Problem 3(LDA, QDA, KNN)

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.

In [1]:
```
#install.packages('ISLR')
#install.packages('aod')
#install.packages('ggplot2')
#install.packages('MASS')
#install.packages('class')
```

In [2]:
```
library(ISLR)
library(aod)
library(ggplot2)
library(MASS)
library(class)
```

In [3]:
```
head(Weekly)
```

Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
1990	0.816	1.572	-3.936	-0.229	-3.484	0.1549760	-0.270	Down
1990	-0.270	0.816	1.572	-3.936	-0.229	0.1485740	-2.576	Down
1990	-2.576	-0.270	0.816	1.572	-3.936	0.1598375	3.514	Up
1990	3.514	-2.576	-0.270	0.816	1.572	0.1616300	0.712	Up
1990	0.712	3.514	-2.576	-0.270	0.816	0.1537280	1.178	Up
1990	1.178	0.712	3.514	-2.576	-0.270	0.1544440	-1.372	Down

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

In [4]:
```
summary(Weekly)
```

```
      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
```

Answer: The differences among lag1, lag2, lag3, lag4 and lag5 seems slightly.
Min and max for both of them are same.
Mean for them also seems similar.

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?

In [5]:
```
glm.fit = glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data = Weekly, fami
summary(glm.fit)
```

```
Call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
    Volume, family = binomial(link = "logit"), 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
```

Answer: Lag1 and Lag2 are the two predictors which will be statistically significant.

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.

In [6]:
```
glm.probs=predict(glm.fit,newdata = Weekly, type="response")
glm.pred=rep("Down",dim(Weekly)[1])
glm.pred[glm.probs>0.5]="Up"
prediction.glm=cbind(Weekly,glm.pred)
colnames(prediction.glm)[10]="Direction prediction"
contrasts(Weekly$Direction)
table(glm.pred, Weekly$Direction)
```

	Up
Down	0
Up	1

```
glm.pred Down  Up
    Down   54  48
    Up    430 557
```

Assume Down - False ('0'); Up - True('1')
TypeI: The model incorrectly predict Up and create 430 errors.
TypeII: The model incorrectly predict Down and create 48 errors.

In [7]:
```
mean(glm.pred == Weekly$Direction)
54/(54+430)
557/(557+48)
```

0.561065197428834

0.111570247933884

0.920661157024793

Accuracy: 0.5610
Sensitivity: 0.1115
Specificity: 0.9206

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).

In [8]:
```
training=Weekly[Weekly$Year <= 2008,]
test=Weekly[Weekly$Year > 2008, ]
```

In [9]:
```
glm.fit=glm(Direction~Lag2,data=training,family=binomial(link = 'logit'))
glm.probs=predict(glm.fit,newdata = test, type="response")
glm.pred=rep("Down",dim(training)[1])
glm.pred[glm.probs>0.5]="Up"
prediction.glm=cbind(training,glm.pred)
colnames(prediction.glm)[10]="Direction prediction"
table(glm.pred, training$Direction)
```

```
glm.pred Down  Up
    Down   61  73
    Up    380 471
```

In [10]:
```
mean(glm.pred == training$Direction)
54/(54+430)
557/(557+48)
```

0.54010152284264

0.111570247933884

0.920661157024793

Accuracy: 0.5401
Sensitivity: 0.1115
Specificity: 0.9206

e) Repeat d) using LDA.

In [11]:
```
lda.fit = lda(Direction~Lag2,data=training)
lda.pred = predict(lda.fit, test)
lda.class = lda.pred$class
prediction.lda=cbind(test, lda.class)
colnames(prediction.lda)[10] = "Direction prediction"
head(prediction.lda)
table(lda.class, test$Direction)
mean(lda.class == test$Direction)
```

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down	Up
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down	Up
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down	Down
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down	Down
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up	Up
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down	Up

```
lda.class Down Up
     Down    9  5
     Up     34 56
```

0.625

Accuracy: 0.625

f) Repeat d) using QDA.

