Choose You Own Project

David Forrester

23/12/2019

# Overview

This project will look into the total stock price data for apple from listing on the stock exchange in 1985 to 2018. 4 different methods will be used to attempt to predict and forecast the close price each day. The data will be split into a training and test set for training and validating the methods.

A root mean square error calculation will be used to assess the performance of the predictions with a goal of returning the smallest value.

## Dataset

The dataset is loaded into r as a comma separated value type from the git repo. This data is then split into a training and test set of 95% to 5%.

library(tidyverse)  
library(lubridate)  
library(stringr)  
library(rvest)  
library(XML)  
library(tidytext)  
library(wordcloud)  
library(dslabs)  
library(caret)  
library(pracma)  
library(MASS)  
library(tseries)  
library(forecast)  
library(tsfknn)  
  
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")  
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")  
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")  
if(!require(tsfknn)) install.packages("tsfknn")  
  
# relative path to file  
file <- ("aapl.us.txt")  
  
# Read csv file into environment  
aapl <- read\_csv(file)  
  
# Split the data set by date ranges 80% train 20% test  
aapl\_train <- aapl[1:round(nrow(aapl) \* 0.95),]  
aapl\_test <- na.omit(aapl[round(nrow(aapl) \* 0.95) + 1:nrow(aapl),])

# Methods and Analysis

## Data Analysis

The first 6 rows and the summary below of the training dataset give insight into the structure of the data. It can be seen that each row represents a day of trading for the apple stock with the open, close, volume and intraday prices.

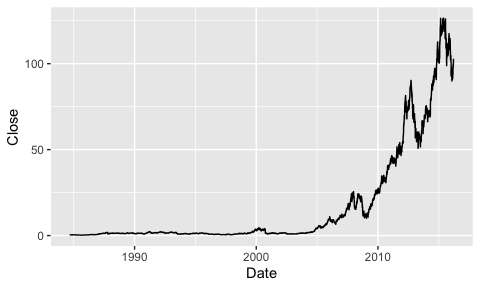
## # A tibble: 6 x 7  
## Date Open High Low Close Volume OpenInt  
## <date> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1984-09-07 0.424 0.429 0.419 0.424 23220030 0  
## 2 1984-09-10 0.424 0.425 0.414 0.421 18022532 0  
## 3 1984-09-11 0.425 0.437 0.425 0.429 42498199 0  
## 4 1984-09-12 0.429 0.432 0.416 0.416 37125801 0  
## 5 1984-09-13 0.439 0.441 0.439 0.439 57822062 0  
## 6 1984-09-14 0.441 0.456 0.441 0.446 68847968 0

## Date Open High   
## Min. :1984-09-07 Min. : 0.2331 Min. : 0.2356   
## 1st Qu.:1992-07-16 1st Qu.: 1.1128 1st Qu.: 1.1371   
## Median :2000-05-28 Median : 1.5355 Median : 1.5660   
## Mean :2000-06-05 Mean : 16.8422 Mean : 17.0179   
## 3rd Qu.:2008-04-27 3rd Qu.: 16.6885 3rd Qu.: 17.0115   
## Max. :2016-03-17 Max. :127.7800 Max. :127.8600   
## Low Close Volume OpenInt   
## Min. : 0.2305 Min. : 0.2305 Min. :0.000e+00 Min. :0   
## 1st Qu.: 1.0886 1st Qu.: 1.1153 1st Qu.:4.743e+07 1st Qu.:0   
## Median : 1.5015 Median : 1.5369 Median :7.834e+07 Median :0   
## Mean : 16.6435 Mean : 16.8330 Mean :1.106e+08 Mean :0   
## 3rd Qu.: 16.3358 3rd Qu.: 16.6465 3rd Qu.:1.373e+08 3rd Qu.:0   
## Max. :125.4200 Max. :126.4900 Max. :2.070e+09 Max. :0

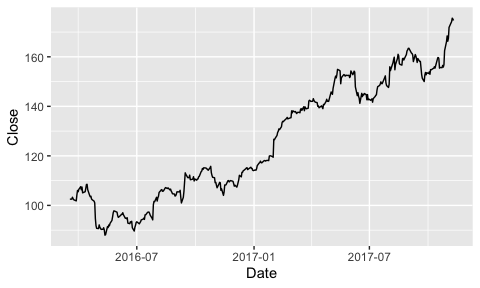
## Modelling Approach

### Model 1: Previous Close

aapl\_train %>%  
 ggplot() +  
 geom\_line(aes(Date, Close))

