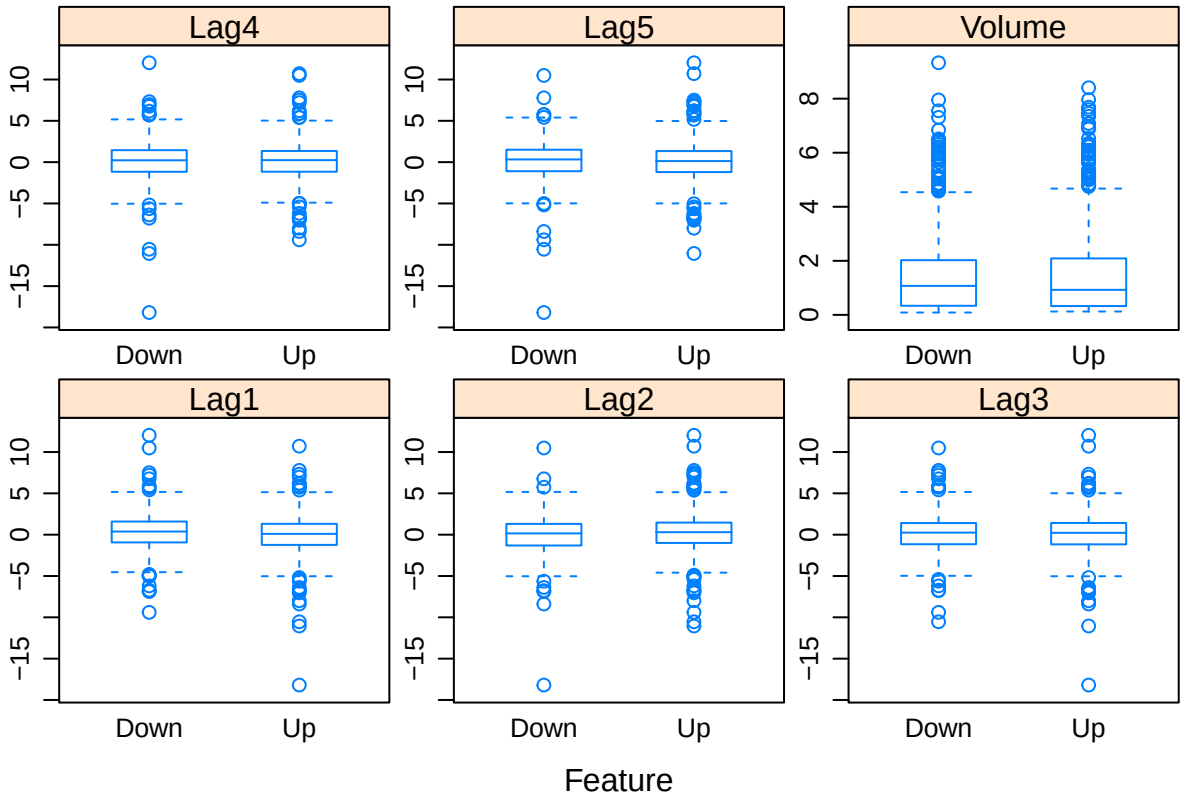
DS2_HW3

Siyan Chen

4/6/2019

(a) Produce some graphical summaries of theWeeklydata.

```
featurePlot(x = df[, 2:7],
            y = df$Direction,
            scales = list(x=list(relation = "free"),
                          y=list(relation = "free")),
            plot = "box", pch = "|")
```



(b) Use the full data set to perform a logistic regression withDirectionas the responseand the five Lagv ariables plus Volumeas predictors. Do any of the predictors appearto be statistically significant? If so, which ones?

```
set.seed(1)
rowTrain = createDataPartition(y = df$Direction,
                               p = 0.75,
                               list = FALSE)
glm_fit = glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,
              data = df,
              subset = rowTrain,
```

```
              family = binomial)
contrasts(df$Direction)
```

```
##      Up
## Down  0
## Up    1
```

```
summary(glm_fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = df, subset = rowTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8407  -1.2503   0.9628   1.0737   1.6492
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.31400    0.10037   3.129  0.00176 **
## Lag1        -0.06315    0.03027  -2.086  0.03694 *
## Lag2         0.07588    0.03136   2.420  0.01553 *
## Lag3         0.00262    0.03144   0.083  0.93358
## Lag4        -0.02396    0.03023  -0.793  0.42807
## Lag5        -0.02942    0.03184  -0.924  0.35547
## Volume      -0.05148    0.04150  -1.241  0.21478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1122.4  on 816  degrees of freedom
## Residual deviance: 1108.2  on 810  degrees of freedom
## AIC: 1122.2
##
## Number of Fisher Scoring iterations: 4
```

Yes, Lag1, Lag2 and Intercept

(c) Compute the confusion matrix and overall fraction of correct predictions. Briefly explain what the confusion matrix is telling you.

