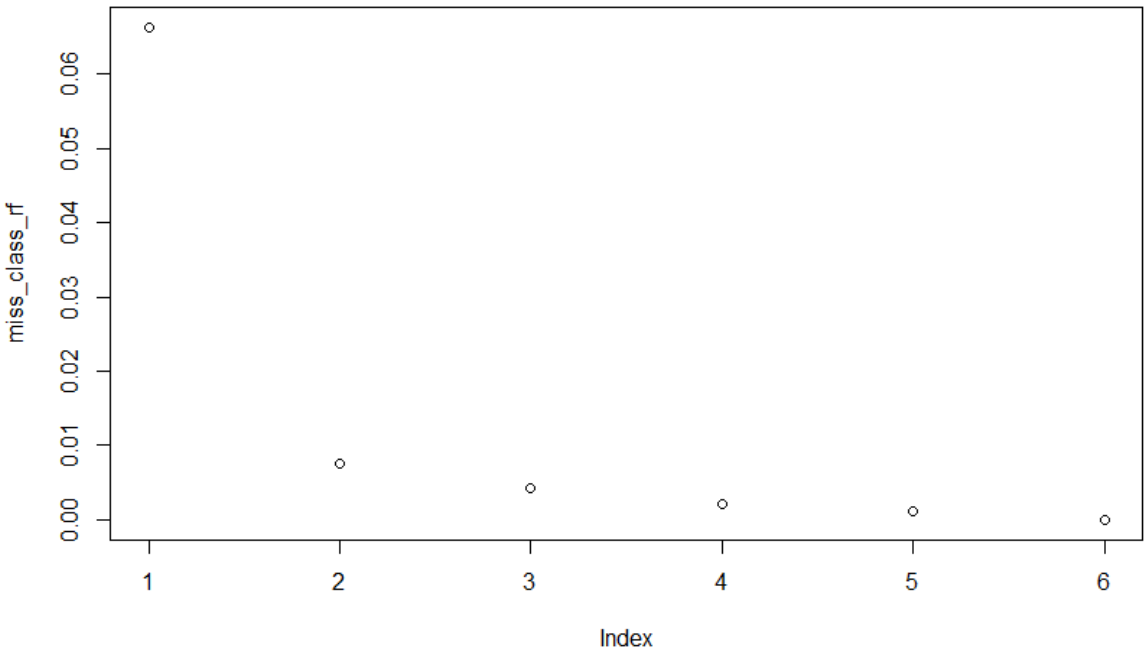
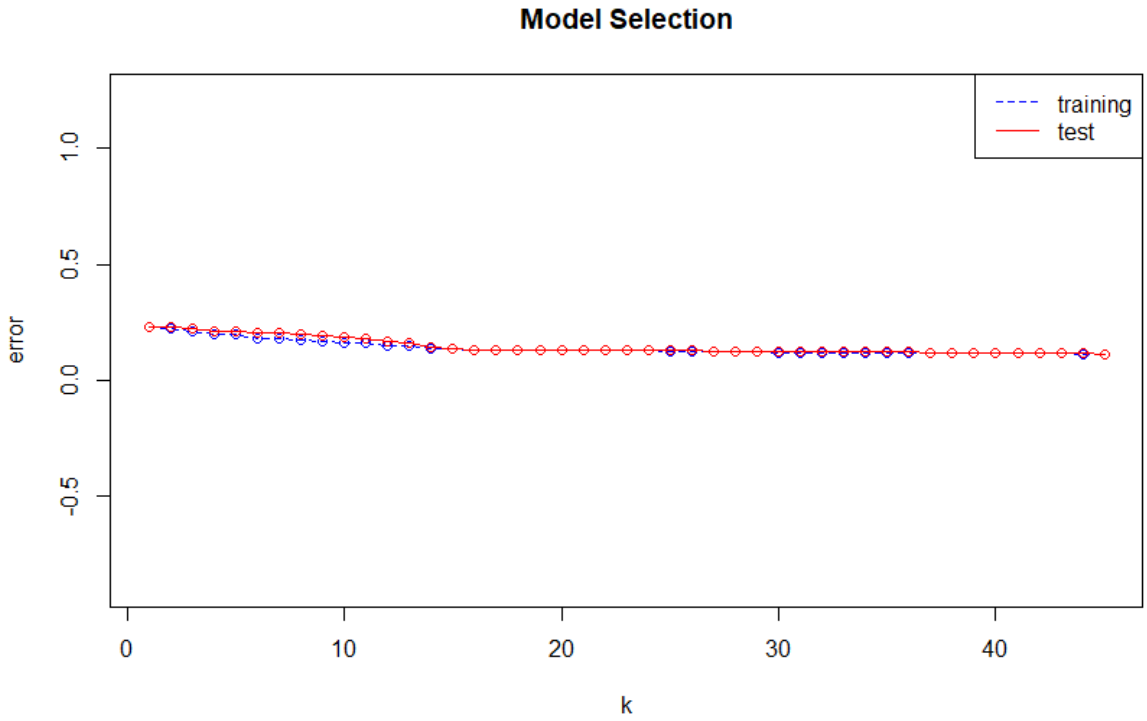


Figure Index – 1 – m = 1, 2-5,3-10,4-15,5-20,6-25

As the m value increases the miss classification rate decreases. We can see that the error rate is about to converge after the values of m greater than 20.

