Assignment3_E

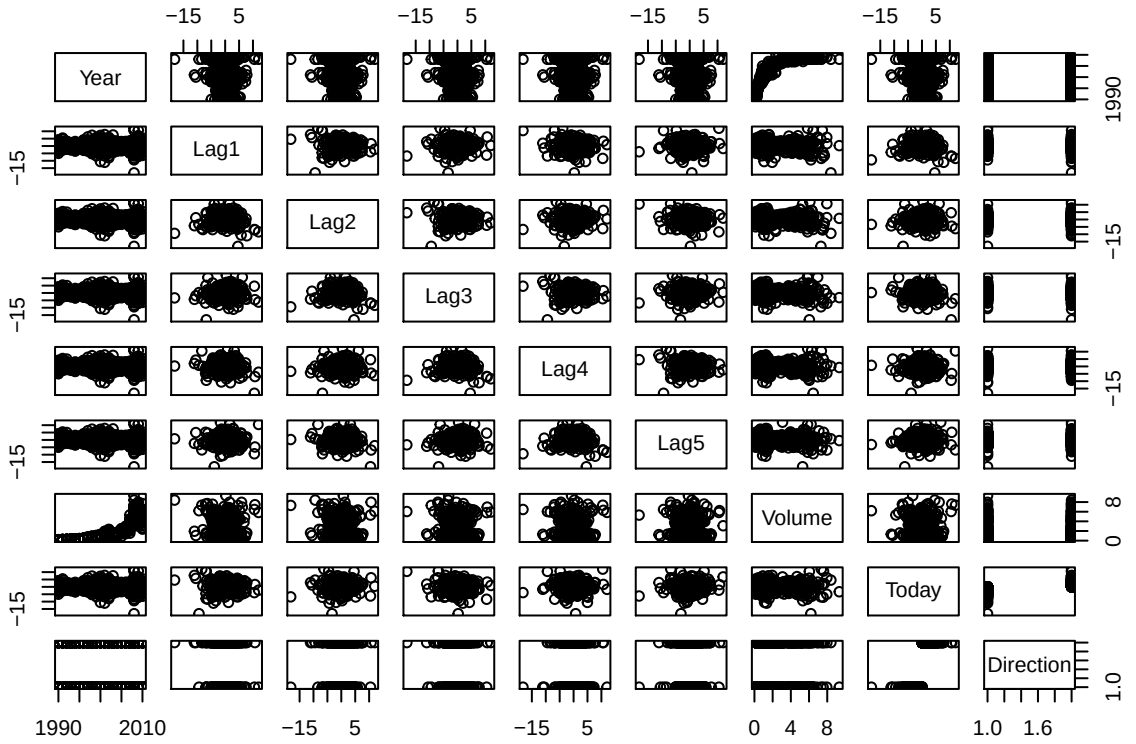
Sangamesh

3 December 2018

Q.1] Consider the Weekly data set, which is part of ISLR package. It contains the weekly stock market returns for 21 years.

a] Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any pattern?

```
##       Year           Lag1               Lag2               Lag3         
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410  
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472  
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090  
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260  
##       Lag4               Lag5              Volume       
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747  
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202  
##  Median :  0.2380   Median :  0.2340   Median :1.00268  
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821  
##      Today          Direction 
##  Min.   :-18.1950   Down:484  
##  1st Qu.: -1.1540   Up  :605  
##  Median :  0.2410             
##  Mean   :  0.1499             
##  3rd Qu.:  1.4050             
##  Max.   : 12.0260             
```



We can observe that the Weekly data from ISLR has Volume and Year taken together has logarithmic distribution.

b] Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appears to be statistically significant? If so, which ones?

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = "binomial", data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

Statistically significant predictor among the given is Lag2 only since the p-value is greater than the significant code attached to it.

c] Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

```
##       Weeklyglm.preds
##        Down  Up
##   Down   54 430
##   Up     48 557
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down  Up
##       Down   54 430
##       Up     48 557
##
##                Accuracy : 0.5611
##                  95% CI : (0.531, 0.5908)
##     No Information Rate : 0.9063
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.035
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.52941
##             Specificity : 0.56434
##          Pos Pred Value : 0.11157
##          Neg Pred Value : 0.92066
##              Prevalence : 0.09366
##          Detection Rate : 0.04959
##    Detection Prevalence : 0.44444
##       Balanced Accuracy : 0.54687
##
##        'Positive' Class : Down
##
```

There are a predominance of Up prediction. The model predicts well the Up direction, but it predict poorly the Down direction.

d] Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010.)

```
##       glm.preds.d
##        Down Up
##   Down    9 34
##   Up      5 56
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down    9 34
##       Up      5 56
##
##                Accuracy : 0.625
##                  95% CI : (0.5247, 0.718)
##     No Information Rate : 0.8654
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1414
##  Mcnemar's Test P-Value : 7.34e-06
##
##             Sensitivity : 0.64286
##             Specificity : 0.62222
##          Pos Pred Value : 0.20930
##          Neg Pred Value : 0.91803
##              Prevalence : 0.13462
##          Detection Rate : 0.08654
##    Detection Prevalence : 0.41346
##       Balanced Accuracy : 0.63254
##
##        'Positive' Class : Down
##
```

