Stock Market Prediction

Aymen Rumi

5/5/2020

Overview

We will try to build a classification model for Stock Maret data for 1089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010. Our goal is to predict the trend of the market as "Up" or "Down" given some predictors; returns for Lag 1 to Lag5 & Volume

Dataset

```r
names(Weekly)
```

```
[1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"
[7] "Volume"    "Today"     "Direction"
```

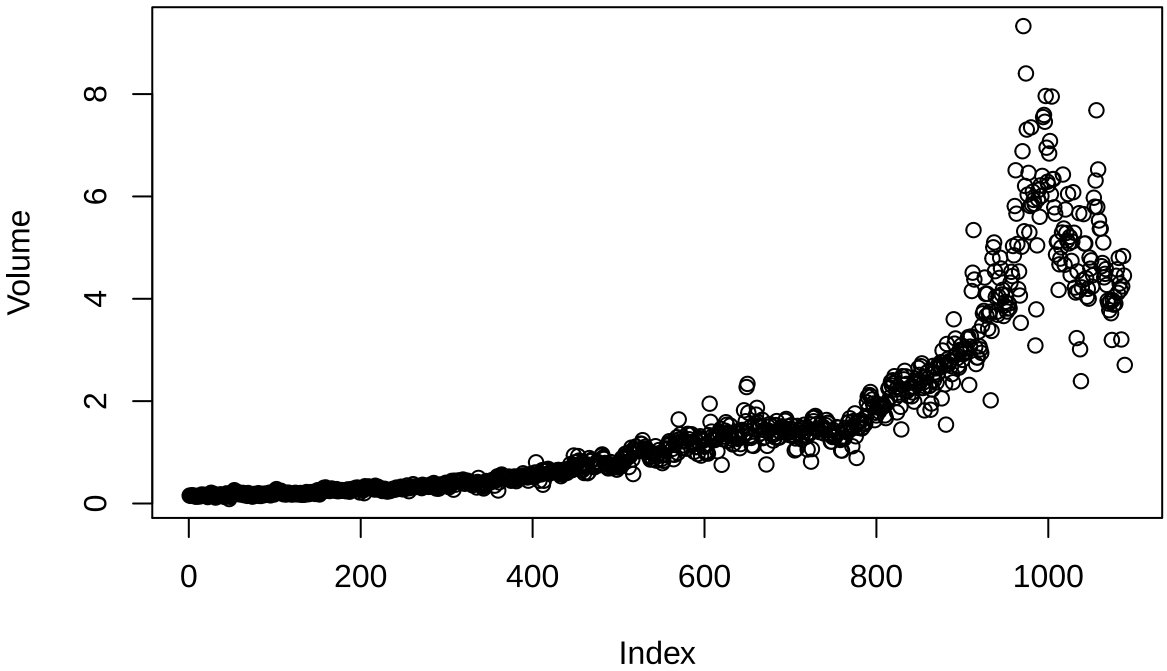
```r
head(Weekly)
```

```
  Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

```r
attach(Weekly)
cor(Weekly[1:8])
```

```
              Year         Lag1        Lag2        Lag3         Lag4
Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
               Lag5      Volume        Today
Year   -0.030519101  0.84194162 -0.032459894
Lag1   -0.008183096 -0.06495131 -0.075031842
Lag2   -0.072499482 -0.08551314  0.059166717
Lag3    0.060657175 -0.06928771 -0.071243639
Lag4   -0.075675027 -0.06107462 -0.007825873
Lag5    1.000000000 -0.05851741  0.011012698
Volume -0.058517414  1.00000000 -0.033077783
Today   0.011012698 -0.03307778  1.000000000
```

```r
plot(Volume)
```



Logistic Regression

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

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
```

```r
contrasts(Direction)
```

```
     Up
Down  0
Up    1
```

```r
glm.probabilities=predict(glm.fit,type="response")

glm.prediction=rep("Down",dim(Weekly)[1])

glm.prediction[glm.probabilities>0.5]="Up"

table(glm.prediction,Direction)
```

```
              Direction
glm.prediction Down  Up
          Down   54  48
          Up    430 557
```

```r
mean(glm.prediction==Direction)
```

```
[1] 0.5610652
```

```r
mean(glm.prediction!=Direction)
```

```
[1] 0.4389348
```

Training & Testing Model

```r
train=Year<2007

Data.predictors=Weekly[!train,]
Data.response=Direction[!train]


glm.fit=glm(Direction~Lag2,data=Weekly,family=binomial,subset=train)

summary(glm.fit)
```

```
Call:
glm(formula = Direction ~ Lag2, family = binomial, data = Weekly,
    subset = train)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.374  -1.277   1.036   1.081   1.261

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.23057    0.06818   3.382 0.000721 ***
Lag2         0.03837    0.03304   1.162 0.245435
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1207.6  on 879  degrees of freedom
Residual deviance: 1206.3  on 878  degrees of freedom
AIC: 1210.3

Number of Fisher Scoring iterations: 4
```

```r
glm.probabilities=predict(glm.fit,Data.predictors,type="response")

glm.prediction=rep("Down",length(Data.response))
glm.prediction[glm.probabilities>0.5]="Up"

table(glm.prediction,Data.response)
```

```
              Data.response
glm.prediction Down  Up
          Down    5   3
          Up     91 110
```

```r
mean(glm.prediction==Data.response)
```

```
[1] 0.5502392
```

```r
mean(glm.prediction!=Data.response)
```