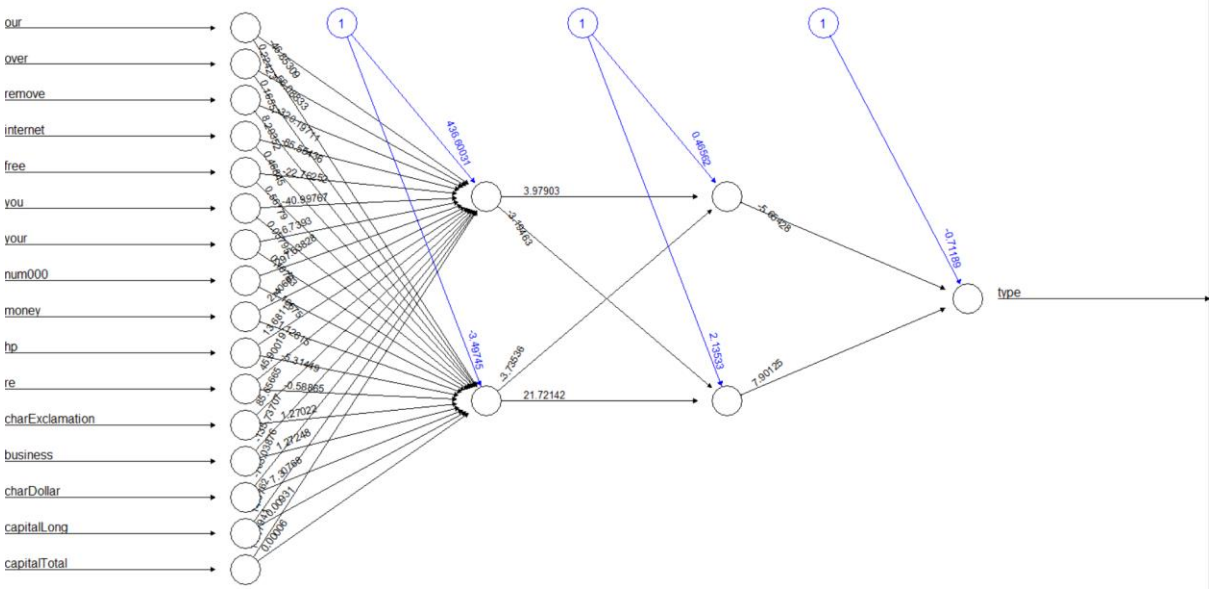
2) (10 points; Exercise 11.7) Fit a neural network to the spam data of Section 9.1.2. The data is available through the package "ElemStatLearn". Use cross-validation or the hold out method to determine the number of neurons to use in the layer. Compare your results to those for the additive model given in the chapter. When making the comparison, consider both the classification performance and interpretability of the final model.

For the dataset spam the output variable type needs to be converted into numeric form for proceedings. The subset of training and testing is divided for the spam dataset. I have chosen holdout method to find the best parameters for data. After using forward and backward subset selections Cp values says 43 variables is best from forward and 43 from backward and from BIC which says 32 variables is best for forward and 30 for backward. Checking the error for different number of best parameters. The plot for the same is as below.



From the graph we can see that the model at 15 i.e. with 15 parameters the error rate gets converges as the number of parameters increases.

So for fitting the data with these 15 parameters to build the model and predict the test data.



Above is the 2 layers 2 node neural network model for the 15 parameter model of neural network.

For which the error rate is found to be 0.0799.