Overall fraction of correct predictions for the held out data is accuracy is 0.625

e] Repeat (d) using linear discriminant analysis (LDA).

```
##
##        Down Up
##   Down    9 34
##   Up      5 56
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down    9 34
##       Up      5 56
##
##                Accuracy : 0.625
##                  95% CI : (0.5247, 0.718)
##     No Information Rate : 0.8654
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1414
##  Mcnemar's Test P-Value : 7.34e-06
```

```
##
##             Sensitivity : 0.64286
##             Specificity : 0.62222
##          Pos Pred Value : 0.20930
##          Neg Pred Value : 0.91803
##              Prevalence : 0.13462
##          Detection Rate : 0.08654
##    Detection Prevalence : 0.41346
##       Balanced Accuracy : 0.63254
##
##        'Positive' Class : Down
##
```

Overall fraction of correct predictions for the held out data is accuracy is 0.625

f] Repeat (d) using quadratic discriminant analysis (QDA).

```
##
##        Down Up
##   Down    0 43
##   Up      0 61
```

```
## [1] 0.5865385
```

Overall fraction of correct predictions for the held out data is accuracy is 0.5865

g] Repeat (d) using KNN with =1.

```
##       knn.pred
##        Down Up
##   Down   21 22
##   Up     30 31
```

```
## [1] 0.5
```

Overall fraction of correct predictions for the held out data is accuracy is 0.5865

h] Which of these methods appears to provide the best results on this data?

The models from letter d and e, respectively Logistic Regression and LDA.

Q.2] This problem involves predicting Salary from the Hitters data set which is part of the ISLR package.

a] Remove the observations for whom the salary information is unknown, and then log-transform the salaries.

```
##      AtBat            Hits         HmRun            Runs
##  Min.   : 16.0   Min.   :  1   Min.   : 0.00   Min.   :  0.00
##  1st Qu.:255.2   1st Qu.: 64   1st Qu.: 4.00   1st Qu.: 30.25
##  Median :379.5   Median : 96   Median : 8.00   Median : 48.00
##  Mean   :380.9   Mean   :101   Mean   :10.77   Mean   : 50.91
##  3rd Qu.:512.0   3rd Qu.:137   3rd Qu.:16.00   3rd Qu.: 69.00
##  Max.   :687.0   Max.   :238   Max.   :40.00   Max.   :130.00
##
##       RBI             Walks            Years            CAtBat
```

```
##  Min.   :  0.00   Min.   :  0.00   Min.   : 1.000   Min.   :   19.0
##  1st Qu.: 28.00   1st Qu.: 22.00   1st Qu.: 4.000   1st Qu.:  816.8
##  Median : 44.00   Median : 35.00   Median : 6.000   Median : 1928.0
##  Mean   : 48.03   Mean   : 38.74   Mean   : 7.444   Mean   : 2648.7
##  3rd Qu.: 64.75   3rd Qu.: 53.00   3rd Qu.:11.000   3rd Qu.: 3924.2
##  Max.   :121.00   Max.   :105.00   Max.   :24.000   Max.   :14053.0
##
##      CHits            CHmRun           CRuns             CRBI
##  Min.   :   4.0   Min.   :  0.00   Min.   :   1.0   Min.   :   0.00
##  1st Qu.: 209.0   1st Qu.: 14.00   1st Qu.: 100.2   1st Qu.:  88.75
##  Median : 508.0   Median : 37.50   Median : 247.0   Median : 220.50
##  Mean   : 717.6   Mean   : 69.49   Mean   : 358.8   Mean   : 330.12
##  3rd Qu.:1059.2   3rd Qu.: 90.00   3rd Qu.: 526.2   3rd Qu.: 426.25
##  Max.   :4256.0   Max.   :548.00   Max.   :2165.0   Max.   :1659.00
##
##      CWalks        League  Division    PutOuts          Assists
##  Min.   :   0.00   A:175   E:157    Min.   :   0.0   Min.   :  0.0
##  1st Qu.:  67.25   N:147   W:165    1st Qu.: 109.2   1st Qu.:  7.0
##  Median : 170.50                    Median : 212.0   Median : 39.5
##  Mean   : 260.24                    Mean   : 288.9   Mean   :106.9
##  3rd Qu.: 339.25                    3rd Qu.: 325.0   3rd Qu.:166.0
##  Max.   :1566.00                    Max.   :1378.0   Max.   :492.0
##
##      Errors          Salary       NewLeague
##  Min.   : 0.00   Min.   :  67.5   A:176
##  1st Qu.: 3.00   1st Qu.: 190.0   N:146
##  Median : 6.00   Median : 425.0
##  Mean   : 8.04   Mean   : 535.9
##  3rd Qu.:11.00   3rd Qu.: 750.0
##  Max.   :32.00   Max.   :2460.0
##                  NA's   :59
```

b] Create a training set consisting of the first 200 observations, and a test set consisting of the remaining observations.

```
## [1] "Training data head: "
```

