ECON-M 524: Out of Sample Analysis of Pfizer

Final Project

Ritwik Budhiraja
MS Data Science
Indiana University, Bloomington
Indiana, USA
rbudhira@iu.edu

Abstract

This project aims to benchmark different time series forecasting methods. Throughout the timeline of this project, three out-of-sample models were built and compared. These models were used on a dataset containing historic time series data of the Pfizer stock (PFE). The Moving average algorithm outperformed other models.

Introduction

Pfizer is a multinational pharmaceutical company that has been in the news lately. The company's stock has seen a series of rises and declines throughout history. It is worthwhile analyzing the trends of the stock. I've compared the time series forecasting capabilities of the historical mean, direct method, and moving average algorithms.

Figure 1: Pfizer

Data Description

- The dataset was downloaded from Yahoo finance.
- It contains historic time series data on the quarterly price and dividend values of the PFE stock ranging from 1985 to the present.
- The dataset required some initial data processing:
  ⇒ Two separate datasets were downloaded for the price and dividend values which were later joined.
  ⇒ Since the dividend values are updated each quarter and you only get daily, monthly, or annual data on price, only the final price of the stock at the end of each quarter was retained and then joined with the dividends' dataset using the date column. These data pre-processing tasks were carried out using MS Excel.

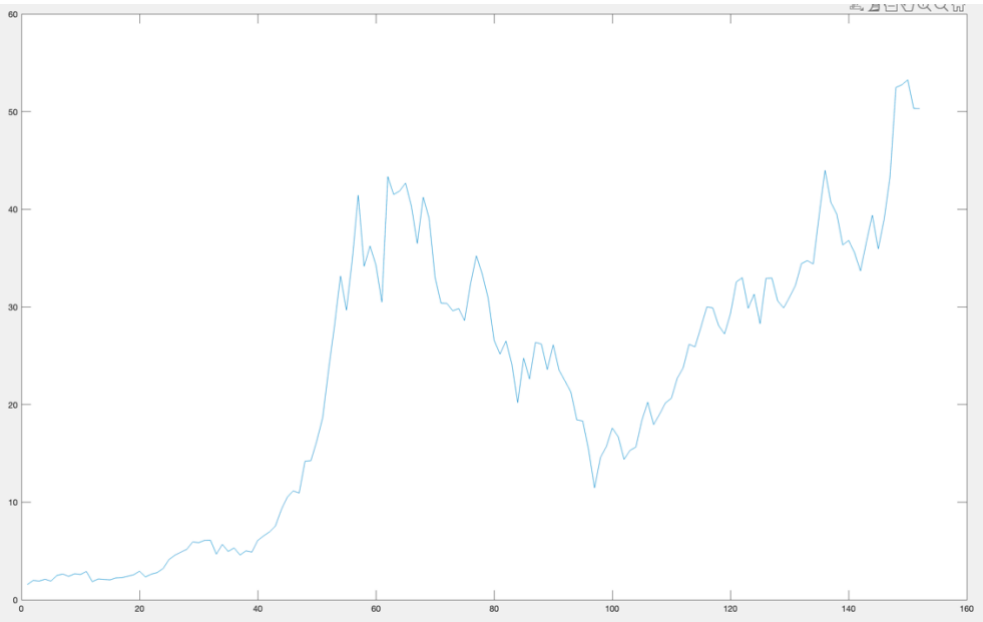


Figure 2: PFE data 1985-2022

This stock was chosen thinking of the volatile impact of COVID and its relationship with the Pfizer vaccine.
There was another interesting boom in the stock's price around the 1990s. This was because Pfizer had some strong product launches, such as the anti-inflammatory medicine Feldene and erectile dysfunction treatment Viagra during that time.