When compared with the tree model that was built as part of the question 1 was 0.042

Also Logistic regression model was built to check the prediction error for the dataset. Its was found to be 0.0684

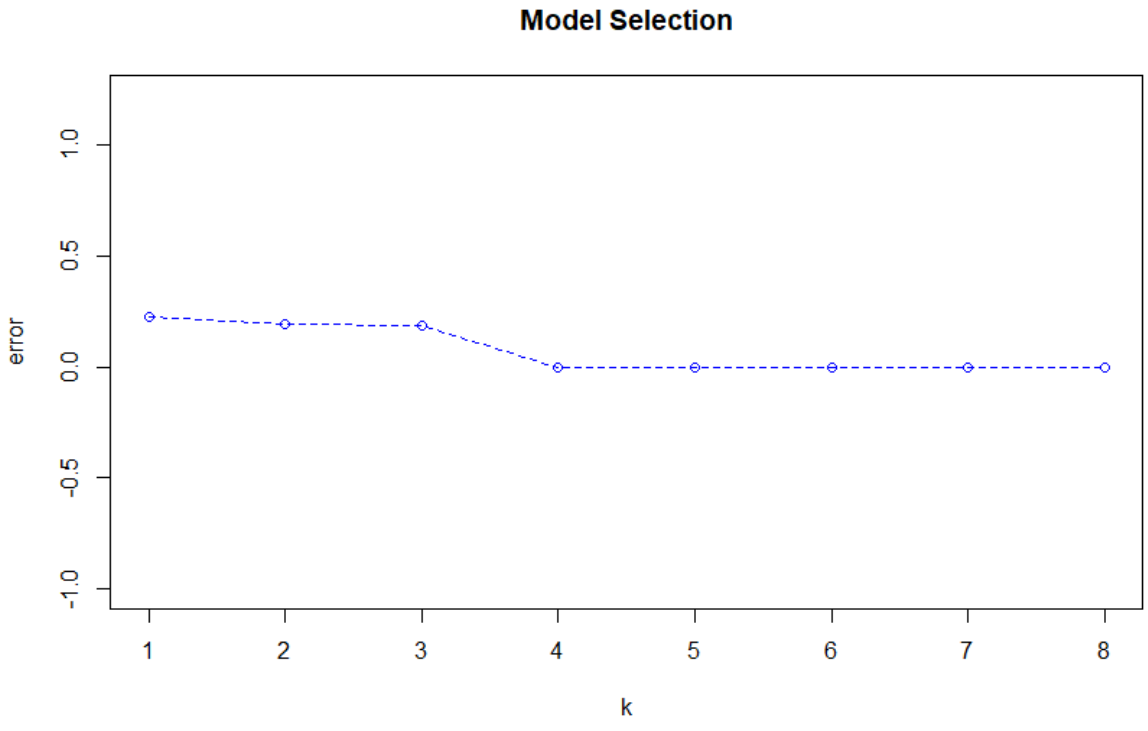
I was able see a lot of variation in neural network outputs with small changes in the model building parameters. It is considered to be tougher than a simplistic models to interpret the output results of a neural network model.

Because it was taking a lot of time to process the holdout method on neural network models. I have used linear model to select the number of parameters.

3) (10 points) Take any classification data set and divide it up into a learning set and a test set. Change the value of one observation on one input variable in the learning set so that the value is now a univariate outlier. Fit separate single hidden-layer neural networks to the original learning-set data and to the learning set data with the outlier. Use cross-validation or the hold out method to determine the number of neurons to use in the layer. Comment on the effect of the outlier on the fit and on its effect on classifying the test set. Shrink the value of that outlier toward its original value and evaluate when the effect of the outlier on the fit vanishes. How far away must the outlier move from its original value that significant changes to the network coefficient estimates occur?