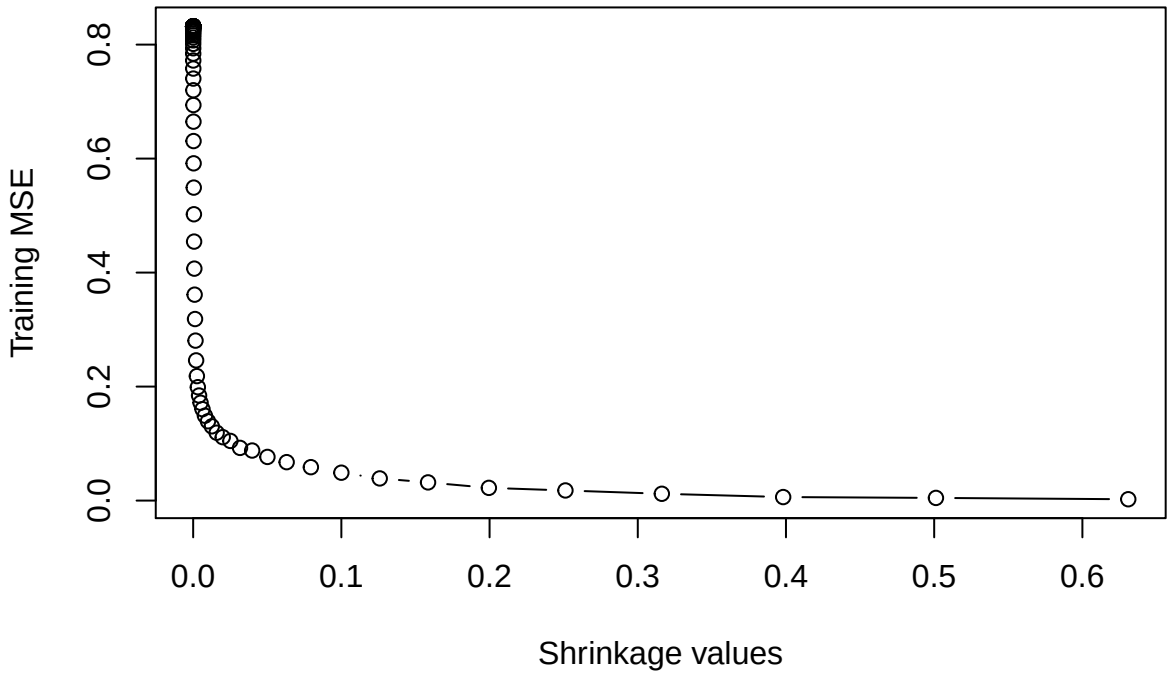
```
##                   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits
## -Alan Ashby         315   81     7   24  38    39    14   3449   835
## -Alvin Davis        479  130    18   66  72    76     3   1624   457
## -Andre Dawson       496  141    20   65  78    37    11   5628  1575
## -Andres Galarraga   321   87    10   39  42    30     2    396   101
## -Alfredo Griffin    594  169     4   74  51    35    11   4408  1133
## -Al Newman          185   37     1   23   8    21     2    214    42
##                   CHmRun CRuns CRBI CWalks League Division PutOuts Assists
## -Alan Ashby           69   321  414    375      N        W     632      43
## -Alvin Davis          63   224  266    263      A        W     880      82
## -Andre Dawson        225   828  838    354      N        E     200      11
## -Andres Galarraga     12    48   46     33      N        E     805      40
## -Alfredo Griffin      19   501  336    194      A        W     282     421
## -Al Newman             1    30    9     24      N        E      76     127
##                   Errors   Salary NewLeague
## -Alan Ashby           10 6.163315         N
```

```
## -Alvin Davis          14 6.173786         A
## -Andre Dawson          3 6.214608         N
## -Andres Galarraga      4 4.516339         N
## -Alfredo Griffin      25 6.620073         A
## -Al Newman             7 4.248495         A
```

```
## [1] "Test data head: "
```

```
##                  AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun
## -Reggie Jackson    419  101    18   65  58    92    20   9528  2510    548
## -Ron Kittle        376   82    21   42  60    35     5   1770   408    115
## -Ray Knight        486  145    11   51  76    40    11   3967  1102     67
## -Rick Leach        246   76     5   35  39    13     6    912   234     12
## -Rick Manning      205   52     8   31  27    17    12   5134  1323     56
## -Rance Mulliniks   348   90    11   50  45    43    10   2288   614     43
##                  CRuns CRBI CWalks League Division PutOuts Assists Errors
## -Reggie Jackson   1509 1659   1342      A        W       0       0      0
## -Ron Kittle        238  299    157      A        W       0       0      0
## -Ray Knight        410  497    284      N        E      88     204     16
## -Rick Leach        102   96     80      A        E      44       0      1
## -Rick Manning      643  445    459      A        E     155       3      2
## -Rance Mulliniks   295  273    269      A        E      60     176      6
##                    Salary NewLeague
## -Reggie Jackson  6.189290         A
## -Ron Kittle      6.052089         A
## -Ray Knight      6.214608         A
## -Rick Leach      5.521461         A
## -Rick Manning    5.991465         A
## -Rance Mulliniks 6.109248         A
```

