Final Project
Brian Ferrell
ECON 641

Introduction

I will forecast the price of the stock for CDW for the next 10 days. I will buy this stock using $1000 today if I predict the price of the stock 10 days from now to be at least 1% higher than the current price today. I will either buy it today if the 10-day-ahead price is 1% higher than the current price and sell it at the end of the 10 days, or put my money in a high yield savings account that accrues daily interest for the next 10 days.

If I predict the price to be 1% greater than the current, I will buy. If I predict the price to not be 1% greater, I will put the money in the high yield savings account (Ally bank .5% daily interest).

According to my forecast, 19 out of 20 times the 10-day ahead price of CDW will be between 165.97 and 200.46. The out of sample forecasted value of that 10-day ahead price is 183.22, thus my decision is to not buy CDW and will put the $1000 in the Ally bank high yield savings account.

Loss Function

My loss function is a function of three variables: current CDW price pt , my 10-day-ahead forecast of the price p^f_{t+10}, the price of CDW 10 days ahead p_{t+10}.

Example 1: I predict the CDW price to be $183.73 10 days from now which is at least 1% higher than the current (181.91) and buy $1000 of shares today at $181.91. If after 10 days the price is $184.00 I will sell my shares making 2 dollars and 9 cent profit per share (or (-2.09 * number of shares) loss). If the 10-ahead-price is $180.91 I will lose 1 dollar per share (loss of 5).

Example 2: I predict the CDW price to be $182 10 days from now which is not at least 1% higher than the current price (181.91) and therefore I do not buy, but I put the $1000 in the high yield interest rate account and after 10 days with .5% APY, I make 50 cents (-.50 loss). If after 10 days the price is $184.00 I've lost an opportunity to make 2 dollars and 9 cent profit per share (or (+2.09 * number of shares) loss), but I still made 50 cents from the savings account so really the loss would be +2.09 * number of shares - .50. If the price after 10 days was 150, my loss would be -.50 because I did not buy and I made money from the savings account.

Details steps:

A. First, when I predict the CDW price to be at least 1% higher than current price ($p^f_{t+10}>$ (pt * .01 + pt)). I buy qt = 1000/pt shares of CDW and make price difference times quantity ($p_{t+10} -$ pt)qt

B. Second, when I predict CDW 10-day-ahead price to not be at least 1% higher than the current price (p^f_{t+10} < (pt * .01 + pt)). Then, I don't buy anything and I put the $1000 in the savings account for 10 days.

- If the price falls below pt I have a loss of whatever amount I made from the savings account (-.50 cents)
- I lose possible gains if the price goes up but it is offsetted by the money made from the savings account; (((p_{t+10} − pt)qt) - 50 cents) when price goes up p_{t+10} > pt

C. Using dummy (indicator) variables $D_1 = D(p^f_{t+10}$ >= ((pt * .01) + pt)) and $D_2 = D(p_{t+10} > pt)$

D. Loss function formulation and written examples

$$L(p_t, p_{t+10}, p^f_{t+10}) = (-D_1(((p_{t+10} - pt)qt))) + ((1 - D_1)D_2(((p_{t+10} - pt)qt)) - .50$$

Example 1: $L(181, 184, 183) = (-\ 1\ (((184 - 181)5))) + ((1 - 1)(1)\ (((184 - 181)5) - .50$

= -15.00

Example 2: $L(181, 180, 183) = (-\ 1\ ((180 - 181)5)) + ((1 - 1)(0)\ (((180 - 181)5) - .50$

= 5

Example 3: $L(181, 184, 181) = (-\ 0\ (((184 - 181)5))) + ((1 - 0)(1)\ (((184 - 181)5)) - .50$

= 14.50

Example 4: $L(181, 150, 179) = (-\ 0\ (((150 - 181)5))) + ((1 - 0)(0)\ (((150 - 181)5))) - .50$

= -.50

This loss function makes sense for this decision because I don't think there is a purpose in buying this stock unless the predicted price is at least 1% higher than the current price, and 1% seemed like an appropriate threshold.

The code for the loss function as well as example 1 implementation can be found on page 1 of Appendix.