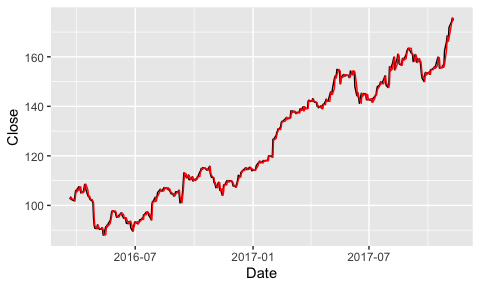


aapl\_test %>%  
 ggplot() +  
 geom\_line(aes(Date, Close))



aapl\_test <- mutate(aapl\_test, PrevClose = lag(Close))  
aapl\_test <- na.omit(aapl\_test)

aapl\_test %>%  
 ggplot(aes(x = Date)) +  
 geom\_line(aes(y = Close), colour = "black") +  
 geom\_line(aes(y = PrevClose), colour = "red")



naive\_rmse <- RMSE(aapl\_test$Close, aapl\_test$PrevClose)  
  
rmse\_results <- data\_frame(method = "Previous Value Prediction", RMSE = naive\_rmse)

## Warning: `data\_frame()` is deprecated, use `tibble()`.  
## This warning is displayed once per session.

rmse\_results %>% knitr::kable()

|  |  |
| --- | --- |
| method | RMSE |
| Previous Value Prediction | 1.51005 |

rmse\_results

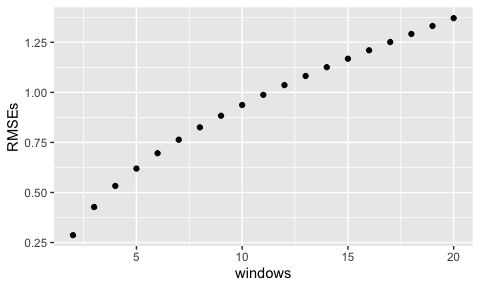
## # A tibble: 1 x 2  
## method RMSE  
## <chr> <dbl>  
## 1 Previous Value Prediction 1.51

### Model 2: Moving Average

windows <- seq(2, 20, 1)

RMSEs <- sapply(windows, function(w){  
   
 aapl\_train <- mutate(aapl\_train, MovingAvg = movavg(aapl\_train$Close, w, type=c("s")))  
   
 return(RMSE(aapl\_train$Close, aapl\_train$MovingAvg))  
})

qplot(windows, RMSEs)

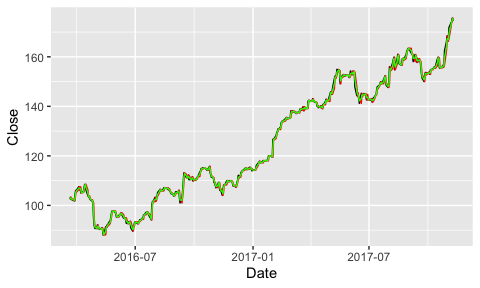


best\_window <- windows[which.min(RMSEs)]  
best\_window

## [1] 2

aapl\_test <- mutate(aapl\_test, MovingAvg = movavg(aapl\_test$Close, best\_window, type=c("s")))

aapl\_test %>%  
 ggplot(aes(x = Date)) +  
 geom\_line(aes(y = Close), colour = "black") +  
 geom\_line(aes(y = PrevClose), colour = "red") +  
 geom\_line(aes(y = MovingAvg), colour = "green")



naive\_rmse <- RMSE(aapl\_test$Close, aapl\_test$MovingAvg)  
rmse\_results <- bind\_rows(rmse\_results,  
 data\_frame(method="Moving Average Prediction",   
 RMSE = naive\_rmse))  
rmse\_results

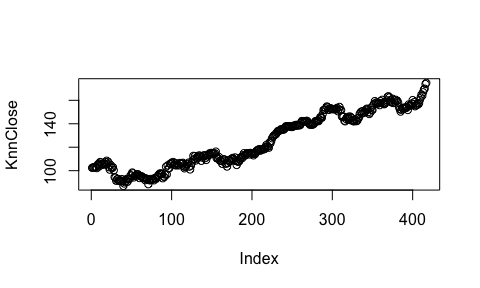
## # A tibble: 2 x 2  
## method RMSE  
## <chr> <dbl>  
## 1 Previous Value Prediction 1.51   
## 2 Moving Average Prediction 0.755

