MBA 6693 Business Analytics

Assignment 2: Classification Models

Name: Vinayak B. Menon

18/07/2020

Objective:

This report aims to model the relationship between direction of the $SP500$ index and 3 variables: the percentage returns from the previous week $Lag1$, the volume of shares traded $Volume$ and the percentage return of this week $Today$. We compare the effectiveness of logistic regression and LDA and arrive at the best representative model.

Data Exploration:

The $Weekly$ data, from the $ISLR$ package, is the dataset under consideration and has weekly observations from 1990 to 2010. We restrict our predictor variables to $Lag1$, $Today$ and $Volume$ and we shall form a model that predicts $Direction$. We summarize the key values of the dataset below.

```
head(week_ret)
```

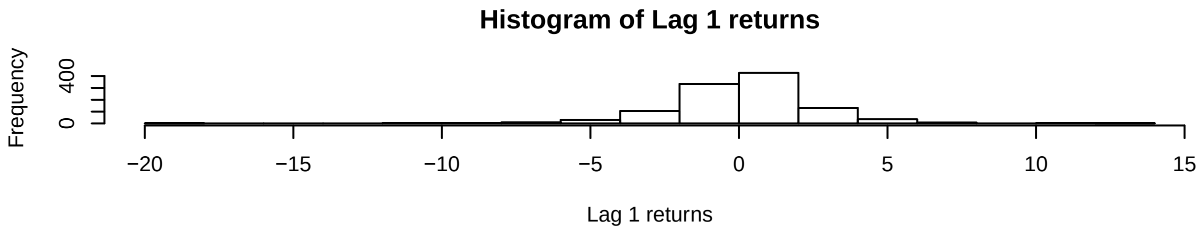
```
##      Lag1    Volume  Today Direction
## 1:  0.816 0.1549760 -0.270      Down
## 2: -0.270 0.1485740 -2.576      Down
## 3: -2.576 0.1598375  3.514        Up
## 4:  3.514 0.1616300  0.712        Up
## 5:  0.712 0.1537280  1.178        Up
## 6:  1.178 0.1544440 -1.372      Down
```

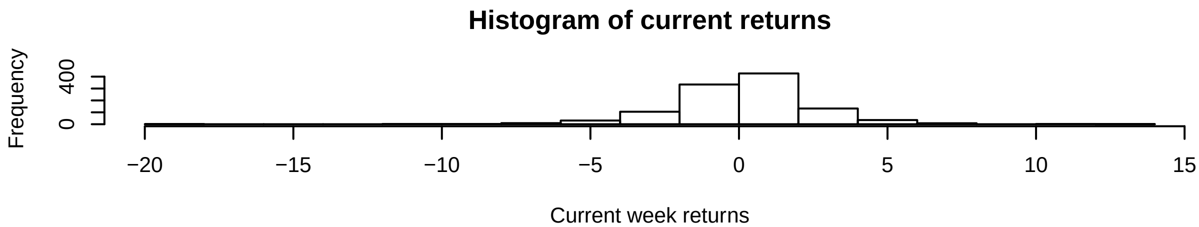
```
summary(week_ret)
```

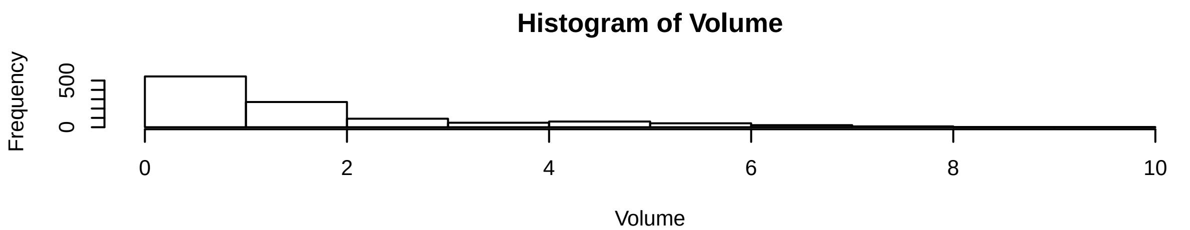
```
##       Lag1              Volume            Today          Direction
##  Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950   Down:484
##  1st Qu.: -1.1540   1st Qu.:0.33202   1st Qu.: -1.1540   Up  :605
##  Median :  0.2410   Median :1.00268   Median :  0.2410
##  Mean   :  0.1506   Mean   :1.57462   Mean   :  0.1499
##  3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050
##  Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260
```

We now plot the histograms for each of the numeric independent factors to see if there are any widely deviating values:

```
#histogram output
par(mfrow=c(3,1))
hist(week_ret$Lag1,xlab = "Lag 1 returns",main="Histogram of Lag 1 returns")
hist(week_ret$Today,xlab = "Current week returns",main="Histogram of current returns")
hist(week_ret$Volume,xlab = "Volume",main="Histogram of Volume")
```