In [12]:
```
qda.fit = qda(Direction~Lag2, data = training)
qda.pred = predict(qda.fit, test)
qda.class = qda.pred$class
prediction.qda=cbind(test, qda.class)
colnames(prediction.qda)[10] = "Direction prediction"
head(prediction.qda)
table(qda.class, test$Direction)
mean(qda.class == test$Direction)
```

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down	Up
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down	Up
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down	Up
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down	Up
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up	Up
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down	Up

```
qda.class Down Up
     Down    0  0
     Up     43 61
```

0.586538461538462

Accuracy = 0.5865

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

In [13]:
```
test.x=cbind(test$index,test$Lag2)
training.x=cbind(training$index,training$Lag2)
test.x=cbind(test$Lag2)
knn.pred=knn(training.x,test.x,training$Direction,k=1)
prediction.knn=cbind(test,knn.pred)
colnames(prediction.knn)[10]="Direction prediction"
head(prediction.knn)
table(knn.pred ,test$Direction)
```

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down	Up
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down	Up
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down	Down
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down	Down
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up	Down
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down	Up

```
knn.pred Down Up
    Down   21 29
    Up     22 32
```

In [23]:
```
mean(knn.pred == test$Direction)
21/(21+22)
31/(31+30)
```

0.480769230769231

0.488372093023256

0.508196721311475

Accuracy: 0.5096
Sensitivity: 0.4883
Specificity: 0.5081

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

Logistic Regression Accuracy: 0.5401
LDA Accuracy: 0.625
QDA Accuracy = 0.5865
KNN Accuracy: 0.5096
So the LDA model provide the best results on this data

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.)

- Test KNN with K = 5

In [15]:
```
test.x=cbind(test$index,test$Lag2)
training.x=cbind(training$index,training$Lag2)
test.x=cbind(test$Lag2)
knn.pred=knn(training.x,test.x,training$Direction,k=5)
prediction.knn=cbind(test,knn.pred)
colnames(prediction.knn)[10]="Direction prediction"
head(prediction.knn)
table(knn.pred ,test$Direction)
mean(knn.pred == test$Direction)
```

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down	Up
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down	Up
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down	Down
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down	Down
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up	Up
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down	Up

```
knn.pred Down Up
    Down   15 22
    Up     28 39
```

0.519230769230769

Accuracy: 0.5192

- Test KNN with K = 10

In [16]:
```
test.x=cbind(test$index,test$Lag2)
training.x=cbind(training$index,training$Lag2)
test.x=cbind(test$Lag2)
knn.pred=knn(training.x,test.x,training$Direction,k=10)
prediction.knn=cbind(test,knn.pred)
colnames(prediction.knn)[10]="Direction prediction"
head(prediction.knn)
table(knn.pred ,test$Direction)
mean(knn.pred == test$Direction)
```

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down	Down
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down	Up
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down	Down
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down	Down
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up	Up
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down	Up

```
knn.pred Down Up
    Down   15 20
    Up     28 41
```

0.538461538461538

Accuracy: 0.5384

- Test Logistic Regression, LDA, QDA, KNN - K=1
- With Training (1990 to 2009) + test(2009 - 2010);
- Predictor Lag2

In [17]:
```
training=Weekly[Weekly$Year <= 2009,]
test=Weekly[Weekly$Year > 2009, ]
```

```
- Logistic Regression
```

In [18]:
```
glm.fit=glm(Direction~Lag2,data=training,family=binomial(link = 'logit'))
glm.probs=predict(glm.fit,newdata = test, type="response")
glm.pred=rep("Down",dim(training)[1])
glm.pred[glm.probs>0.5]="Up"
prediction.glm=cbind(training,glm.pred)
colnames(prediction.glm)[10]="Direction prediction"
table(glm.pred, training$Direction)
mean(glm.pred == training$Direction)
```

```
glm.pred Down  Up
    Down   20  40
    Up    444 533
```