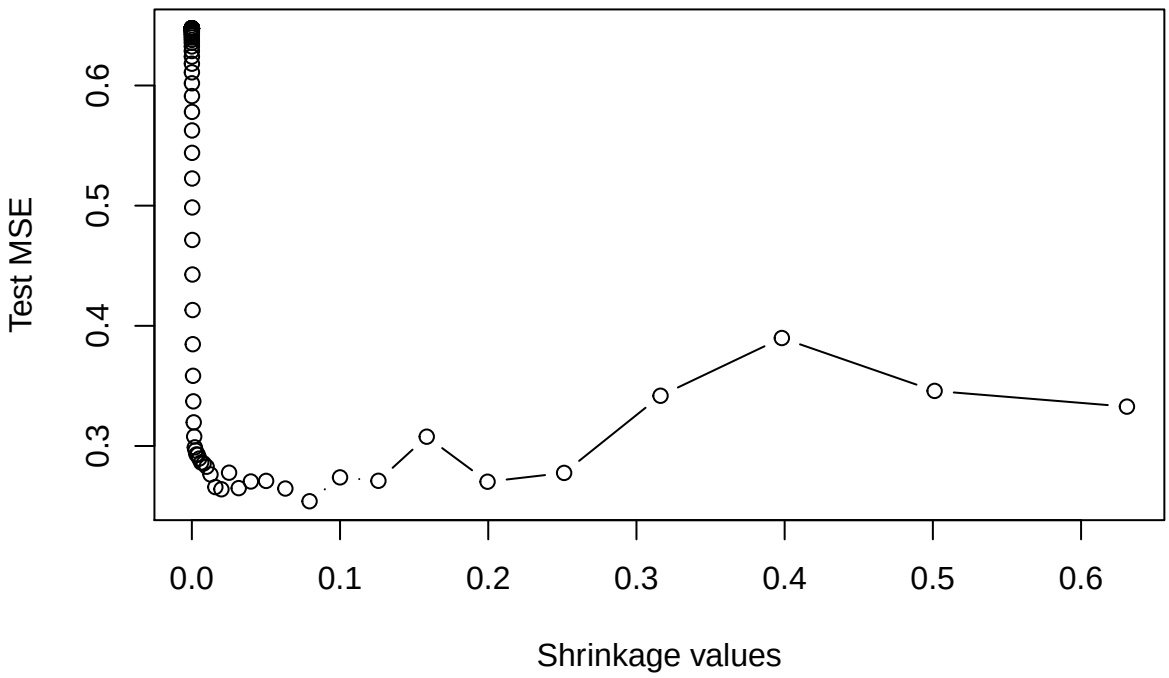
c] Perform boosting on the training set with 1000 trees for a range of values of the shrinkage parameter lambda. Produce a plot with different shrinkage values on the x-axis and the corresponding training set MSE on the y-



axis.

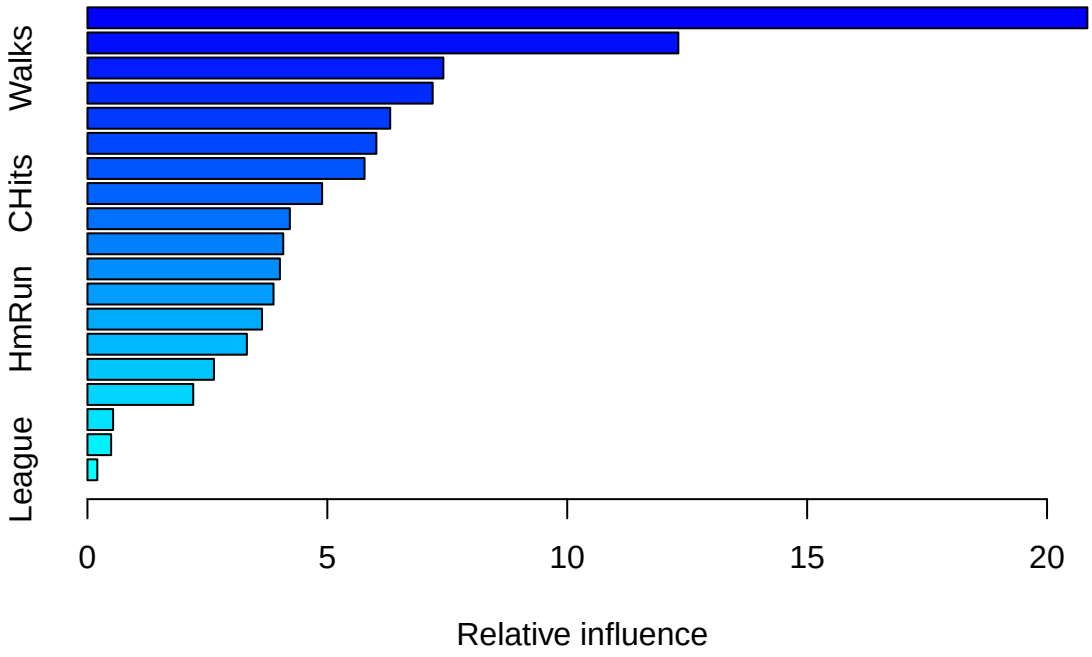
We observe, as shrinkage value increases the training MSE value exponentially decreases.