Methods

- First, I calculated continuously compounded return using the price column, i.e., the logarithm of the ratio of the price today to the price yesterday and the logarithm of the dividend-price ratio.
- Then I split the data into two halves and then used three algorithms to predict the latter 50% of the data:
  ⇒ Historical mean: This method takes the mean of the previous target values to predict the current value.
  ⇒ Direct method: This method uses the previous predictors and the target values to estimate the model for the current value.
  ⇒ Moving average: This method finds the rolling means of a few previous target values to predict the current value.
- I compared all these models for two different setups of Horizon-h regression: 1 and 12. Multiple-horizon regression is used to forecast the sum of the price of the asset for the next h days.
- Furthermore, I tried two different window sizes of moving averages: 10 and 50.

Results and Conclusion

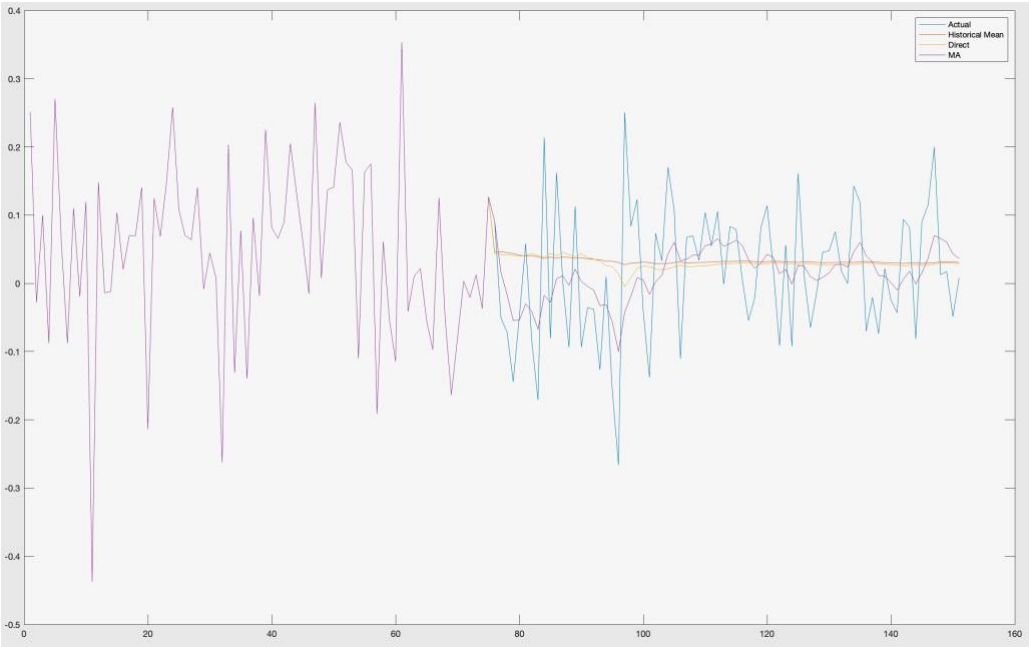


Figure 3: HORIZON-1 | MA 10 PLOT

```
MSE_hm =

    0.0100

MSE_direct =

    0.0101

R2_direct =

   -0.0157

MSE_MA =

    0.0071

R2_MA =

    0.2865
```