Data & Stationarity & Exogeneity

1. Data

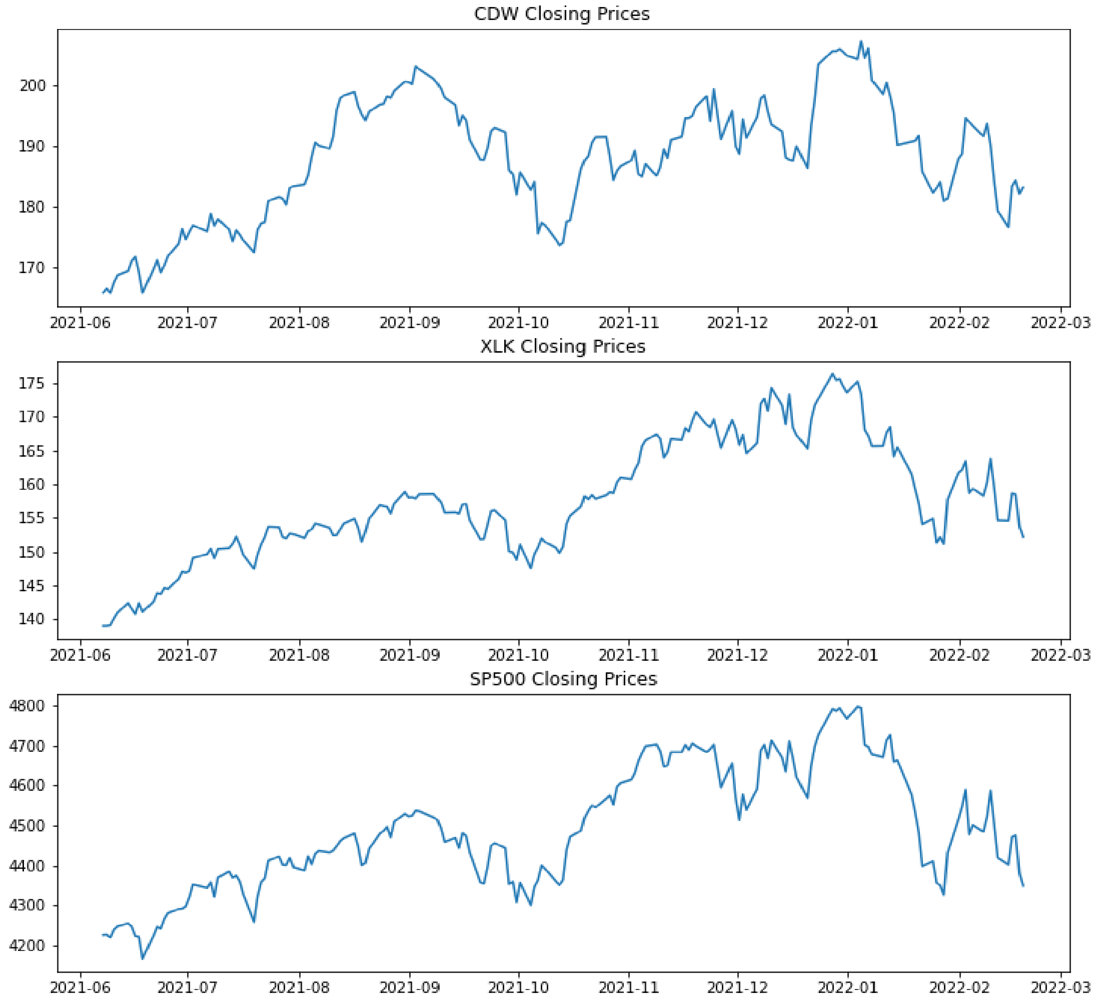

Figure 1: Plot of each data series

Above shows a subplot figure for the daily closing prices of the stock/ETF symbols: CDW, XLK, and SP500. All three show a slight linear trend, so there is no need for any transformations such as log transformation. CDW is a provider of technology products and services for business, government and education, XLK is an index fund that tracks the performance of the technology sector, and SP500 is an index that tracks the performance of 500 large companies listed on stock exchanges in the United States; all of which seemingly mirror each other in this figure.