d] Produce a plot with different shrinkage values on the x-axis and the corresponding test set MSE on the y-axis.




```
## [1] "The minimum test MSE is  0.254026510444201  , and is obtained for lambda = 0.0794328234724282"
```

e] Which variable appear to be the most important predictors in the boosted model?




```
##                 var    rel.inf
## CAtBat       CAtBat 20.8404970
## CRBI           CRBI 12.3158959
## Walks         Walks  7.4186037
## PutOuts     PutOuts  7.1958539
## Years         Years  6.3104535
## CWalks       CWalks  6.0221656
## CHmRun       CHmRun  5.7759763
## CHits         CHits  4.8914360
## AtBat         AtBat  4.2187460
## RBI             RBI  4.0812410
## Hits           Hits  4.0117255
## Assists     Assists  3.8786634
## HmRun         HmRun  3.6386178
## CRuns         CRuns  3.3230296
## Errors       Errors  2.6369128
## Runs           Runs  2.2048386
## Division   Division  0.5347342
## NewLeague NewLeague  0.4943540
## League       League  0.2062551
```

We see that CAtBat is most important variable in all the variables list, relatively. Also, relative influence of Walks is found to be highest.

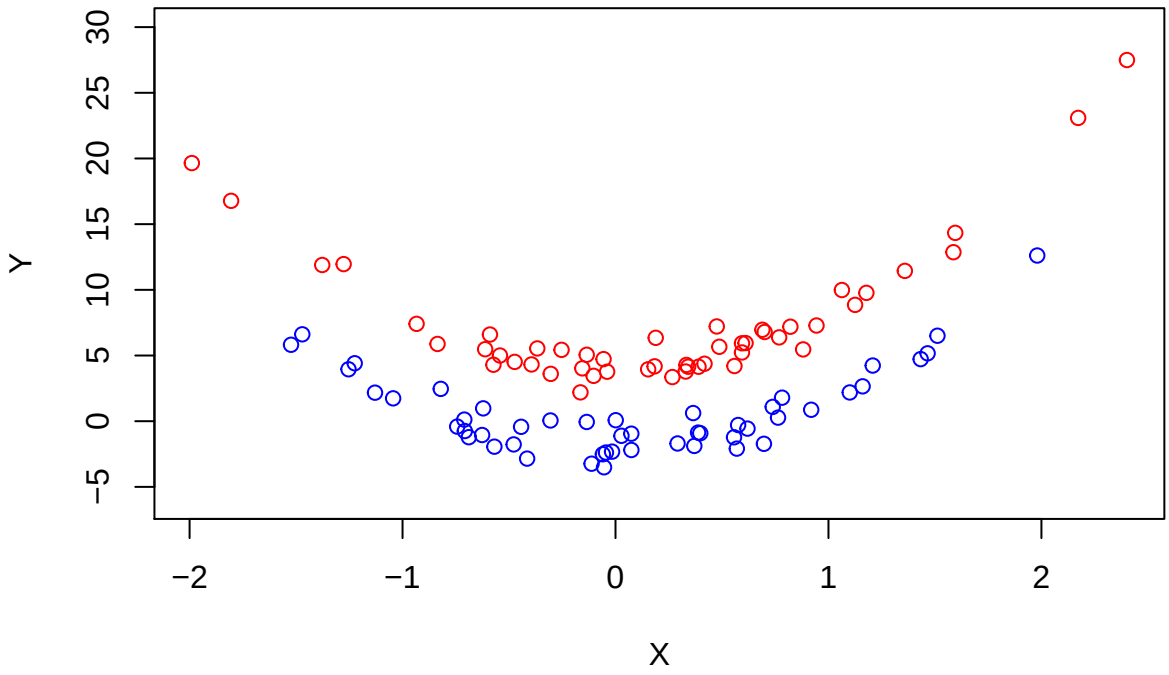
f] Apply bagging to the training set. What is the test set MSE for this approach.

```
## [1] "The test MSE for bagging is 0.22993242086693, which is slightly lower than the test MSE for boos
```

