# Homework 3

*Matthew Perrotta*

*April 10, 2019*

Load Libraries

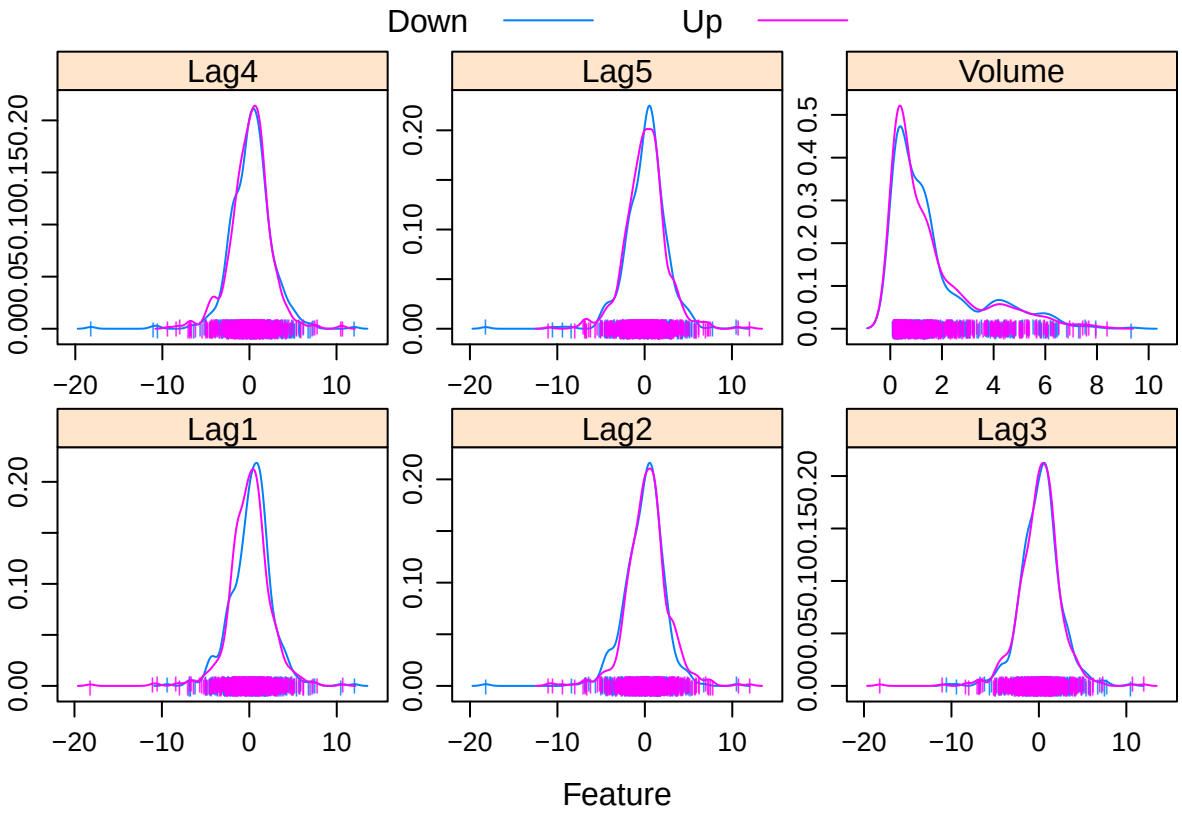
```
library(ISLR)
library(tidyverse)
library(caret)
library(corrplot)
library(pROC)
library(MASS)
```

```
data(Weekly)

x = model.matrix(Direction~., Weekly)[,3:8]

y = Weekly$Direction
```

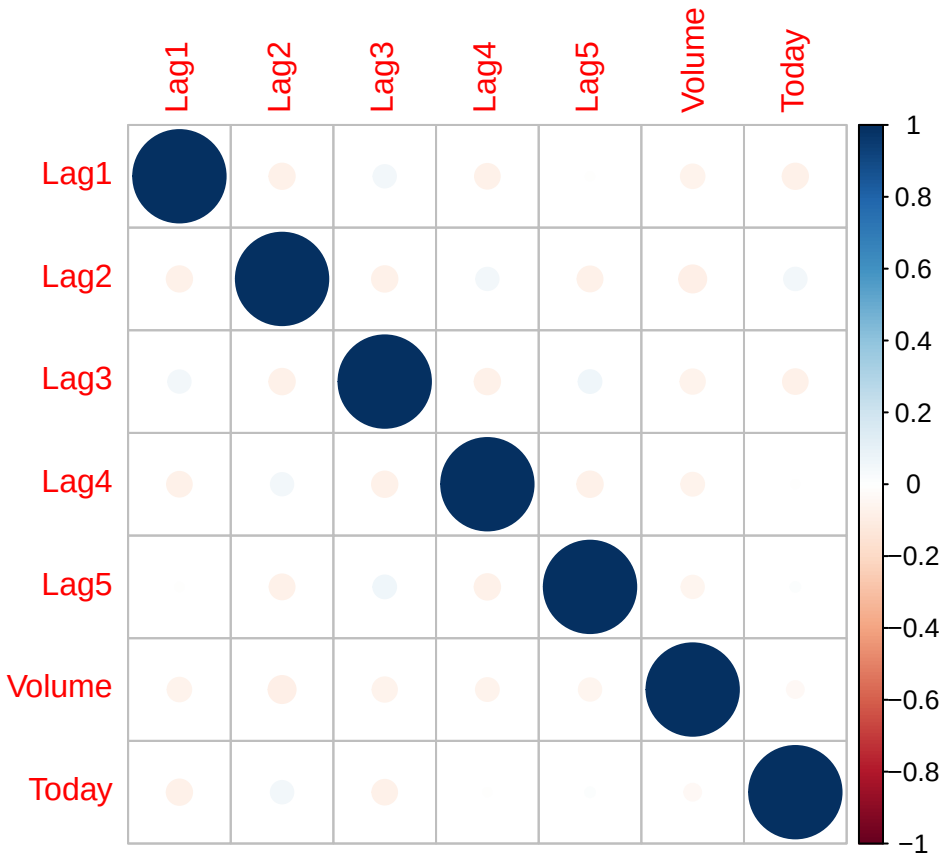
## Problem (a)