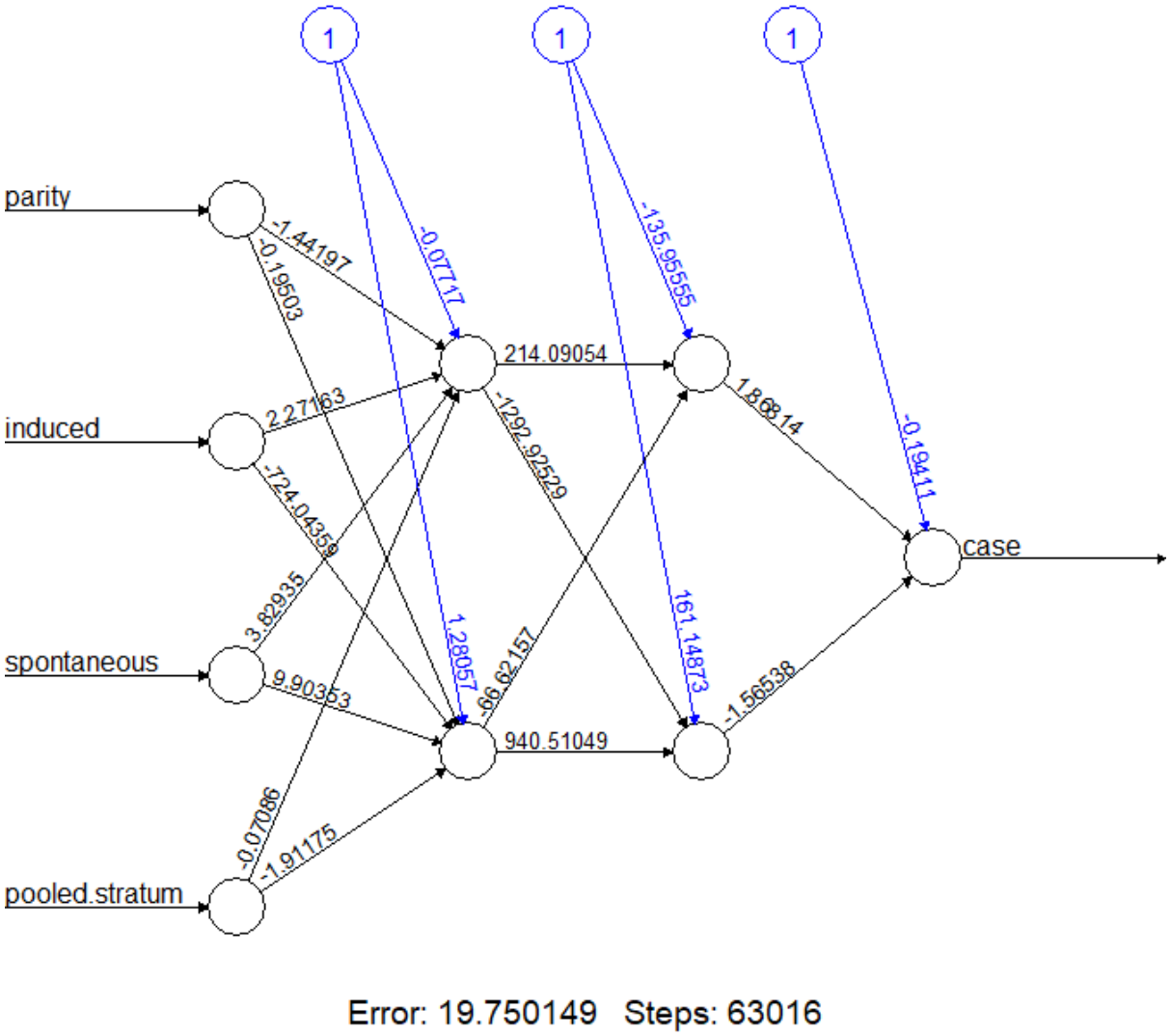
Answer:

I have taken infert data to analyse the above given problem. To start with the dataset summary has been visualized and then the dataset is divided into training and testing datasets. For which the exhaustive subset selection was done to find the best models for different number of parameters. Cross validation for linear model was done to find the best number of parameters with minimum error.



From the below graph it is visualized that from 4 parameters that error rate got converged.