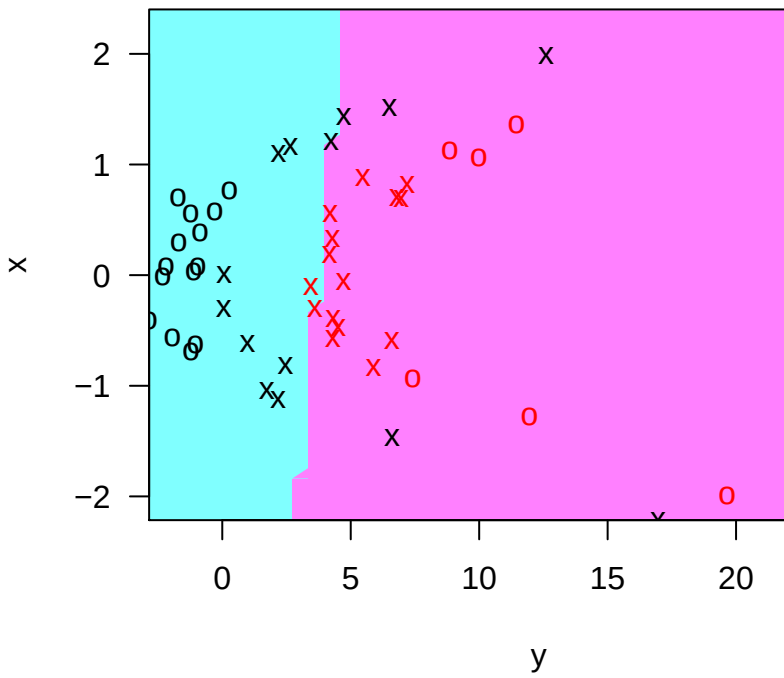
g] Apply random forests to the training set. What is the test set MSE for this approach.

```
## [1] "The test MSE for Random Forest is 0.214033998567829, which is slightly lower than the test MSE f
```

Q.3] Generate a simulated two-class data set with 100 observations and two features in which there is a visible but non-linear seperation between the classes. Show that in this setting, a support vector machine with a polynomial kernal (with degree greater than 1) or a radial kernal will outperform a support vector classifier on the training data. Which technique performs best on the test data? Make plots and report training and test error rates in order to back up your assertions.



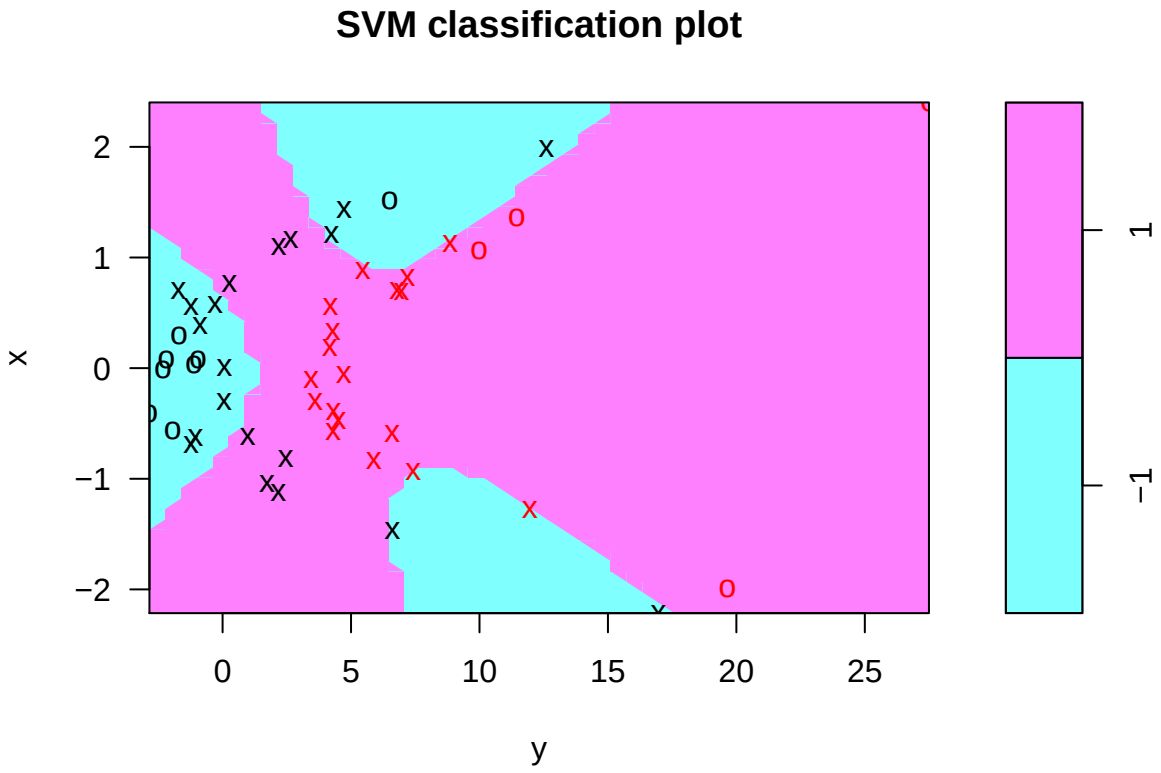
We can clearly see the seperation between two classes - Non linear