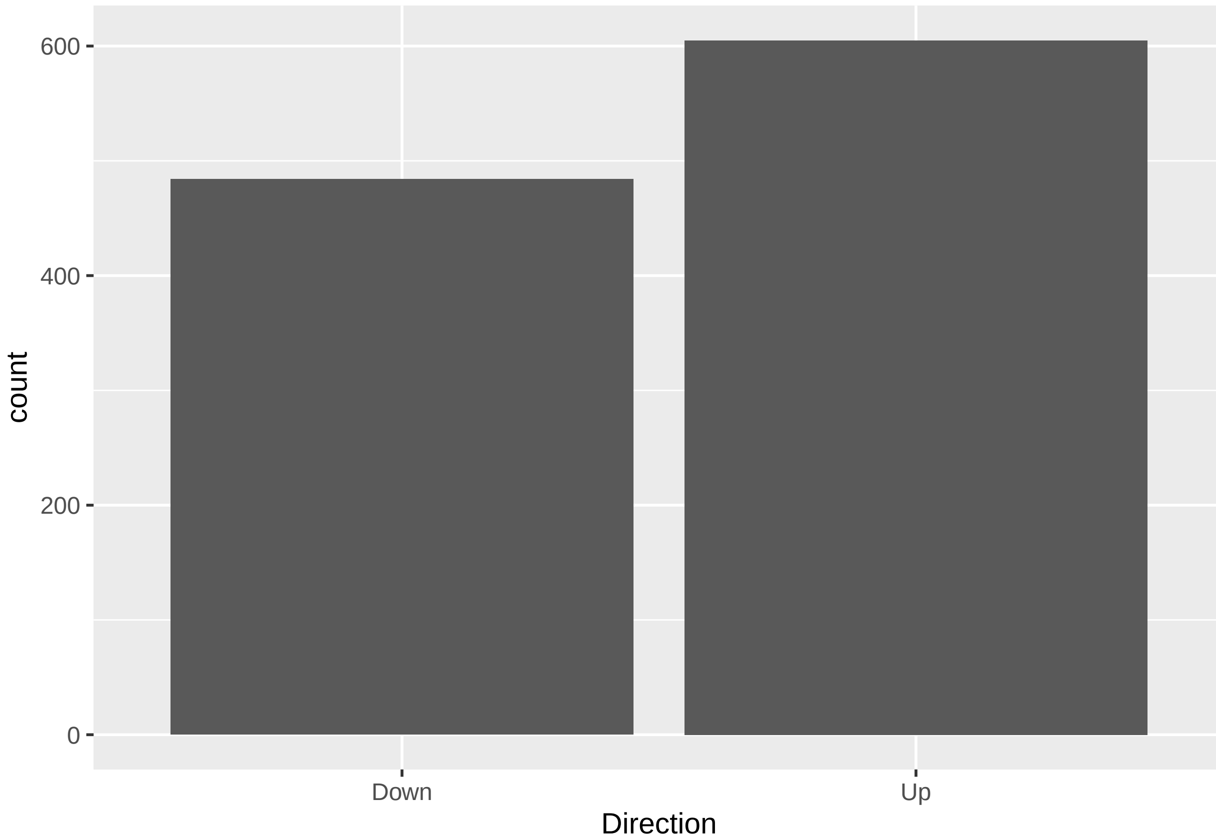






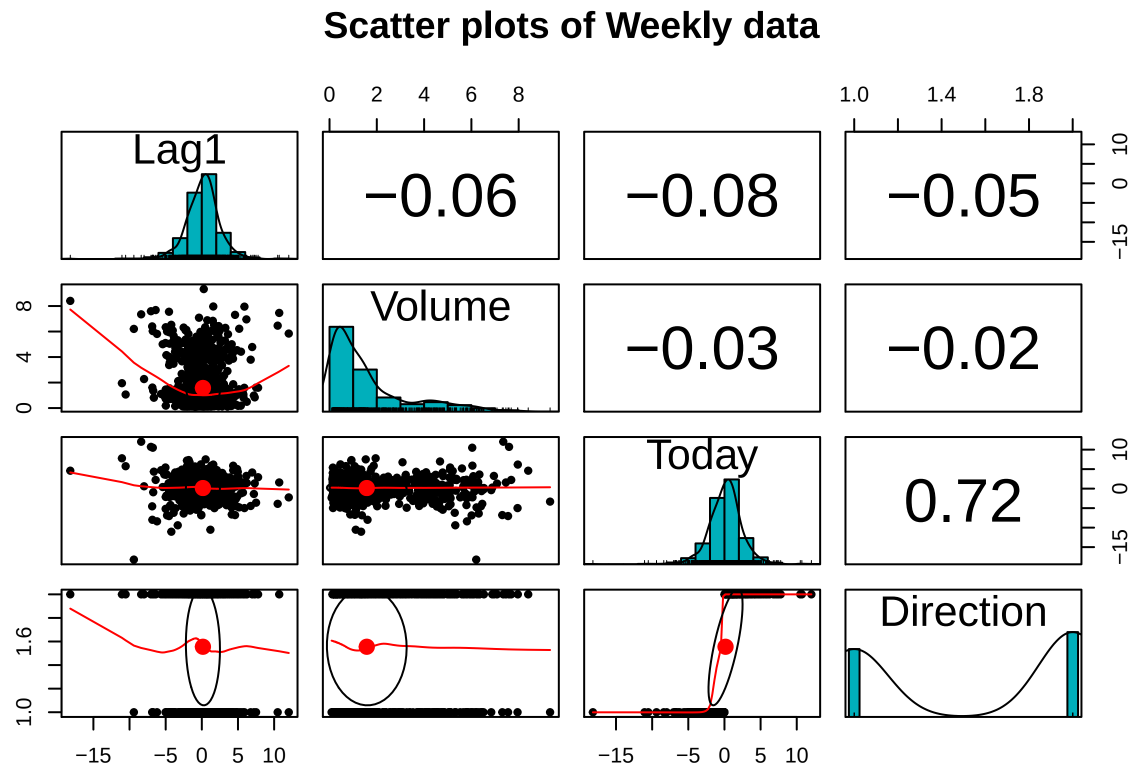
Note the decreasing trend in the Volume histogram. This is most likely due to the high volume of trades that would have occured specifically in crisis situations. In normal markets, the trades lie on the lower end. Now we plot the bar graph of the market direction.

```
#barplot
ggplot(data = week_ret) +
  geom_bar(mapping = aes(x = Direction))
```