### Model 3: KNN Forecasting

KnnClose <- aapl[7946,]$Close

for (n in seq(1, 416, 1)) {  
 PredKnn <- knn\_forecasting(aapl[1:7946 + n,]$Close, h = 1, lags = 1:2, k = 2)  
 KnnClose <- append(KnnClose, PredKnn$prediction)  
}

plot(KnnClose)



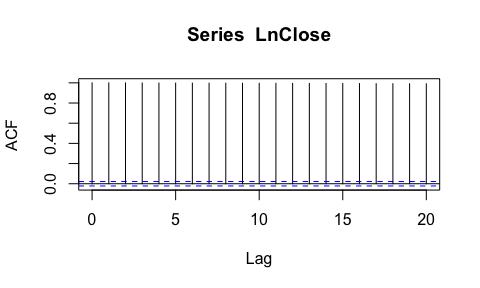
df <- data.frame(aapl\_test$Close, KnnClose)  
naive\_rmse <- RMSE(df$aapl\_test.Close, df$KnnClose)  
rmse\_results <- bind\_rows(rmse\_results,  
 data\_frame(method="KNN Prediction",   
 RMSE = naive\_rmse))  
rmse\_results

## # A tibble: 3 x 2  
## method RMSE  
## <chr> <dbl>  
## 1 Previous Value Prediction 1.51   
## 2 Moving Average Prediction 0.755  
## 3 KNN Prediction 2.59

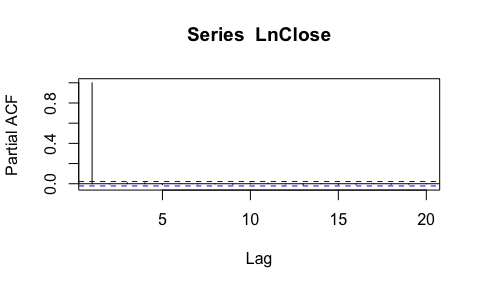
### Model 4: ARIMA

LnClose = log(aapl\_train$Close)

acf(LnClose, lag.max = 20)



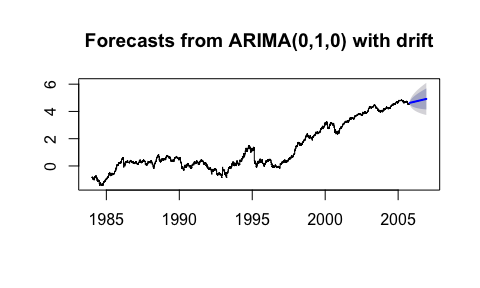
pacf(LnClose, lag.max = 20)



CloseArima <- ts(LnClose, start = c(1984, 09), frequency = 365)  
FitCloseLn <- auto.arima(CloseArima)  
FitCloseLn

## Series: CloseArima   
## ARIMA(0,1,0) with drift   
##   
## Coefficients:  
## drift  
## 7e-04  
## s.e. 3e-04  
##   
## sigma^2 estimated as 0.0008698: log likelihood=16722.29  
## AIC=-33440.57 AICc=-33440.57 BIC=-33426.61

ForecastValueLn = forecast(FitCloseLn, h = 417)  
plot(ForecastValueLn)



ForecastValuesExtracted = as.numeric(ForecastValueLn$mean)  
FinalForecastValues = exp(ForecastValuesExtracted)

df <- data.frame(aapl\_test$Close, FinalForecastValues)  
naive\_rmse <- RMSE(df$aapl\_test.Close, df$FinalForecastValues)  
rmse\_results <- bind\_rows(rmse\_results,  
 data\_frame(method="ARIMA Prediction",   
 RMSE = naive\_rmse))  
rmse\_results

## # A tibble: 4 x 2  
## method RMSE  
## <chr> <dbl>  
## 1 Previous Value Prediction 1.51   
## 2 Moving Average Prediction 0.755  
## 3 KNN Prediction 2.59   
## 4 ARIMA Prediction 16.3

# Results

# Conclusion