Now, we fit a support vector classifier on the training data

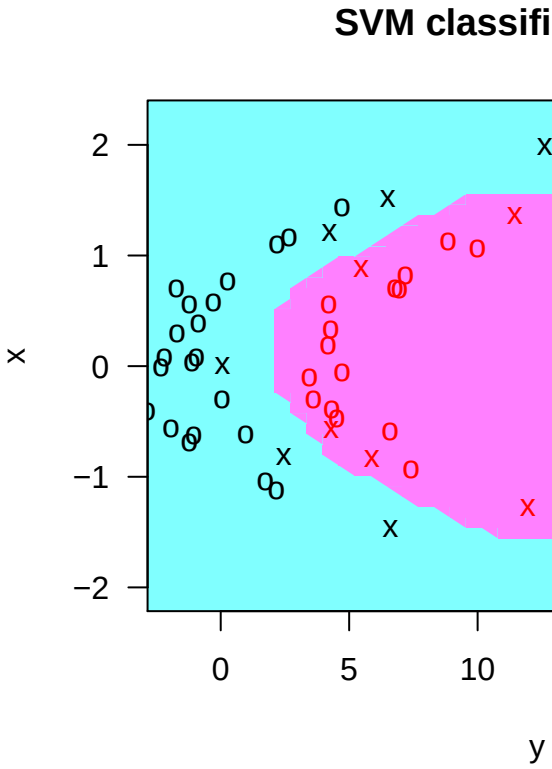
```
##         truth
## predict -1  1
##      -1 22  0
##      1   6 22
```

The support vector classifier makes 6 errors on the training data. Next, we fit a support vector machine with a polynomial kernel.




```
##         truth
## predict -1  1
##      -1 19  0
##      1   9 22
```

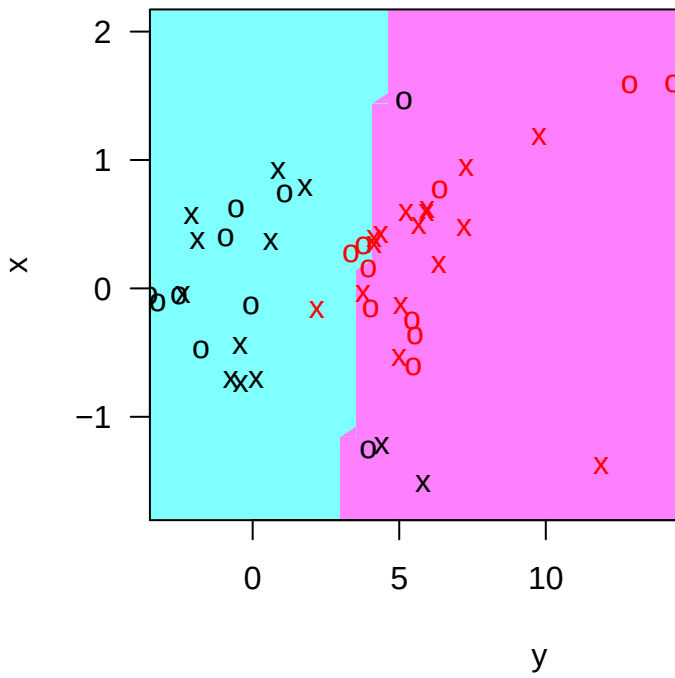
The support vector machine with a polynomial kernel of degree 3 makes 9 errors on the training data.



Finally, we fit a support vector machine with a radial kernel and a gamma of 1.

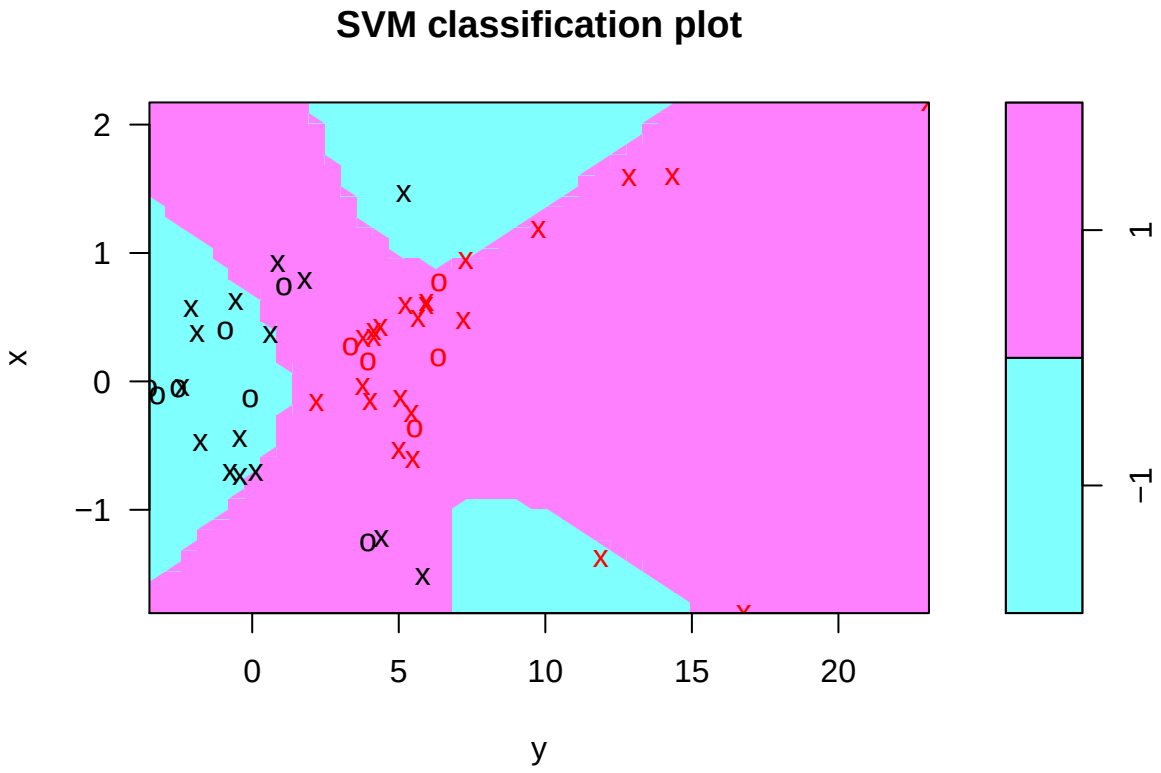
```
##         truth
## predict -1  1
##      -1 28  0
##      1   0 22
```

