**NAME**

**COLLEGE NUMBER**

**Abstract**

Statisticians and mathematicians together with data scientists require data in every day life to support industry. Whereas it’s true that data is the gold of the future, a lot of research and attention is currently ongoing in into the industry to help support the managerial and top leadership of these industries in making better and less risky decisions on maintaining the business. Models and algorithms have been built to help banks and financial institutions forecast sales and predict the next direction the sales industry will take, by understanding how previously the company has performed and then putting the same to determine how likely the company is to perform in the near future. Of the several models and functions that exist, the linear logistic regression model, the LDA and the QDA models have been widely been used as part of the processes in determining the next values in the models. In this study, the researchers’ aims to use the three models outlined above and try to predict the Smarket dataset as a comparison is made between the three choices and how they are highly applicable in this scenario.

**Introduction**

In this project, the research is going to apply models into the Smarket dataset that contains information on the various trading performance of the various durations. The models shall monitor the S&P 500 that is considered by the United States Stock exchange as the yardstick for measuring the performance of the various company market shares and portfolios based on their trading index. The dataset has been measured across four years since 2001 to 2005 giving a total of 1250 days of dataset. As the companies continued to trade across the years, their returns on the various trading’s were recorded and given out as Lags ranging from 1 to 5, the volume of these transactions in billions was also recorded and given out on scale, along with “Today”, which is confirmed as the returns in the expected day and “Direction” which helped to indicate whether the market was up or down on that particular date and day.

**Problem statement**

The purpose of this project is to come up with scientific and statistical models and algorithms to solve the “Smarket” case scenario by monitoring past behaviour and predicting future performance and values. This shall be done based on the column values available on the dataset. The below models have been proposed for usage during the project:

* Logistic regression
* LDA
* QDA
* KNN

**Methods**

The methods of probability sampling applied on this dataset includes:

**Stratified random sampling**

The raw values of these datasets composed of the S&P 500 of the top performing companies according to YAHOO finance dataset and their values were converted to percentages and lagged as Lag 1 Lag2 Lag3 Lag4 and Lag5 representing the past consecutive 5 days of working.

A quick load of the dataset revealed the dataset as follows:

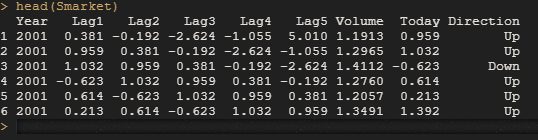
> data(Smarket)

> names(Smarket)

[1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"

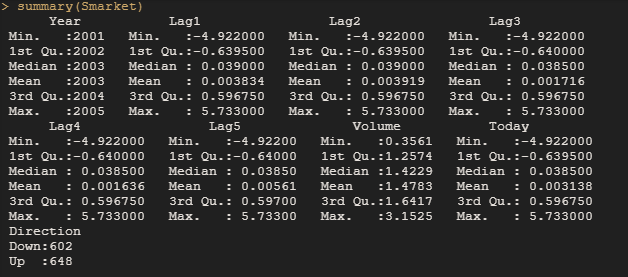
[7] "Volume"    "Today"     "Direction"

> head(Smarket)



A quick snip into the dataset reveals that for instance on row 1, in the year 2001 across the first 5 consecutive days that cut across Lag1, Lag2, Lag3, Lag4 and Lag5 indicate that the market was down 3 times and up twice closing the day with a positive in 0.96 Billions making the status an upward trend.

A quick summary of the dataset revealed the following:



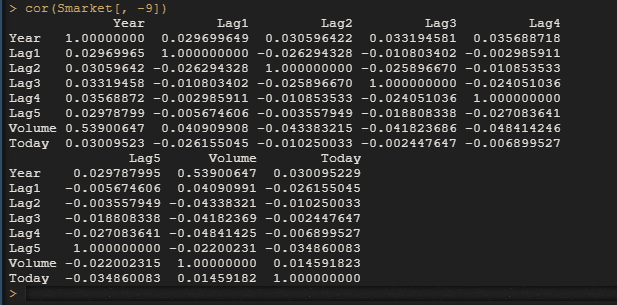
And we take this further by establishing the existing correlation in the dataset;