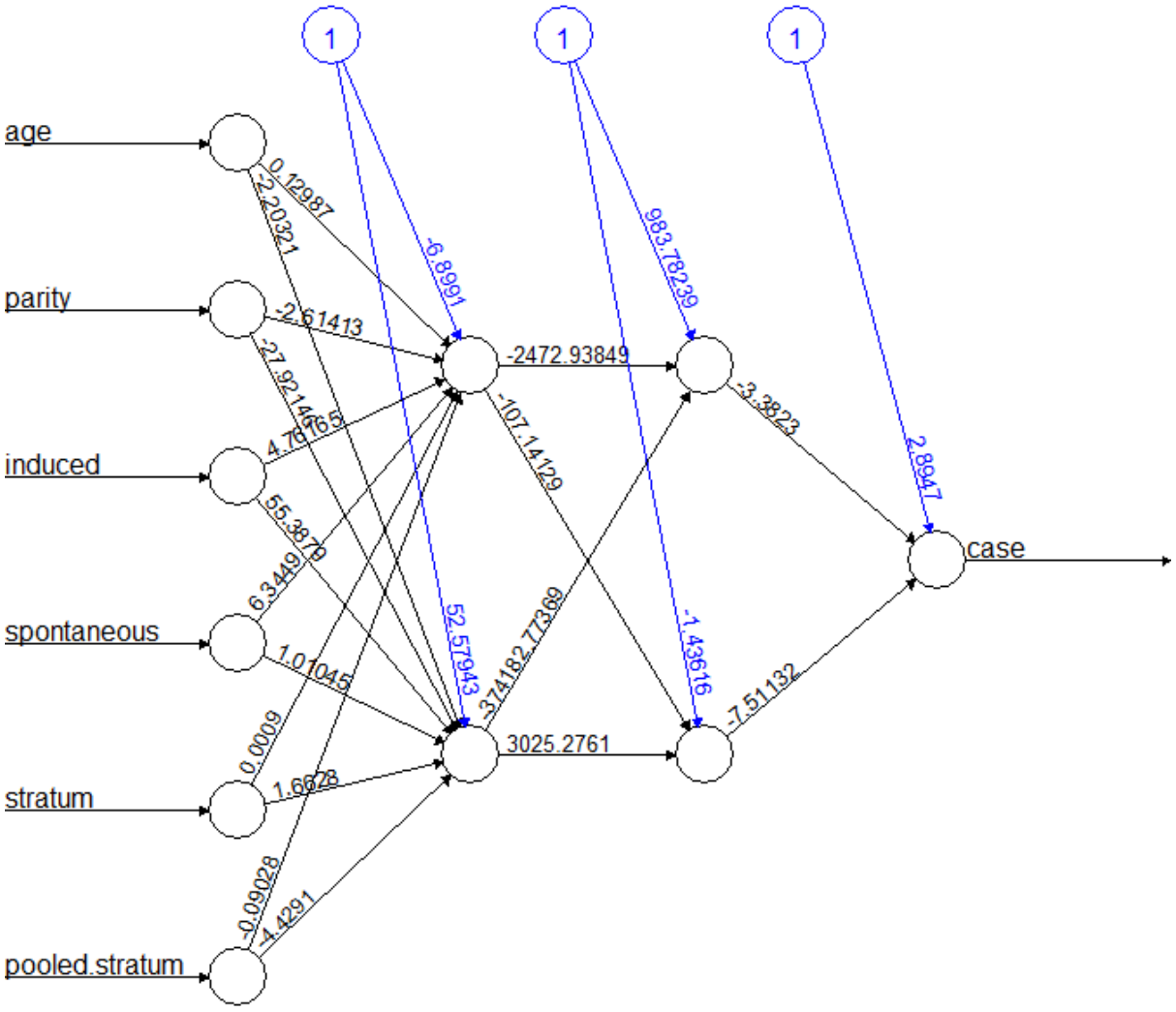
Using the best 4 parameters the neural network model was built. Below is the neural network plot for the same.



Error: 19.750149 Steps: 63016

The prediction error for the same was 0.2298.

New inducing an outlier for the column pooled.stratum which was previously 13 and edited to 10 - 26.95,30 - 20.63,50 - 20.63,70 - 18.02,90 – 17.84 and the error changed as mentioned with respect the value changes. It is seen that the error values decreases as the values becomes outlier. The value can move outisde the maximum value of the particular column. Once the values changes to anything less than maximum value the error spices up.



Above is the plot when values is changes from 13 to 90(outlier).

Which clearly says that the neural networks model is very robust to the outliers.

4) (10 points; ISLR modified Ch9ex8) This problem involves the OJ data set in the ISLR package. We are interested in the prediction of "Purchase". Divide the data into test and training.

(A) Fit a support vector classifier with varying cost parameters over the range [0.01, 10]. Plot the training and test error across this spectrum of cost parameters, and determine the optimal cost.

(B) Repeat the exercise in (A) for a support vector machine with a radial kernel. (Use the default parameter for gamma). Repeat the exercise again for a support vector machine with a polynomial kernel of degree=2. Reflect on the performance of the SVM with different kernels, and the support vector classifier, i.e., SVM with a linear kernel.

Answer:

