HW3_P1_RMD

Naveen Nagarajan

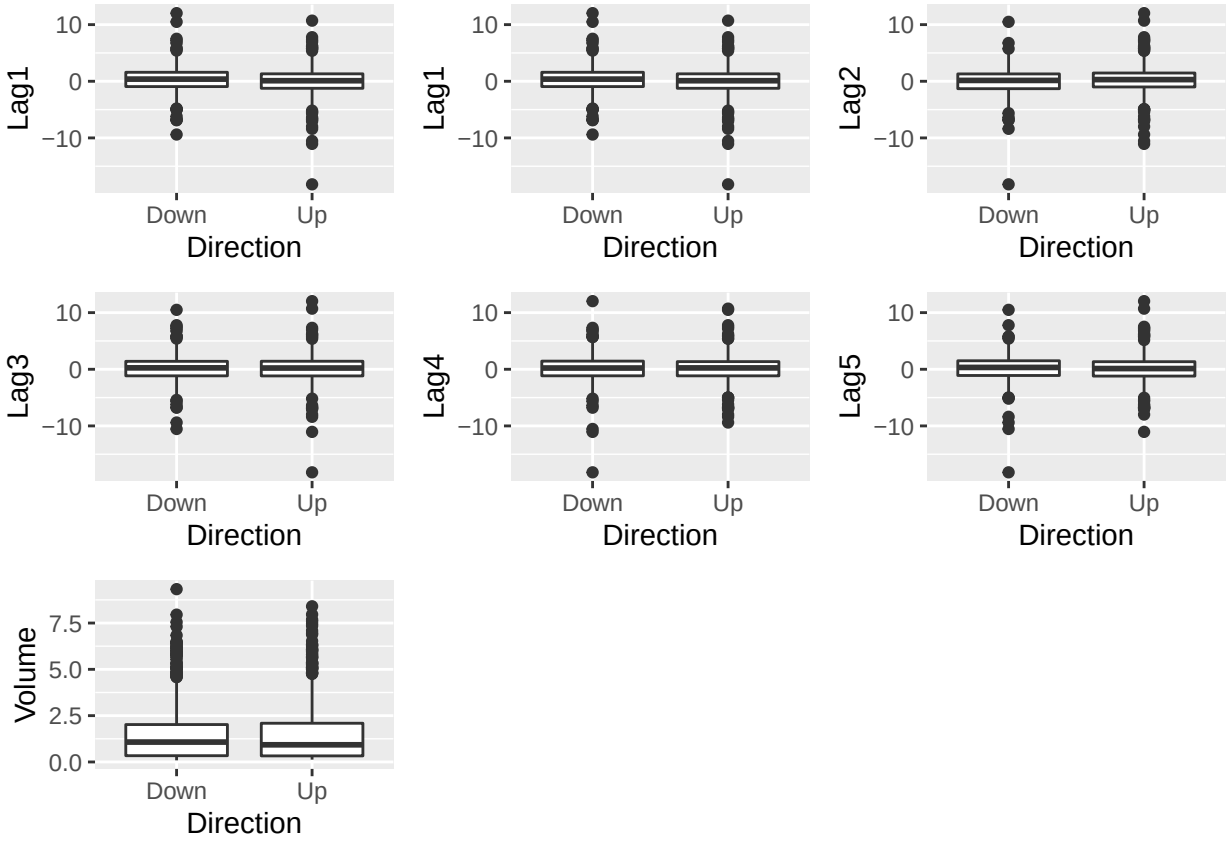
4/5/2021

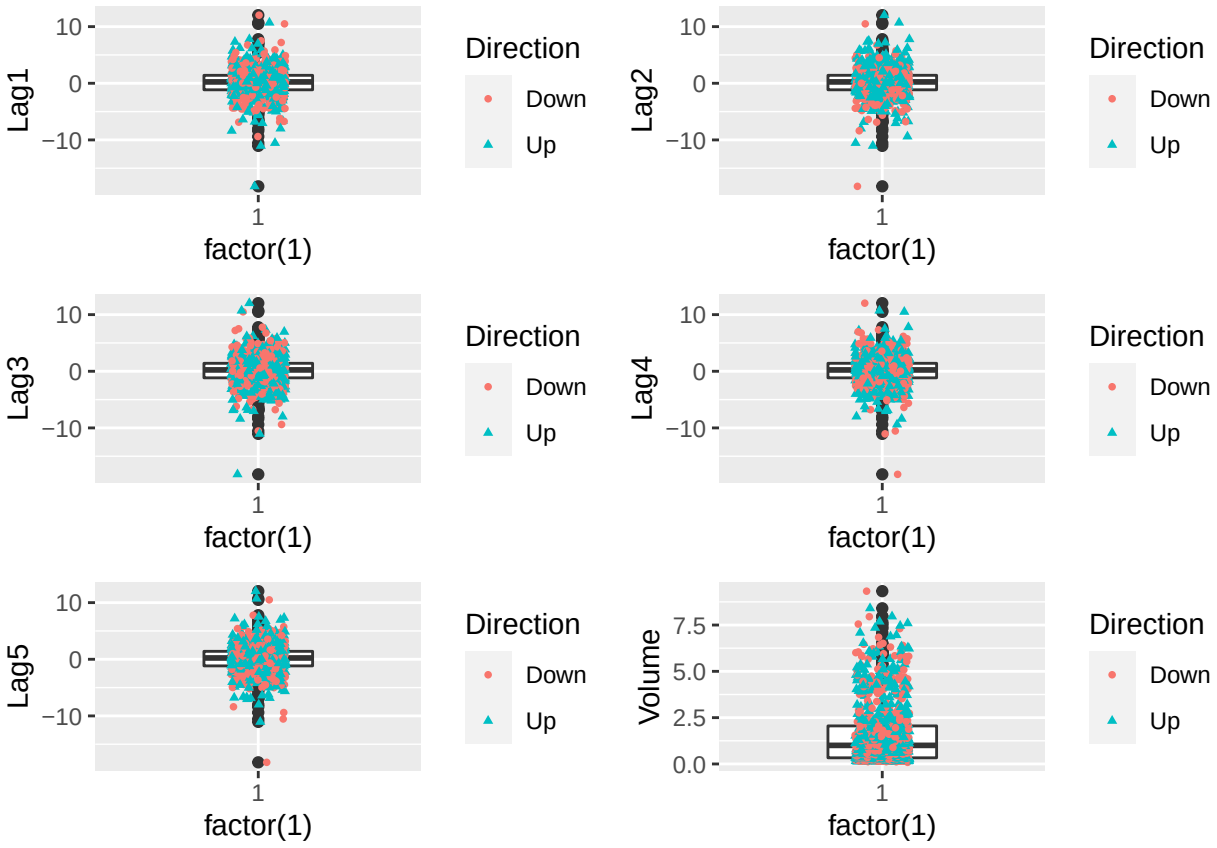
This question should be answered using the Weekly data set, which is part of the ISLR package in R. The file have been included in the assignment as Weekly.csv. It contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