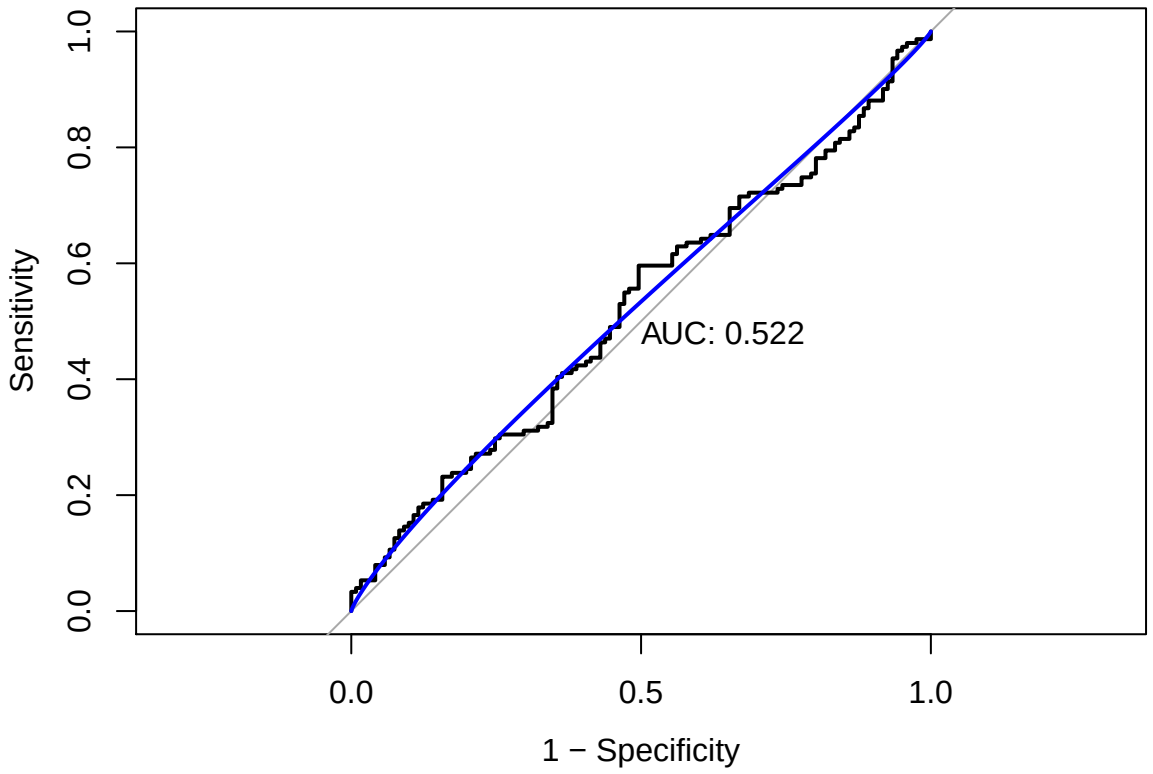
```
# Bayes classigier(cutoff 0.5)
test.pred.prob = predict(glm_fit, newdata = df[-rowTrain,],
                         type = "response")
test.pred = rep("Down", length(test.pred.prob))
test.pred[test.pred.prob>0.5] = "Up"

# confustionMatrix
confusionMatrix(data = as.factor(test.pred),
                reference = as.factor(df$Direction[-rowTrain]),
                positive = "Up")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down  Up
##       Down   16  28
##       Up    105 123
##
##                Accuracy : 0.511
##                  95% CI : (0.4499, 0.5719)
##     No Information Rate : 0.5551
##     P-Value [Acc > NIR] : 0.9361
##
##                   Kappa : -0.0568
##  Mcnemar's Test P-Value : 4.397e-11
##
##             Sensitivity : 0.8146
##             Specificity : 0.1322
##          Pos Pred Value : 0.5395
##          Neg Pred Value : 0.3636
##              Prevalence : 0.5551
##          Detection Rate : 0.4522
##    Detection Prevalence : 0.8382
##       Balanced Accuracy : 0.4734
##
##        'Positive' Class : Up
##
```

(d) Plot the ROC curve using the predicted probability from logistic regression and reportthe AUC.

```
roc_glm = roc(df$Direction[-rowTrain], test.pred.prob)
plot(roc_glm, legacy.axes = TRUE, print.auc = TRUE)
plot(smooth(roc_glm), col = 4, add = TRUE)
```



(e) Now fit the logistic regression model using a training data period from 1990 to 2008,with-Lag1andLag2as the predictors. Plot the ROC curve using the held out data(that is, the data from 2009 and 2010) and report the AUC.

(f) Repeat (e) using LDA and QDA.

(g) Repeat (e) using KNN. Briefly discuss your results.