0.533269045323047

Accuracy: 0.5332

```
- LDA
```

In [19]:
```
lda.fit = lda(Direction~Lag2,data=training)
lda.pred = predict(lda.fit, test)
lda.class = lda.pred$class
prediction.lda=cbind(test, lda.class)
colnames(prediction.lda)[10] = "Direction prediction"
head(prediction.lda)
table(lda.class, test$Direction)
mean(lda.class == test$Direction)
```

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
1038	2010	-1.010	2.178	-0.356	0.039	1.328	2.390427	2.680	Up	Up
1039	2010	2.680	-1.010	2.178	-0.356	0.039	4.223070	-0.782	Down	Up
1040	2010	-0.782	2.680	-1.010	2.178	-0.356	4.363246	-3.897	Down	Up
1041	2010	-3.897	-0.782	2.680	-1.010	2.178	5.654582	-1.639	Down	Up
1042	2010	-1.639	-3.897	-0.782	2.680	-1.010	5.079534	-0.715	Down	Up
1043	2010	-0.715	-1.639	-3.897	-0.782	2.680	5.082238	0.874	Up	Up

```
lda.class Down Up
     Down    3  0
     Up     17 32
```

0.673076923076923

Accuracy: 0.6730

```
- QDA
```

In [20]:
```
qda.fit = qda(Direction~Lag2, data = training)
qda.pred = predict(qda.fit, test)
qda.class = qda.pred$class
prediction.qda=cbind(test, qda.class)
colnames(prediction.qda)[10] = "Direction prediction"
head(prediction.qda)
table(qda.class, test$Direction)
mean(qda.class == test$Direction)
```

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
1038	2010	-1.010	2.178	-0.356	0.039	1.328	2.390427	2.680	Up	Up
1039	2010	2.680	-1.010	2.178	-0.356	0.039	4.223070	-0.782	Down	Up
1040	2010	-0.782	2.680	-1.010	2.178	-0.356	4.363246	-3.897	Down	Up
1041	2010	-3.897	-0.782	2.680	-1.010	2.178	5.654582	-1.639	Down	Up
1042	2010	-1.639	-3.897	-0.782	2.680	-1.010	5.079534	-0.715	Down	Up
1043	2010	-0.715	-1.639	-3.897	-0.782	2.680	5.082238	0.874	Up	Up

```
qda.class Down Up
     Down    0  0
     Up     20 32
```

0.615384615384615

Accuracy: 0.6153

- KNN(k=1)

In [21]:
```
test.x=cbind(test$index,test$Lag2)
training.x=cbind(training$index,training$Lag2)
test.x=cbind(test$Lag2)
knn.pred=knn(training.x,test.x,training$Direction,k=1)
prediction.knn=cbind(test,knn.pred)
colnames(prediction.knn)[10]="Direction prediction"
head(prediction.knn)
table(knn.pred ,test$Direction)
mean(knn.pred == test$Direction)
```

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
1038	2010	-1.010	2.178	-0.356	0.039	1.328	2.390427	2.680	Up	Down
1039	2010	2.680	-1.010	2.178	-0.356	0.039	4.223070	-0.782	Down	Down
1040	2010	-0.782	2.680	-1.010	2.178	-0.356	4.363246	-3.897	Down	Up
1041	2010	-3.897	-0.782	2.680	-1.010	2.178	5.654582	-1.639	Down	Up
1042	2010	-1.639	-3.897	-0.782	2.680	-1.010	5.079534	-0.715	Down	Down
1043	2010	-0.715	-1.639	-3.897	-0.782	2.680	5.082238	0.874	Up	Up

```
knn.pred Down Up
    Down    8 15
    Up     12 17
```

0.480769230769231

Predictor: Lag2 With Training (1990 to 2008) + test(2009 - 2010); Logistic Regression Accuracy: 0.5401
LDA Accuracy: 0.625
QDA Accuracy = 0.5865
KNN Accuracy: 0.5096

Predictor: Lag2 With Training (1990 to 2005) + test(2006 - 2010); Logistic Regression Accuracy: 0.5332
LDA Accuracy: 0.6730
QDA Accuracy = 0.6153
KNN Accuracy: 0.4807

In [22]:
```
#End
```