Figure 4 HORIZON-1 | MA 10 STATS

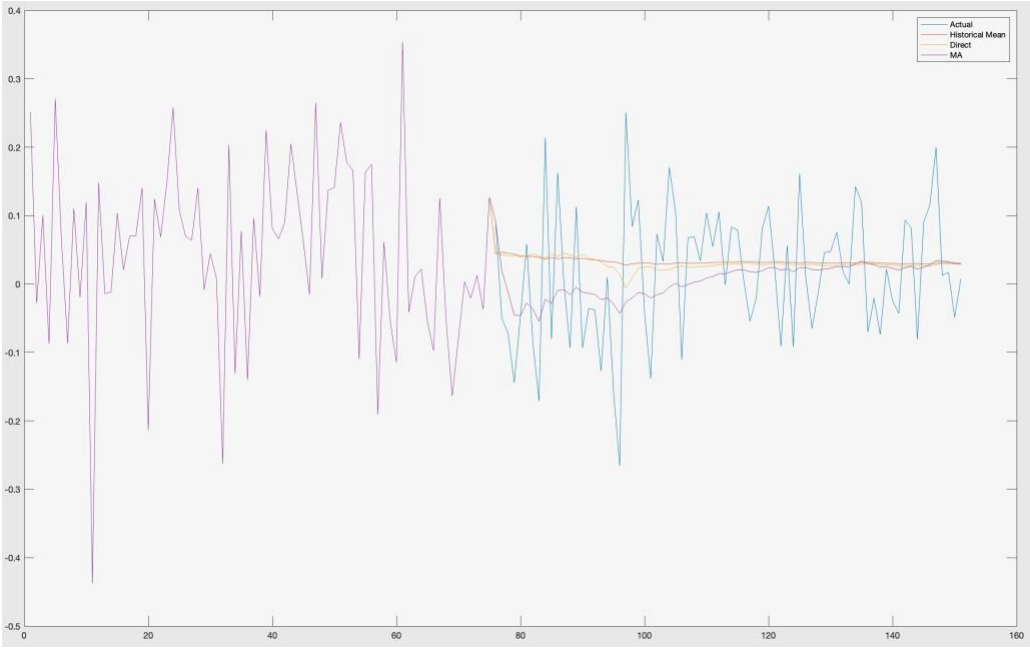


Figure 5: HORIZON-1 | MA 50 PLOT

```
MSE_MA =

    0.0085

R2_MA =

    0.1485
```

Figure 6: HORIZON-1 | MA 50 STATS

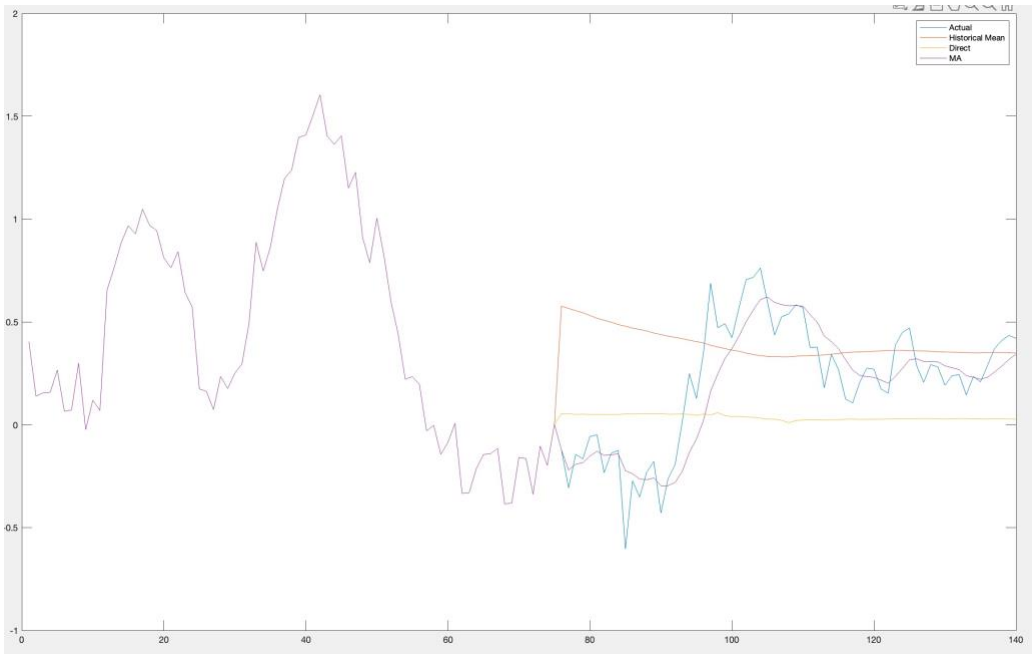


Figure 7 HORIZON-12 | MA 10 PLOT

```
MSE_hm =

    0.1652

MSE_direct =

    0.1276

R2_direct =

    0.2273

MSE_MA =

    0.0219

R2_MA =

    0.8675
```