EDA

```
featurePlot(x,
            y,
            scales = list(x=list(relation="free"),
                          y=list(relation="free")),
            plot = "density", pch = "|",
            auto.key = list(columns = 2))
```




```
corrplot::corrplot(cor(Weekly[2:8]))
```




## Problem (b)

Logistic Regression

```
glm.fit <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
               data = Weekly,
               family = binomial)

summary(glm.fit)
```

```
##
## 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
```

```
ctrl <- trainControl(method = "repeatedcv",
                     repeats = 5,
                     summaryFunction = twoClassSummary,
                     classProbs = TRUE)
set.seed(1)
model.glm <- train(x,
                   y,
                   method = "glm",
                   metric = "ROC",
                   trControl = ctrl)

summary(model.glm)
```

```
##
## Call:
## NULL
##
## 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
```

# Problem (c)

Cofusion Matrix

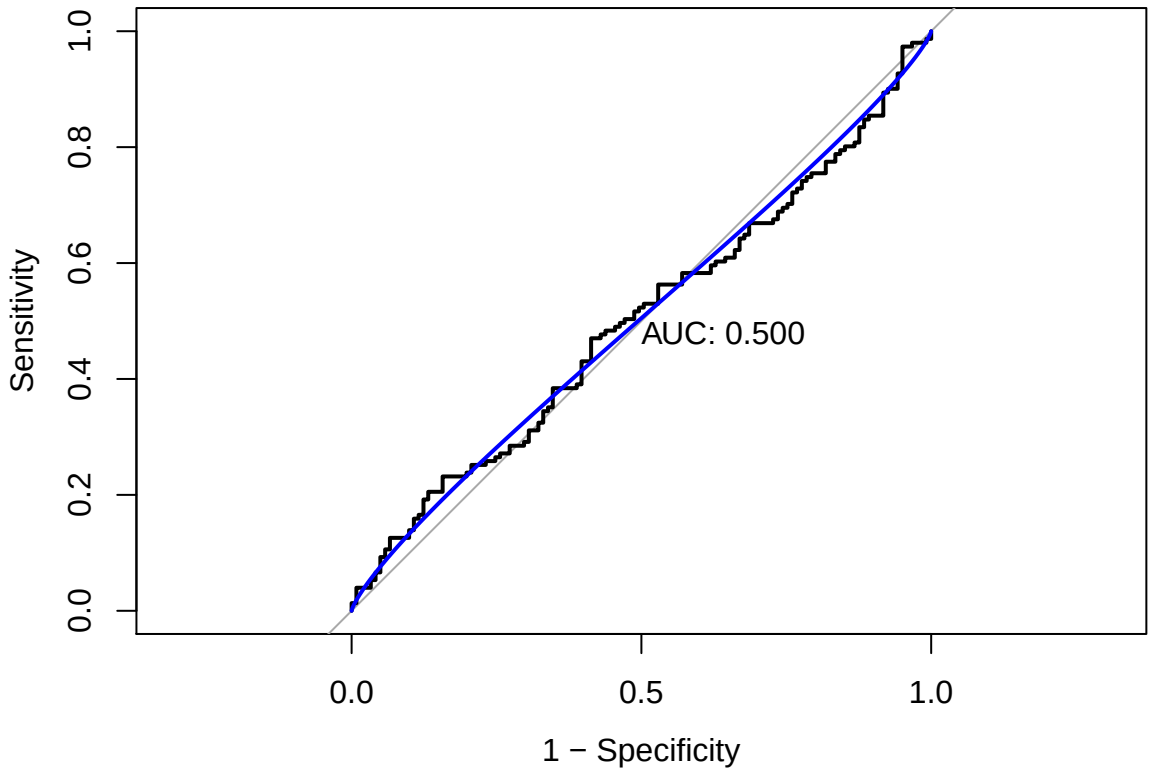
```
set.seed(1)
rowTrain <- createDataPartition(y,
                                p = 0.75,
                                list = FALSE)

test.pred.prob  <- predict(glm.fit, newdata = Weekly[-rowTrain,],
                           type = "response")
test.pred <- rep("Down", length(test.pred.prob))
test.pred[test.pred.prob > 0.5] <- "Up"

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

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down  Up
##       Down    7  16
##       Up    114 135
##
##                Accuracy : 0.5221
##                  95% CI : (0.4609, 0.5827)
##     No Information Rate : 0.5551
##     P-Value [Acc > NIR] : 0.8767
##
##                   Kappa : -0.0523
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.89404
##             Specificity : 0.05785
##          Pos Pred Value : 0.54217
##          Neg Pred Value : 0.30435
##              Prevalence : 0.55515
##          Detection Rate : 0.49632
##    Detection Prevalence : 0.91544
##       Balanced Accuracy : 0.47595
##
##        'Positive' Class : Up
##
```

# Problem (d)

ROC Curve

```r
roc.glm <- roc(y[-rowTrain], test.pred.prob)

plot(roc.glm, legacy.axes = TRUE, print.auc = TRUE)
plot(smooth(roc.glm), col = 4, add = TRUE)
```




## Problem (e)

Logistic Regression