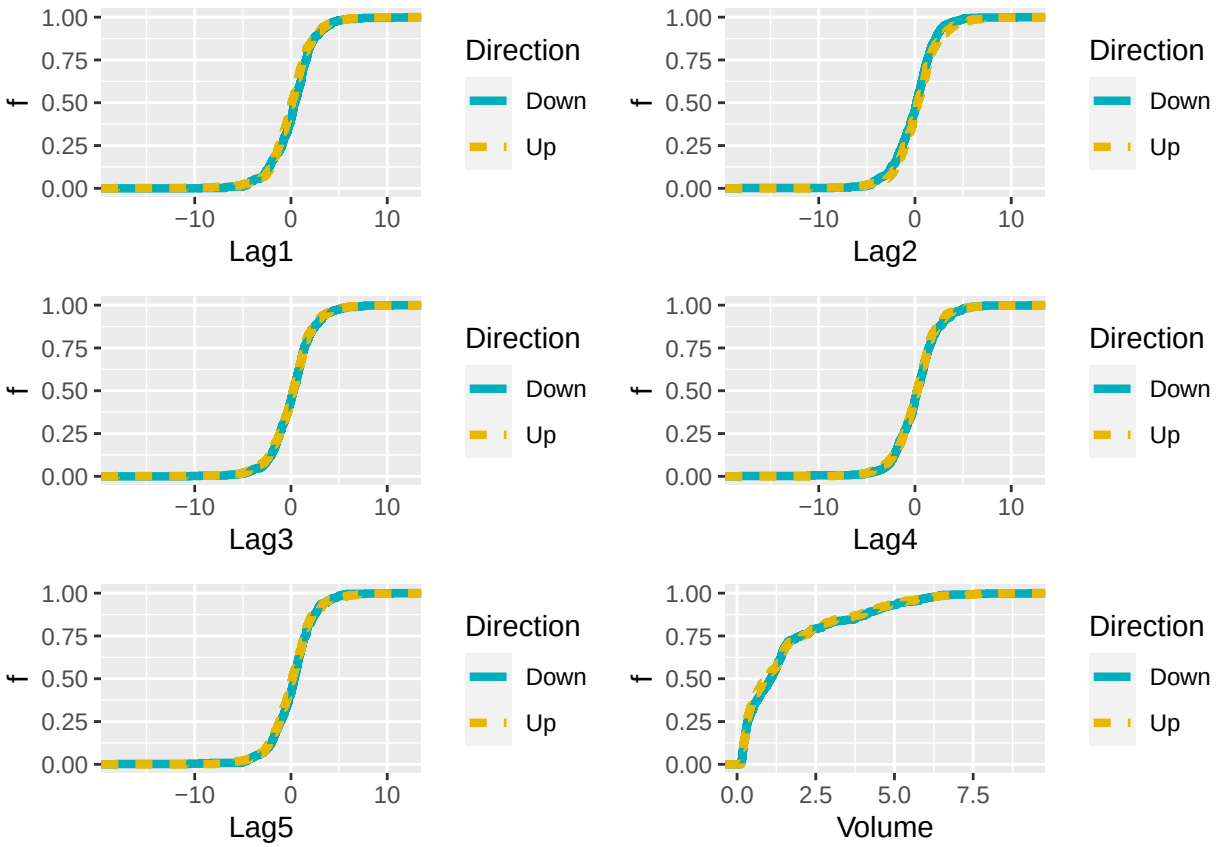
a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

```
##       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            Today        
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950  
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540  
##  Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410  
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260  
##  Direction 
##  Down:484  
##  Up  :605  
##            
##            
##            
##            
```

```
## adding dummy grobs
```


Lag1
10
0
−10
Down
Up
Direction
Lag1
10
0
−10
Down
Up
Direction
Lag2
10
0
−10
Down
Up
Direction
Lag3
10
0
−10
Down
Up
Direction
Lag4
10
0
−10
Down
Up
Direction
Lag5
10
0
−10
Down
Up
Direction
Volume
7.5
5.0
2.5
0.0
Down
Up
Direction


Lag1
10
0
−10
1
factor(1)
Direction
Down
Up
Lag2
10
0
−10
1
factor(1)
Direction
Down
Up
Lag3
10
0
−10
1
factor(1)
Direction
Down
Up
Lag4
10
0
−10
1
factor(1)
Direction
Down
Up
Lag5
10
0
−10
1
factor(1)
Direction
Down
Up
Volume
7.5
5.0
2.5
0.0
1
factor(1)
Direction
Down
Up



From the plots we can see there is no correlation between lags and volume against Direction

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 appear to be statistically significant? If so, which ones?

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = ISLR::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
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %     97.5 %
## (Intercept)  0.098808746 0.43580101
## Lag1        -0.093477110 0.01029269
## Lag2         0.006197597 0.11169774
## Lag3        -0.068653910 0.03604309
## Lag4        -0.079952378 0.02401603
## Lag5        -0.066495108 0.03711989
## Volume      -0.095051949 0.04979338
```

```
##                    2.5 %     97.5 %
## (Intercept)  0.098445204 0.43528308
## Lag1        -0.093032105 0.01049422
## Lag2         0.005787254 0.11109610
## Lag3        -0.068319640 0.03619735
## Lag4        -0.079657357 0.02407694
## Lag5        -0.066185275 0.03724115
## Volume      -0.095060526 0.04957746
```

From above Deviance table we can see p for lag2 is between 0.01 and 0.05 which is somewhat significant for the model. For response variable Direction which is a factor of (Down,Up) last value is considered as event

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.