#GETTING THE CORRELATIONSHIP IN THE DATASET;

> `?`(Smarket)

> attach(Smarket)

> cor(Smarket[, -9])



From the dataset correlation above, we can see that there is minimal to no correlation at all between the different lags and the todays returns, this is observed by noting that the values are somewhat far away from 0 with some values going way closed to -1. On the other hand though, it is clearly visible that there is somewhat a positive correlation between the Years and the volumes of returns in US billion dollars. We can affirm this by illustrating that the as the number of years increase so do the vogues of transactions alongside, this means that there was significant increase in the volumes of returns across the years of market trade since 2001 to 2005.

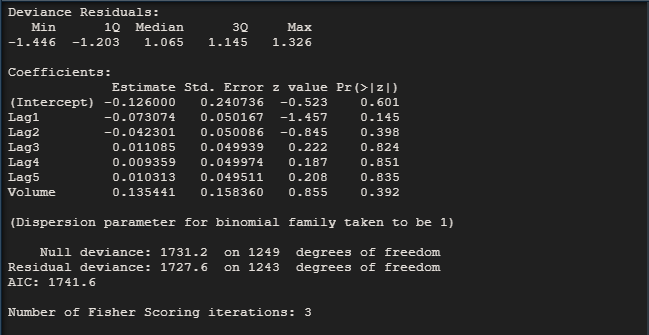
**Logistic regression model**

We predict the future direction of the stock market taking entries from lag 1 through to lag 5. We use R inbuilt function to fit the linear model for based on our logistic regression above.

#UNDERATAKNG LOGISTIC REGRESSION

> glm.fit = glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = Smarket, family = binomial)

>



And then using the predict function. Based on our training dataset initially provided, we sample out the first 10 records with in initial perception that the market will go up;

# ANALYSING THE TARINING SET

> glm.probs = predict(glm.fit, type = "response")

> length(glm.probs)

[1] 1250

> glm.probs[1:10]

        1         2         3         4         5         6         7

0.5070841 0.4814679 0.4811388 0.5152224 0.5107812 0.5069565 0.4926509

        8         9        10

0.5092292 0.5176135 0.4888378

> contrasts(Direction)

     Up

Down  0

Up    1

#CLACULATE THE CORRECT PREDICTION WAS RIGHT

> glm.pred = rep("Down", 1250)

> glm.pred[glm.probs > 0.5] = "Up"

> # Produce a confusion matrix

> table(glm.pred, Direction)

        Direction

glm.pred Down  Up

    Down  145 141

    Up    457 507

> mean(glm.pred == Direction)

[1] 0.5216

**Splitting the dataset into training and testing samples;**

We shall emulate a vector that takes the period from 2001 to 2004 from the original dataset

glm.probs <- predict(glm.fit, Smarket.2005, type = "response")

glm.pred = rep("Down", nrow(Smarket.2005))

glm.pred[glm.probs > 0.5] = "Up"

table(glm.pred, Direction.2005)

mean(glm.pred == Direction.2005)

mean(glm.pred != Direction.2005)

#REFFITITNG THE MODEL

glm.fit = glm(Direction ~ Lag1 + Lag2, data = Smarket, family = binomial, subset = train)

glm.probs = predict(glm.fit, Smarket.2005, type = "response")

glm.pred = rep("Down", nrow(Smarket.2005))

glm.pred[glm.probs > 0.5] = "Up"

table(glm.pred, Direction.2005)