Below is the summary for the OJ dataset.

```
Purchase WeekofPurchase        StoreID           PriceCH          PriceMM          DiscCH
CH:653   Min.   :227.0000   Min.   :1.000000   Min.   :1.690000   Min.   :1.690000   Min.   :0.00000000
MM:417   1st Qu.:240.0000   1st Qu.:2.000000   1st Qu.:1.790000   1st Qu.:1.990000   1st Qu.:0.00000000
         Median :257.0000   Median :3.000000   Median :1.860000   Median :2.090000   Median :0.00000000
         Mean   :254.3813   Mean   :3.959813   Mean   :1.867421   Mean   :2.085411   Mean   :0.05185981
         3rd Qu.:268.0000   3rd Qu.:7.000000   3rd Qu.:1.990000   3rd Qu.:2.180000   3rd Qu.:0.00000000
         Max.   :278.0000   Max.   :7.000000   Max.   :2.090000   Max.   :2.290000   Max.   :0.50000000
    DiscMM          SpecialCH          SpecialMM          LoyalCH          SalePriceMM      SalePriceCH
Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000110   Min.   :1.190000   Min.   :1.390000
1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.3252572   1st Qu.:1.690000   1st Qu.:1.750000
Median :0.0000000   Median :0.0000000   Median :0.0000000   Median :0.6000000   Median :2.090000   Median :1.860000
Mean   :0.1233645   Mean   :0.1476636   Mean   :0.1616822   Mean   :0.5657823   Mean   :1.962047   Mean   :1.815561
3rd Qu.:0.2300000   3rd Qu.:0.0000000   3rd Qu.:0.0000000   3rd Qu.:0.8508727   3rd Qu.:2.130000   3rd Qu.:1.890000
Max.   :0.8000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :0.9999470   Max.   :2.290000   Max.   :2.090000
  PriceDiff         Store7      PctDiscMM          PctDiscCH          ListPriceDiff        STORE
Min.   :-0.670000   No :714   Min.   :0.00000000   Min.   :0.00000000   Min.   :0.0000000   Min.   :0.000000
1st Qu.: 0.000000   Yes:356   1st Qu.:0.00000000   1st Qu.:0.00000000   1st Qu.:0.1400000   1st Qu.:0.000000
Median : 0.230000             Median :0.00000000   Median :0.00000000   Median :0.2400000   Median :2.000000
Mean   : 0.146486             Mean   :0.05929844   Mean   :0.02731384   Mean   :0.2179907   Mean   :1.630841
3rd Qu.: 0.320000             3rd Qu.:0.11267600   3rd Qu.:0.00000000   3rd Qu.:0.3000000   3rd Qu.:3.000000
Max.   : 0.640000             Max.   :0.40201000   Max.   :0.25268800   Max.   :0.4400000   Max.   :4.000000
```

Diving the dataset into training and testing.

Modelling the data with svm for the cost in range 0.01-10. The model output is shown below.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost
 8.01

- best performance: 0.1805555556
```

Which says that the model performance is best at cost 8.01.

The summary for the same is also shown below.

```
- Detailed performance results:
   cost         error    dispersion
1  0.01 0.1848200313 0.05789590898
2  1.01 0.1847809077 0.05473466000
3  2.01 0.1875782473 0.05786187092
4  3.01 0.1819640063 0.05880637283
5  4.01 0.1847613459 0.05648795730
6  5.01 0.1847613459 0.05648795730
7  6.01 0.1819444444 0.05491948059
8  7.01 0.1819444444 0.05832625324
9  8.01 0.1805555556 0.05816065417
10 9.01 0.1819640063 0.05686938188
```

From the summary we can see that the error at 8.01 cost is 0.1805 and comparing to other errors it is not pretty efficient as all the values are very close.

Also best gamma value and number of support vectors that are required are also shown below.

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  8.01
      gamma:  0.05555555556

Number of Support Vectors:  307
```

Same procedure when repeated with radial kernel.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma
 1.01   0.5

- best performance: 0.1958920188
```

Above is the cost and performance for the same.

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1.01
      gamma:  0.5

Number of Support Vectors:  402
```

Also number support vectors used is shown above.

Same procedure when seen for the polynomial kernel.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma degree
 2.01   0.5      2

- best performance: 0.1722809077
```

Cost is optimal for cost of 2.01. and performance is less when compared with linear and radial kernel.

Also the number of support vectors being utilized is shown below.

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  2.01
     degree:  2
      gamma:  0.5
     coef.0:  0

Number of Support Vectors:  273
```

Performance is better for polynomial kernel with degree 2.