Figure 8 HORIZON-12 | MA 10 STATS

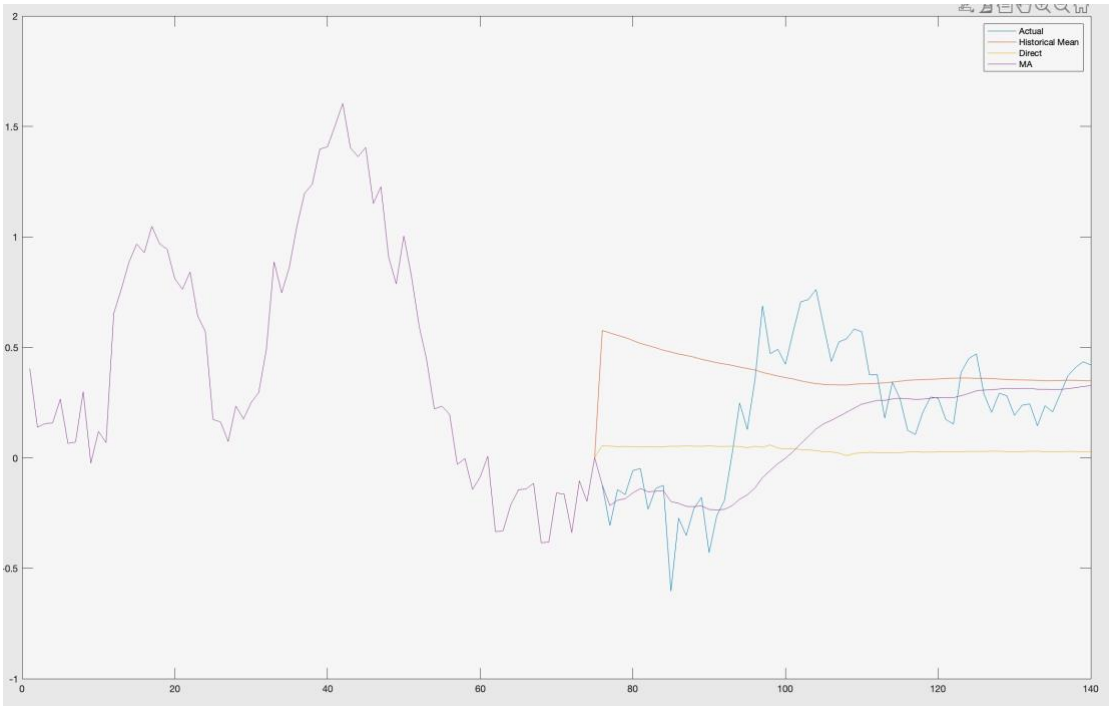


Figure 9 HORIZON-12 | MA 50 PLOT

```
MSE_MA =

    0.0719

R2_MA =

    0.5646
```

Figure 10 HORIZON-12 | MA 50 STATS

- It is evident that the historical mean and the direct method fail to make any good predictions in all scenarios. This could be probably because of the short size of the data.
- Moving average closely follows the actual values of our target variable. That is because the mean of only a previous few values is used to perform the prediction.
- The smaller the moving average window, the more local the prediction.
- Out-of-sample R-squared is used to measure the performance of each model. Mean squared error of the historical mean method is used as a benchmark.

- In the case when h was equal to 1, the direct method performed even worse than the historical mean method, and the R-squared value for it was negative.
- The best R-squared value was achieved when h was 12 and the moving average window size was 10. This is because the prediction was more local to the neighboring points due to the small size of the moving average window.
- Based on this information, dividend price ratio wasn't a good predictor for my data, and moving average performed the best out of the models in hand.