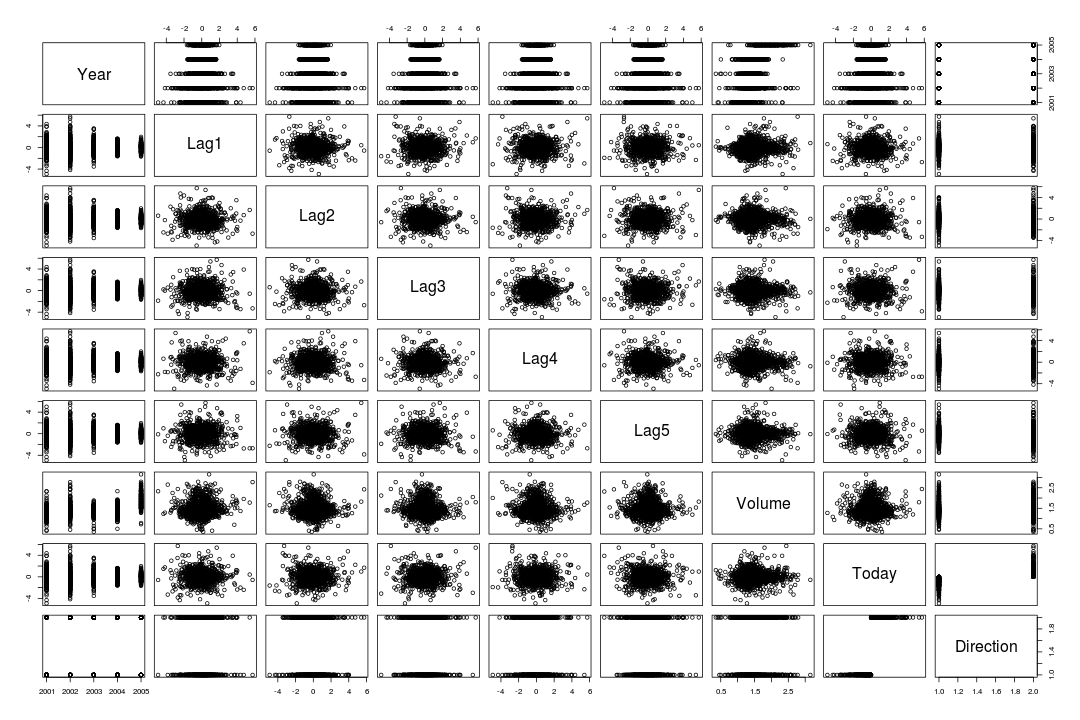
#THEN WE PREDICT

predict(glm.fit, newdata = data.frame(Lag1 = c(1.2, 1.5), Lag2 = c(1.1, -0.8)),

type = "response")