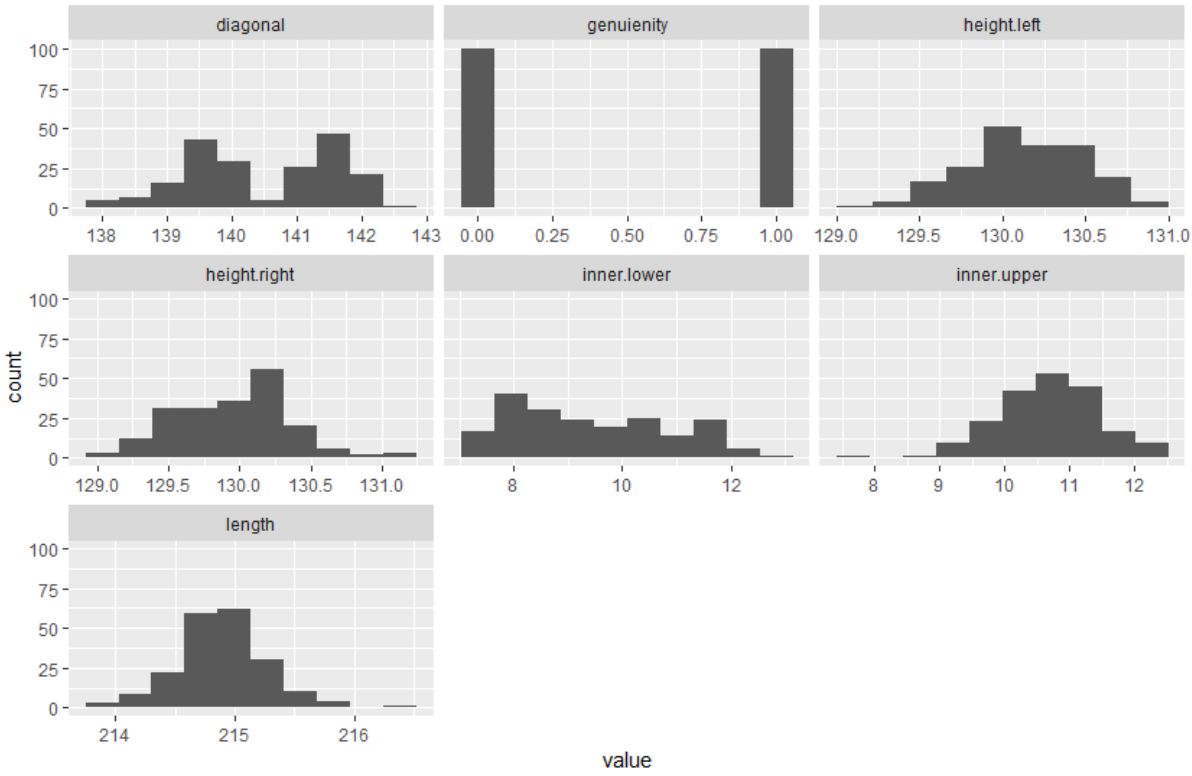
5) (10 points) Access the SwissBankNotes data (posted with assignment). The data consists of six variables measured on 200 old Swiss 1,000-franc bank notes. The first 100 are genuine and the second 100 are counterfeit. The six variables are length of the bank note, height of the bank note, measured on the left, height of the bank note measured on the right, distance of the inner frame to the lower border, distance of inner frame to upper border, and length of the diagonal. Carry out a PCA of the 100 genuine bank notes, of the 100 counterfeit bank notes, and all of the 200 bank notes combined. Do you notice any differences in the results? Show all work in the selection of Principal Components, including diagnostic plots.

Answer:

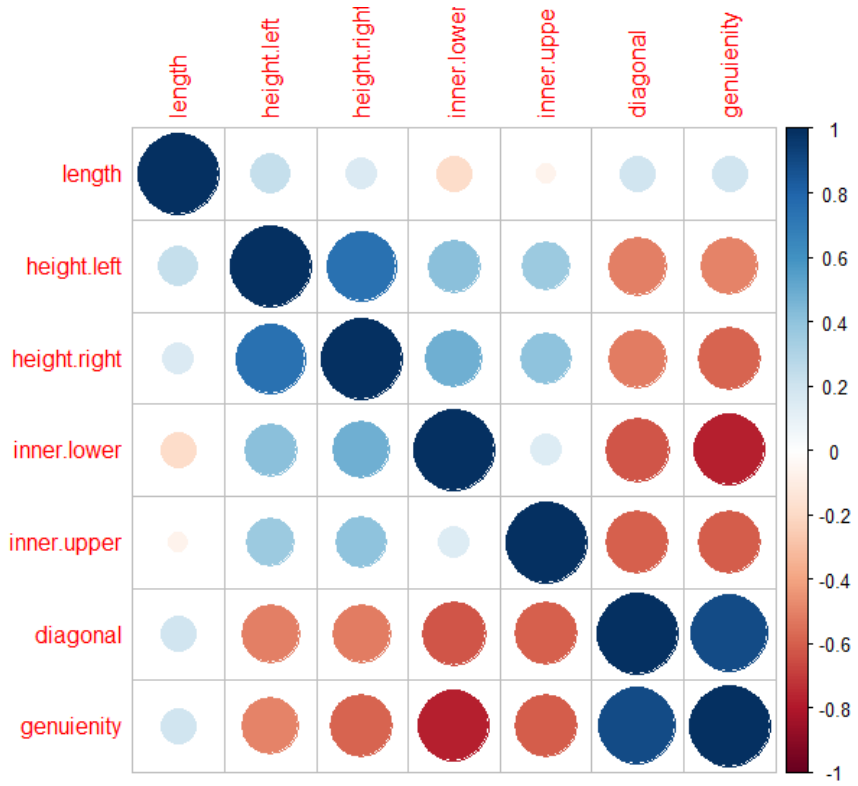
Below is the summary for the dataset after adding the genuinity column.

```
     length        height.left      height.right      inner.lower      inner.upper       diagonal
Min.   :213.800   Min.   :129.0000   Min.   :129.0000   Min.   : 7.2000   Min.   : 7.7000   Min.   :137.8000
1st Qu.:214.600   1st Qu.:129.9000   1st Qu.:129.7000   1st Qu.: 8.2000   1st Qu.:10.1000   1st Qu.:139.5000
Median :214.900   Median :130.2000   Median :130.0000   Median : 9.1000   Median :10.6000   Median :140.4500
Mean   :214.896   Mean   :130.1215   Mean   :129.9565   Mean   : 9.4175   Mean   :10.6505   Mean   :140.4835
3rd Qu.:215.100   3rd Qu.:130.4000   3rd Qu.:130.2250   3rd Qu.:10.6000   3rd Qu.:11.2000   3rd Qu.:141.5000
Max.   :216.300   Max.   :131.0000   Max.   :131.1000   Max.   :12.7000   Max.   :12.3000   Max.   :142.4000
  genuienity
Min.   :0.0
1st Qu.:0.0
Median :0.5
Mean   :0.5
3rd Qu.:1.0
Max.   :1.0
```