2. AC and PACF (no transformation)

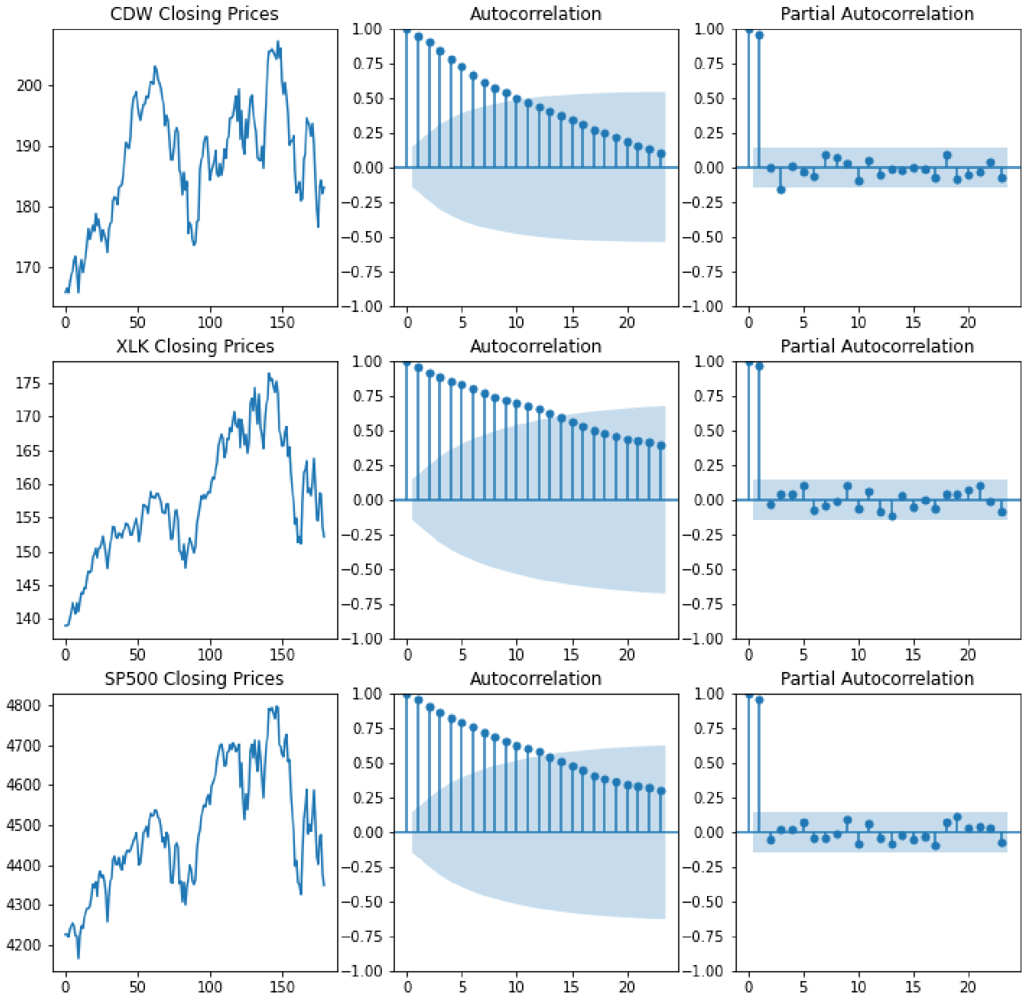

Figure 2: AC and PACF plots (page 7 of Appendix)

All three AC plots show that most of the points significantly differ from 0 and have a slow linear decrease, suggesting nonstationarity. In addition, ignoring that first spike from the PACF plots, our observation is that the spike next to it is close to one for each data series, suggesting nonstationarity.

3. Unit Root Tests (no transformation)

The selected deterministic regressor for all three variables will be a table for τ_τ because there is a trend.

CDW Unit Root

	coef	std err	t	P>|t|	[0.025	0.975]
Level.L1	-0.0595	0.029	-2.049	0.042	-0.117	-0.002
Diff.L1	-0.0092	0.079	-0.117	0.907	-0.165	0.147
Diff.L2	0.1629	0.078	2.078	0.039	0.008	0.318
Diff.L3	-0.0161	0.079	-0.204	0.838	-0.172	0.139
Diff.L4	0.0433	0.081	0.534	0.594	-0.117	0.203
Diff.L5	0.0989	0.081	1.221	0.224	-0.061	0.259
Diff.L6	-0.0771	0.082	-0.941	0.348	-0.239	0.085
Diff.L7	-0.0917	0.082	-1.118	0.265	-0.254	0.070
Diff.L8	-0.0975	0.082	-1.187	0.237	-0.260	0.065
const	11.3279	5.234	2.164	0.032	0.991	21.664
trend	9.316e-06	0.005	0.002	0.999	-0.011	0.011

Table 1: ADF Regression Summary for CDW

The selected number of lags started at 8 but ended up being 2 due to significance level. Since we are inside the critical value zone (higher than critical value) according to the ADF test on page 2 of Appendix, we cannot reject the null hypothesis that we do have a unit root, suggesting that the data is nonstationary and we have a unit root.