The *Up* direction seems to be more in frequency as compared to *Down* by atleast a 100. We further check the bivariate analysis of the variables under consideration.

```
#paired plots
pairs.panels(week_ret,
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             main="Scatter plots of Weekly data",)
```



Notice that there is a strong relationship between the current weekly returns and the market direction. This is obvious since the market movement is essentially a measure of the current return with respect to the previous return.

We first begin with the logistic regression model.

```
#Model 1,
logistic_fit <- glm(Direction ~Lag1+Volume+Today, data = week_ret_train,family = binomial,maxit=1)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
summary(logistic_fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Volume + Today, family = binomial,
##     data = week_ret_train, maxit = 1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2550  -0.8011   0.3330   0.7959   1.1244
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.105491   0.111065   0.950    0.342
```

```
## Lag1        -0.006294   0.036995  -0.170    0.865
## Volume       0.034166   0.050851   0.672    0.502
## Today        0.745867   0.034672  21.512   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1123.36  on 815  degrees of freedom
## Residual deviance:  538.72  on 812  degrees of freedom
## AIC: 546.72
##
## Number of Fisher Scoring iterations: 1
```

Including all 3 variables, we see that there is a strong positive relationship between the current week returns and the market movement. Based on the p-values, the remaining two variables does not significantly affect the direction.
We now move on the LDA model with the same variables.

```r
#Model 4
lda_fit <- lda(Direction ~Volume+Lag1+Today, data = week_ret_train)
lda_fit
```

```
## Call:
## lda(Direction ~ Volume + Lag1 + Today, data = week_ret_train)
##
## Prior probabilities of groups:
##      Down        Up
## 0.4509804 0.5490196
##
## Group means:
##        Volume      Lag1     Today
## Down 1.511636 0.2281495 -1.739644
## Up   1.444983 0.1957522  1.633788
##
## Coefficients of linear discriminants:
##                 LD1
## Volume  0.028215088
## Lag1   -0.005197651
## Today   0.615956032
```

Now we proceed to calculating the error rates between the LDA and all logistic models under consideration. We first create the confusion matrix for the logistic regression model:

```r
#Creating dataframe for out of sample error rate
out_sample_err <- data.frame(matrix(0,nrow = 1,ncol = 2),row.names = c("Error Rate"))
colnames(out_sample_err) <- c("Logistic","LDA")
#calculating the probabilities associated for each category
logistic_prob <- round(predict(logistic_fit, week_ret_test, type = "response"))
#assigning categories over the responses
Model_1 <- rep("Down",nrow(week_ret_test))
Model_1[logistic_prob > 0.5] <- "Up"
#confusion matrix
table(Model_1, week_ret_test$Direction)
```

```
##
## Model_1 Down  Up
##    Down  105   0
##    Up     11 157
```

Notice that there are 13 errors noted here. We obtain the confusion matrix of the LDA model predictions:

```r
#Predicting the values from LDA
lda_pred <- predict(lda_fit, week_ret_test)
#Confusion matrix
table(lda_pred$class, week_ret_test$Direction)
```

```
##
##         Down  Up
##   Down  105   0
##   Up     11 157
```

The matrix looks the same as the logistic model. Now we calculate the out of sample error rate which would likely be the same.

```r
#Out of Sample error
out_sample_err$Logistic <- mean(Model_1!=week_ret_test$Direction)
out_sample_err$LDA <- mean(lda_pred$class!=week_ret_test$Direction)
out_sample_err
```

```
##              Logistic        LDA
## Error Rate 0.04029304 0.04029304
```

We see that both the logistic and LDA models manage to perform equally well in representing the market direction based on the three considered variables.