Also histogram plot for each column is shown below.

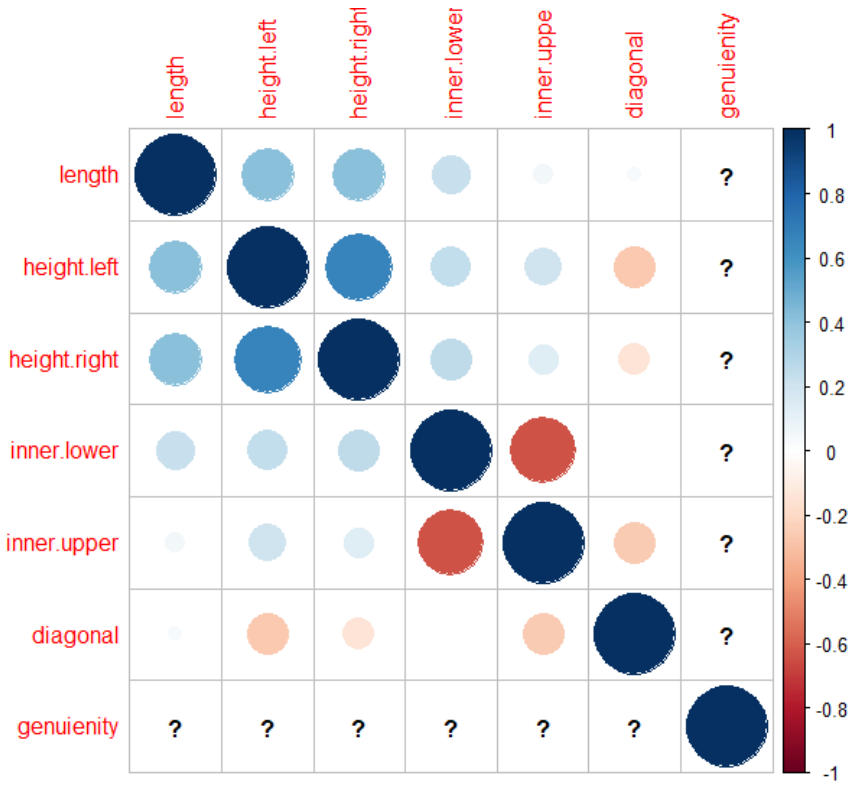


Below is the correlation plot for the same.



We can see that diagonal is well correlated with other parameters except for the length.

But for genuine set of rows from below correlation plot.



We can visualize that the parameters are not well correlated.

Below is the summary for the PCA for the given data.

```
Data:   X dimension: 200 6
        Y dimension: 200 1
Fit method: svdpc
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV          0.5025   0.2466   0.1923   0.1932   0.1436   0.1429   0.1436
adjCV       0.5025   0.2464   0.1920   0.1930   0.1434   0.1426   0.1433

TRAINING: % variance explained
            1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X             49.09    70.39    84.88    92.37    96.85   100.00
genuienity    76.20    85.66    85.66    92.13    92.36    92.42
```

It is evident that all the parameters are crucial for building a model for this dataset.

From the genuine data analysis

```
Data:   X dimension: 100 6
        Y dimension: 100 1
Fit method: svdpc
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV               0        0        0        0        0        0        0
adjCV            0        0        0        0        0        0        0

TRAINING: % variance explained
            1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X             36.73    65.01    81.11    90.82    96.26      100
genuienity      NaN      NaN      NaN      NaN      NaN      NaN
```

No conclusions car to be made as the PCA analysis we can see that the components are given equal weights.