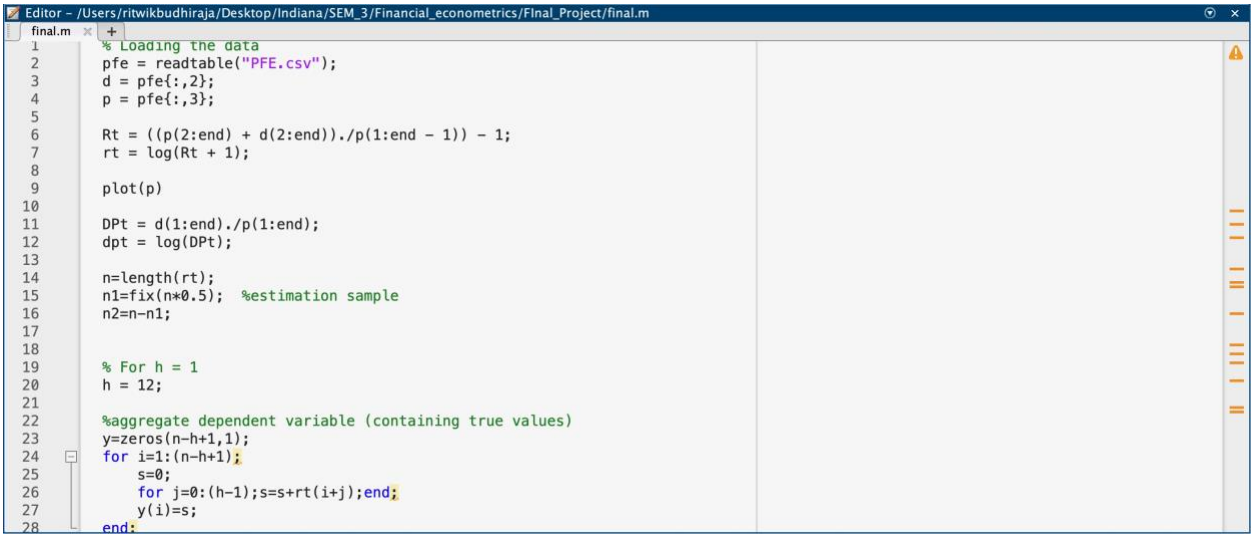
Future Scope

Models like ARMA, ARIMA, and SARIMA can be used to capture further seasonality in the data.

References

1. https://finance.yahoo.com/quote/PFE/
2. https://www.nasdaq.com/articles/pfizer-stock-history-how-drugmaker-became-industry-giant-2017-02-28
3. https://towardsdatascience.com/time-series-models-d9266f8ac7b0

Appendix



Editor - /Users/ritwikbudhiraja/Desktop/Indiana/SEM_3/Financial_econometrics/FInal_Project/final.m

final.m

```
y_true=y((n1+1):(n-h+1));

%historical mean
y_hm=zeros(n2-h+1,1);
for i=1:length(y_hm);
    y_hm(i)=mean(y(1:(n1-h+i)));  %expanding window
end;
MSE_hm=mean((y_hm-y_true).^2)

%direct method
y_direct=zeros(n2-h+1,1);
for i=1:length(y_direct);
    res = ols(rt(2:(n1-h+i)),[ones(n1-h+i-1,1), dpt(1:(n1-h+i-1))]);
    a_lh = res.beta(1);
    b_lh = res.beta(2);
    y_direct(i)=a_lh+b_lh*dpt(n1-1+i);
end;

MSE_direct = mean((y_direct-y_true).^2)
R2_direct =1-MSE_direct/MSE_hm

%Moving Average Method
y_MA=zeros(n2-h+1,1);
type = 'linear';
y_MA = movavg(y((n1+1):(n-h+1)),type,50);

MSE_MA = mean((y_MA-y_true).^2)
```

```
R2_MA =1-MSE_MA/MSE_hm



yplot_true=[y(1:n1);y_true];  %These are for plotting purpose
yplot_hm=[y(1:n1);y_hm];
yplot_direct=[y(1:n1);y_direct];
yplot_MA=[y(1:n1);y_MA];

plot(yplot_true)
hold on
plot(yplot_hm)
plot(yplot_direct)
plot(yplot_MA)
hold off
legend('Actual', 'Historical Mean', 'Direct', 'MA');

```

You can change the values of h and the moving average window size to get the desired outputs.