XLK Unit Root

	coef	std err	t	P>|t|	[0.025	0.975]
Level.L1	-0.0253	0.030	-0.853	0.395	-0.084	0.033
Diff.L1	0.0203	0.080	0.254	0.800	-0.138	0.179
Diff.L2	-0.0172	0.081	-0.212	0.832	-0.177	0.143
Diff.L3	-0.0355	0.081	-0.439	0.661	-0.195	0.124
Diff.L4	-0.2134	0.083	-2.578	0.011	-0.377	-0.050
Diff.L5	0.1131	0.080	1.407	0.161	-0.046	0.272
Diff.L6	0.0611	0.082	0.744	0.458	-0.101	0.223
Diff.L7	-0.0078	0.083	-0.094	0.925	-0.171	0.156
Diff.L8	-0.2008	0.085	-2.365	0.019	-0.368	-0.033
const	4.3208	4.364	0.990	0.324	-4.298	12.940
trend	-0.0026	0.005	-0.523	0.602	-0.013	0.007

Table 2: ADF Regression Summary for XLK

The selected number of lags started at 8 but ended up being 4 due to significance level. Since we are inside the critical value zone (higher than critical value) according to the ADF test on page 2 of Appendix, we cannot reject the null hypothesis that we do have a unit root, suggesting that the data is nonstationary and we have a unit root.

SP500 Unit Root

	coef	std err	t	P>|t|	[0.025	0.975]
Level.L1	-0.0348	0.033	-1.063	0.289	-0.099	0.030
Diff.L1	0.0504	0.080	0.630	0.530	-0.108	0.208
Diff.L2	0.0379	0.081	0.469	0.640	-0.122	0.198
Diff.L3	-0.0183	0.081	-0.226	0.822	-0.178	0.142
Diff.L4	-0.1735	0.083	-2.095	0.038	-0.337	-0.010
Diff.L5	0.0547	0.081	0.679	0.498	-0.104	0.214
Diff.L6	0.0914	0.082	1.120	0.264	-0.070	0.253
Diff.L7	0.0195	0.082	0.237	0.813	-0.143	0.182
Diff.L8	-0.2202	0.084	-2.610	0.010	-0.387	-0.054
const	160.4789	141.264	1.136	0.258	-118.503	439.461
trend	-0.0332	0.094	-0.353	0.725	-0.219	0.153

Table 3: ADF Regression Summary for SP500

The selected number of lags started at 8 and still ended up being 8 due to significance level. Since we are inside the critical value zone (higher than critical value) according to the ADF test on pages 2-3 of Appendix, we cannot reject the null hypothesis that we do have a unit root, suggesting that the data is nonstationary and we have a unit root.

4. Unit Root Tests (transformation)

Based on the previous unit root tests, which indicated nonstationarity, I conclude that they need to undergo first order differencing.