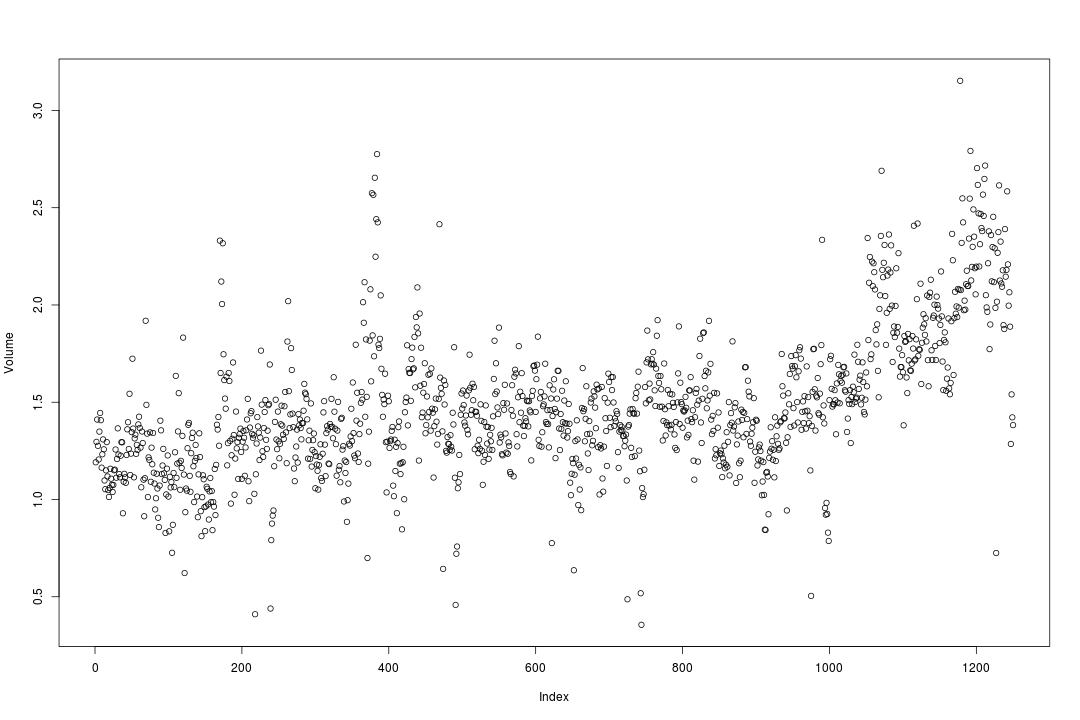
#PLOT THE SCATTER MATRIX

pairs(Smarket)



#WE PLOT THE VOLUME GRAPH

plot(Volume)



As earlier discussed in our logistic regression model, it is again observed that volumes of retunes in Billions tend to increase with significant increase in duration of time.

**LDA MODEL**

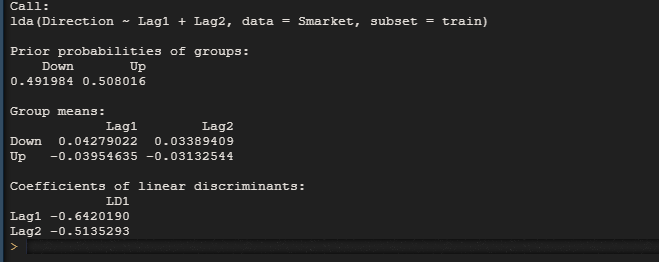
We start by fitting the model to our training dataset.

#FIT AND TRAIN DATA

> library(MASS)

> lda.fit = lda(Direction ~ Lag1 + Lag2, data = Smarket, subset = train)

> lda.fit



#APPLYING THE PREDICT FUNCTIONS ON THE TEST

> Smarket.2005 = Smarket[!train, ]

> lda.pred = predict(lda.fit, Smarket.2005)

> names(lda.pred)

[1] "class"     "posterior" "x"

> lda.class = lda.pred$class

> Direction.2005 = Direction[!train]

> table(lda.class, Direction.2005)

         Direction.2005

lda.class Down  Up

     Down   35  35

     Up     76 106

> mean(lda.class == Direction.2005)

[1] 0.5595238

> sum(lda.pred$posterior[, 1] >= 0.5)

[1] 70

> sum(lda.pred$posterior[, 1] < 0.5

+ )

[1] 182

#PLOT THE MODEL DATA

> lda.class[1:20]

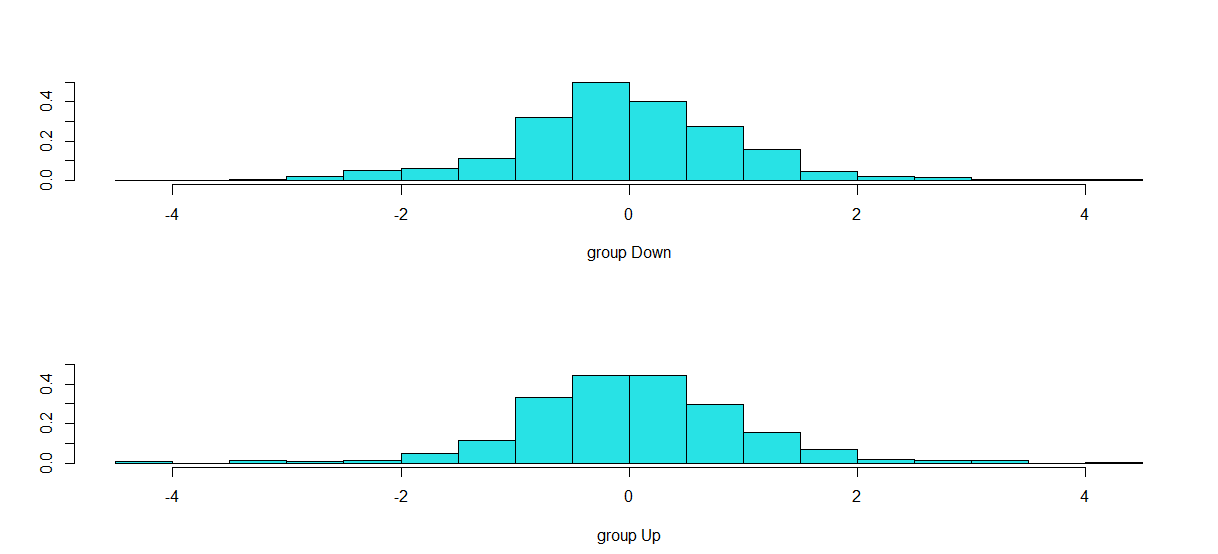
 [1] Up   Up   Up   Up   Up   Up   Up   Up   Up   Up   Up   Down Up   Up