The support vector machine with a radial kernel makes 0 error on the training data.

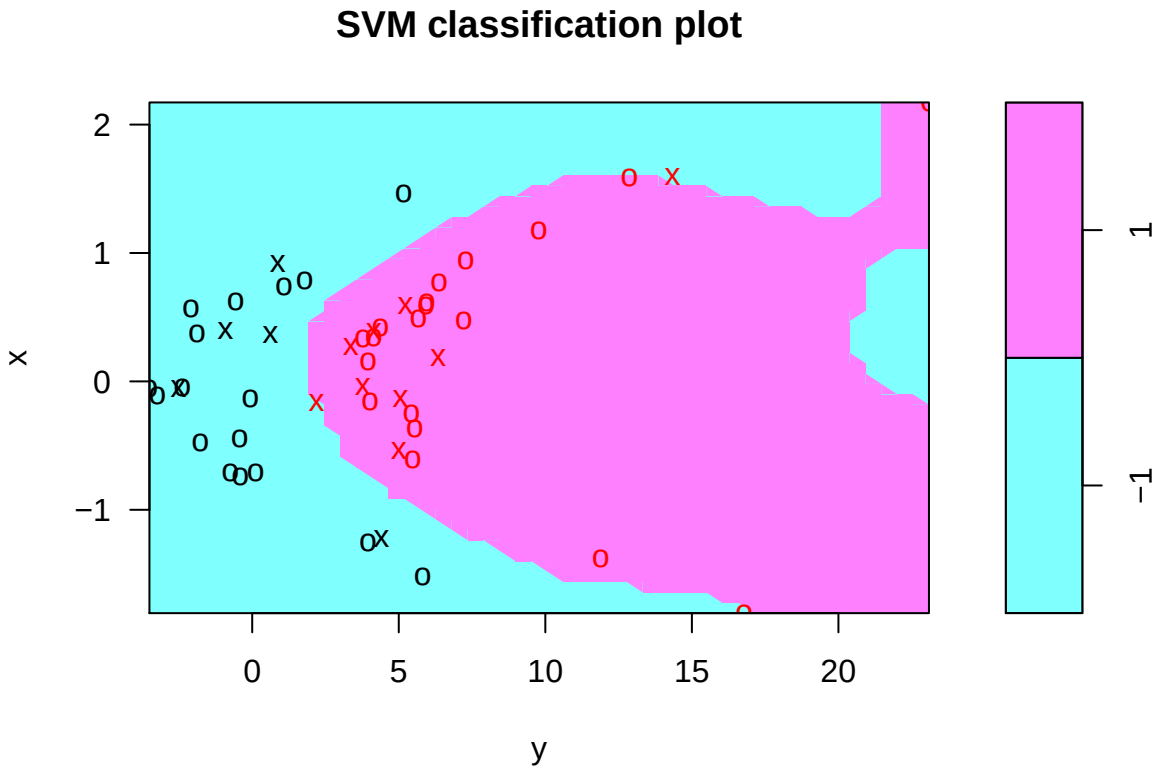


Now, we check how these models fare when applied to the test data.

```
##         truth
## predict -1  1
##      -1 18  2
##      1   4 26
```




```
##         truth
## predict -1  1
##      -1 14  1
##       1  8 27
```




```
##         truth
## predict -1  1
##      -1 22  1
##      1   0 27
```

We may see that the linear, polynomial and radial support vector machines classify respectively 9, 6 and 1 observations incorrectly. So, radial kernel is the best model in this setting.