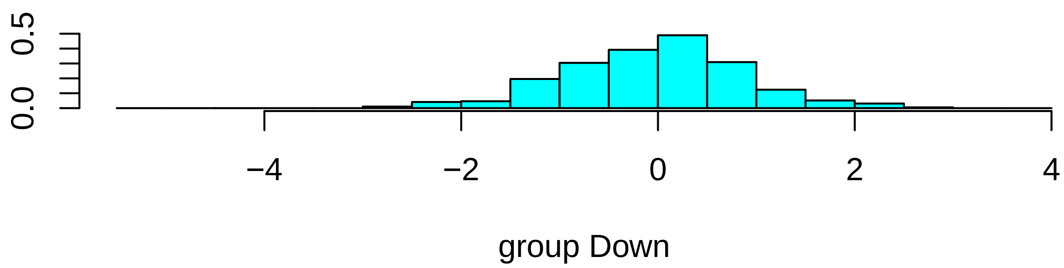
```
[1] 0.4497608
```

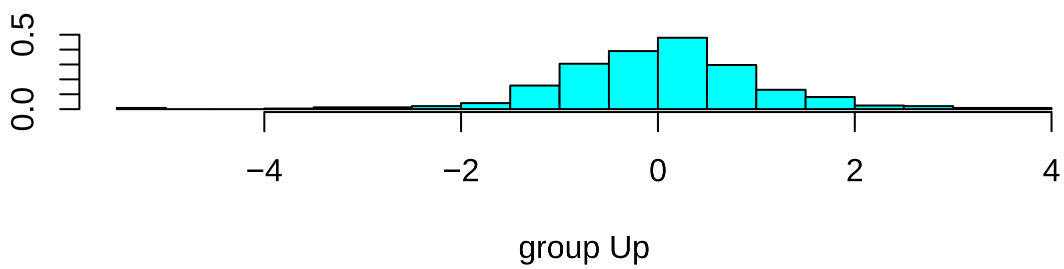
Linear Discriminant Analysis

```r
lda.fit=lda(Direction~Lag2,data=Weekly,subset=train)

lda.fit
```

```
Call:
lda(Direction ~ Lag2, data = Weekly, subset = train)

Prior probabilities of groups:
     Down        Up
0.4409091 0.5590909

Group means:
          Lag2
Down 0.0982732
Up   0.2610650

Coefficients of linear discriminants:
           LD1
Lag2 0.4849425
```

```r
plot(lda.fit)
```






```r
lda.prediction=predict(lda.fit, Data.predictors)
lda.class=lda.prediction$class

table(lda.class,Data.response)
```

```
         Data.response
lda.class Down  Up
     Down    5   3
     Up     91 110
```

```r
mean(lda.class==Data.response)
```

```
[1] 0.5502392
```

```r
mean(lda.class!=Data.response)
```

```
[1] 0.4497608
```

Quadratic Discriminant Analysis

```r
qda.fit=qda(Direction~Lag2,data=Weekly,subset=train)

qda.fit
```

```
Call:
qda(Direction ~ Lag2, data = Weekly, subset = train)

Prior probabilities of groups:
     Down        Up 
0.4409091 0.5590909 

Group means:
          Lag2
Down 0.0982732
Up   0.2610650
```

```r
qda.prediction=predict(qda.fit, Data.predictors)
qda.class=lda.prediction$class

table(qda.class,Data.response)
```

```
         Data.response
qda.class Down  Up
     Down    5   3
     Up     91 110
```

```r
mean(qda.class==Data.response)
```

```
[1] 0.5502392
```

```r
mean(qda.class!=Data.response)
```

```
[1] 0.4497608
```

K-Nearest Neighbors

```r
train.X=cbind(Lag1,Lag2)[train,]


test.X=cbind(Lag1,Lag2)[!train,]


train.Direction=Direction[train]


set.seed(1)
knn.pred=knn(train.X,test.X,train.Direction,k=1)
table(knn.pred,Data.response)
```

```
        Data.response
knn.pred Down Up
    Down   52 50
    Up     44 63
```

```r
mean(knn.pred==Data.response)
```

```
[1] 0.5502392
```

```r
knn.pred=knn(train.X,test.X,train.Direction,k=3)
table(knn.pred,Data.response)
```

```
        Data.response
knn.pred Down Up
    Down   53 50
    Up     43 63
```

```r
mean(knn.pred==Data.response)
```

```
[1] 0.5550239
```

Conclusions

We have compared the performance on test data for 2008 to 2010 on different classifiers trained from 1990 to 2007. The classification alrotighms used were Logistic Regression, Linear Discriminant Analysis, Quadratic Distriminant Analysis & K-Nearest Neighbors and they all perform the same, having a correct classification score of 55% which is only slightly better than random guessing thus we can conclude that this dataset is not a great predictor to beat the market