```
##   Sensitivity Specificity Error rate
## 1   0.9206612   0.1115702  0.4389348
```

```
##
##        Down  Up
##   Down   54 430
##   Up     48 557
```

Model has low specificity and high error rate, when there are more than one predictor variable Logistic regression isn't suitable

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).

```
##   Sensitivity Specificity Error rate
## 1   0.9180328   0.2093023      0.375
```

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

e) Repeat d) using LDA

```
##   Sensitivity Specificity Error rate
## 1   0.9180328   0.2093023      0.375
```

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

f) Repeat d) using QDA

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

```
##   Sensitivity Specificity Error rate
## 1           1           0  0.4134615
```

g) Repeat d) using KNN with K = 1.

```
##
##        Down Up
##   Down   21 22
##   Up     29 32
```

```
##   Sensitivity Specificity Error rate
## 1   0.5245902   0.4883721  0.4903846
```

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

```
##   Sensitivity Specificity Error rate    class predictors
## 1   0.9180328   0.2093023  0.3750000 logistic       Lag2
## 2   0.9180328   0.2093023  0.3750000      lda       Lag2
## 3   1.0000000   0.0000000  0.4134615      qda       Lag2
## 4   0.5245902   0.4883721  0.4903846      knn       Lag2
```

Lda and Logistic regression seems to better in terms of sensitivity. Knn-1 did better with specificity but error rate is about 50%

i) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

Knn optimized

- Knn-2 with Lag2

```
##
##        Down Up
##   Down   22 21
##   Up     29 32
```

```
##   Sensitivity Specificity Error rate
## 1   0.5245902   0.5116279  0.4807692
```

- Knn-3 with Lag2

```
##
##        Down Up
##   Down   16 27
##   Up     19 42
```

```
##   Sensitivity Specificity Error rate
## 1   0.6885246    0.372093  0.4423077
```

- Knn-5 with Lag2

```
##
##        Down Up
##   Down   16 27
##   Up     20 41
```

```
##   Sensitivity Specificity Error rate
## 1   0.6721311    0.372093  0.4519231
```

- Knn-2 with Lag1+Lag2

```
##
##        Down Up
##   Down   21 22
##   Up     35 26
```

```
##   Sensitivity Specificity Error rate
## 1   0.4262295   0.4883721  0.5480769
```

- Knn-3 with Lag1+Lag2

```
##
##        Down Up
##   Down   16 27
##   Up     19 42
```

```
##   Sensitivity Specificity Error rate
## 1   0.6885246    0.372093  0.4423077
```

- GLM with Lag1+Lag2

```
##
##        Down Up
##   Down    7 36
##   Up      8 53
```

```
##   Sensitivity Specificity Error rate
## 1   0.8688525   0.1627907  0.4230769
```

- LDA with Lag1+(Lag2*Lag2)

```
##
##        Down Up
##   Down    7 36
##   Up      8 53
```

```
##   Sensitivity Specificity Error rate
## 1   0.8688525   0.1627907  0.4230769
```

- QDA with Lag1+(Lag2*Lag2)

```
##
##        Down Up
##   Down    7 36
##   Up     10 51
```

```
##   Sensitivity Specificity Error rate
## 1   0.8360656   0.1627907  0.4423077
```

- Performance of all models together

```
##    Sensitivity Specificity Error rate    class       predictors
## 1    0.9180328   0.2093023  0.3750000 logistic             Lag2
## 2    0.9180328   0.2093023  0.3750000      lda             Lag2
## 3    1.0000000   0.0000000  0.4134615      qda             Lag2
## 4    0.5245902   0.4883721  0.4903846      knn             Lag2
## 5    0.5245902   0.5116279  0.4807692    knn-2             Lag2
## 6    0.6885246   0.3720930  0.4423077    knn-3             Lag2
## 7    0.6721311   0.3720930  0.4519231    knn-5             Lag2
## 8    0.4262295   0.4883721  0.5480769    knn-2        Lag1+Lag2
## 9    0.6885246   0.3720930  0.4423077    knn-3        Lag1+Lag2
## 10   0.8688525   0.1627907  0.4230769 logistic        Lag1+Lag2
## 11   0.8688525   0.1627907  0.4230769      lda Lag1+(Lag2*Lag2)
## 12   0.8360656   0.1627907  0.4423077      qda Lag1+(Lag2*Lag2)
```