[15] Up   Up   Up   Down Up   Up

Levels: Down Up

> plot(lda.fit)

>



**QDA MODEL**

We will now apply quadratic discriminant analysis on the dataset. In order to so this, we call the qda()

Function which is written inside R s MASS packages follows:

> library(MASS)

> library(dplyr)

> library(ISLR)

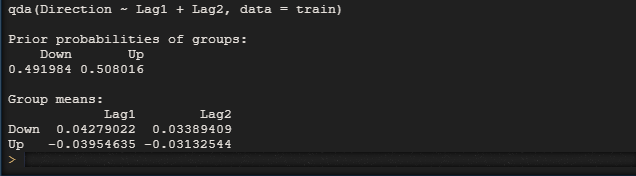
> select <- dplyr::select

> train = Smarket %>%

+ filter(Year < 2005)

> model\_QDA = qda(Direction~Lag1+Lag2, data = train)

> model\_QDA



We apply the predict function on the coefficient of the linear discriminants. The predict function will be applied just as same as the one for LDA as below since we had separated our data into train and test, we apply the function on the test data and determine the direction of the prediction

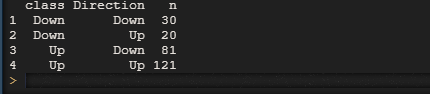
> predictions\_QDA = data.frame(predict(model\_QDA, test))

> test = Smarket %>% filter(Year >= 2005)

> predictions\_QDA = data.frame(predict(model\_QDA, test))

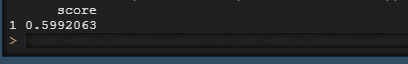
> predictions\_QDA = cbind(test, predictions\_QDA)

> predictions\_QDA %>% count(class, Direction)



Finally predict the direction the stock market will take

> predictions\_QDA %>% summarize(score = mean(class == Direction))



As observed from the model, the QDA gave an prediction of close to 60% accuracy on the dataset of the model.

**Observation**

Both the LDA and the Logistic regression models applied above tend to perform well when the decision boubndreies are linear, on the other hand the QDA gave a nonlinear with a quadratic decision boundary.

**REFERENCES**

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in R, https://www.statlearning.com, Springer-Verlag, New York

Chenail, R. J. (2012). Conducting qualitative data analysis: Qualitative data analysis as a metaphoric process. *Qualitative Report*, *17*(1), 248-253.

Juárez‐Barrientos, J. M., de Montserrat Tejeda‐Paz, M., de Jesús Ramírez‐Rivera, E., Aguirre‐Cruz, A., Rodríguez‐Miranda, J., Martínez‐Sánchez, C. E., & Herman‐Lara, E. (2019). Use of quantitative descriptive analysis (QDA) coupled with multivariate statistical methods to detection and discrimination of adulterated fresh cheeses. *Journal of Sensory Studies*, *34*(1), e12479.

Lee, L. C., Liong, C. Y., & Jemain, A. A. (2017, May). Q-mode versus R-mode principal component analysis for linear discriminant analysis (LDA). In *AIP Conference Proceedings* (Vol. 1842, No. 1, p. 030024). AIP Publishing LLC.

Manning, C. (2007). Logistic regression (with R). *Ver http://nlp. stanford. edu/mannlng/courses/ling28*, *9*.

Lewis, R. B., & Maas, S. M. (2007). QDA Miner 2.0: Mixed-model qualitative data analysis software. *Field methods*, *19*(1), 87-108.