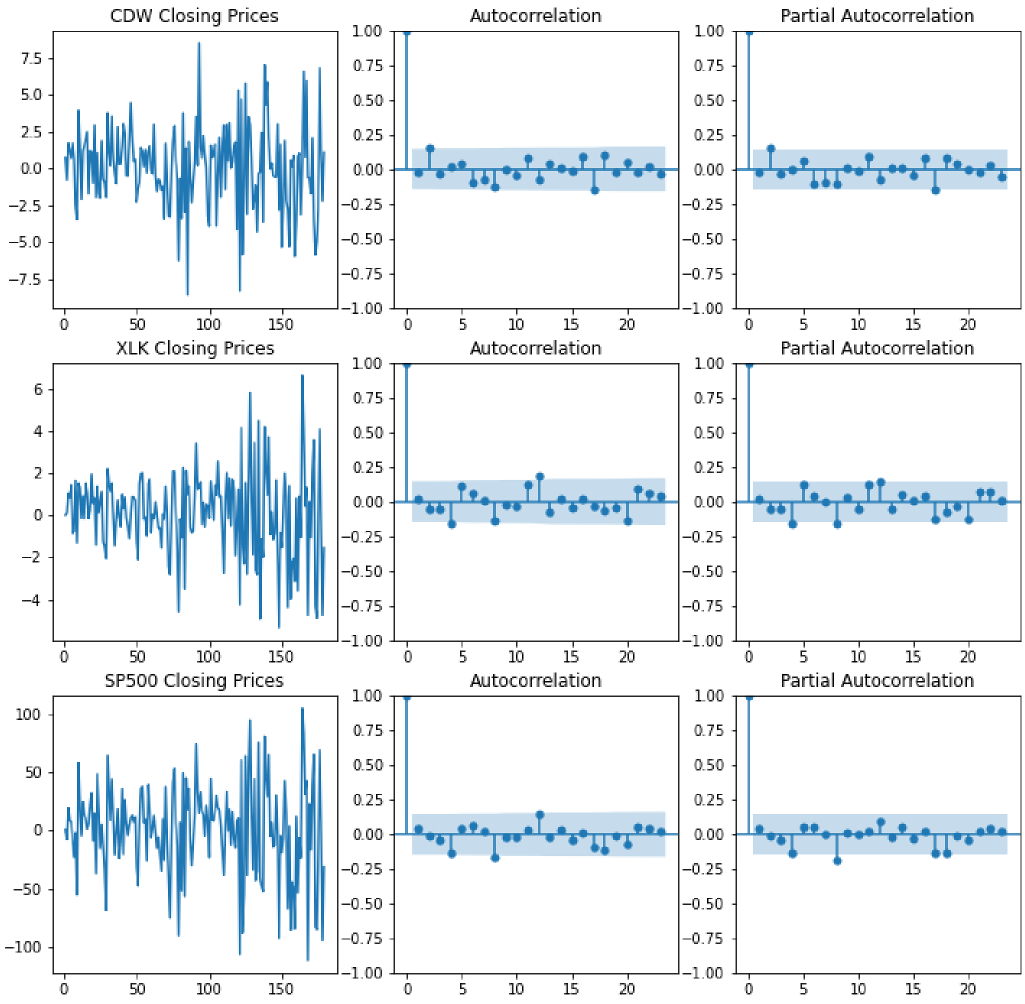

Figure 3: AC, PACF, transformation plots (page 8 Appendix)

The first-order differencing plots show that each crosses the mean often, suggesting that these first difference values are stationary. In addition, ignoring the first spikes, all three AC and PACF plots show nothing significant and look like white noises, suggesting stationarity.

Using unit roots tests, the selected deterministic regressor for all three variables will be a table for τ because there is no more trend.

CDW Unit Root

	coef	std err	t	P>|t|	[0.025	0.975]
Level.L1	-1.1345	0.227	-4.996	0.000	-1.583	-0.686
Diff.L1	0.0954	0.209	0.457	0.648	-0.317	0.508
Diff.L2	0.2414	0.192	1.260	0.210	-0.137	0.620
Diff.L3	0.2094	0.178	1.179	0.240	-0.141	0.560
Diff.L4	0.2295	0.168	1.368	0.173	-0.102	0.561
Diff.L5	0.3094	0.155	1.992	0.048	0.003	0.616
Diff.L6	0.2152	0.140	1.539	0.126	-0.061	0.491
Diff.L7	0.0970	0.120	0.808	0.420	-0.140	0.334
Diff.L8	-0.0140	0.083	-0.169	0.866	-0.177	0.149

Table 4: ADF Regression Summary for CDW

The selected number of lags started at 8 but ended up being 0 due to significance level. Since we are outside the critical value zone (lower than critical value) according to the ADF test on page 3 of Appendix, we can reject the null hypothesis that we do have a unit root, suggesting that the data is stationary and we don’t have a unit root.

XLK Unit Root

	coef	std err	t	P>|t|	[0.025	0.975]
Level.L1	-1.1988	0.259	-4.633	0.000	-1.710	-0.688
Diff.L1	0.2297	0.240	0.956	0.340	-0.245	0.704
Diff.L2	0.2115	0.226	0.937	0.350	-0.234	0.657
Diff.L3	0.1728	0.210	0.824	0.411	-0.241	0.587
Diff.L4	-0.0358	0.188	-0.190	0.850	-0.408	0.336
Diff.L5	0.0965	0.160	0.601	0.549	-0.220	0.413
Diff.L6	0.1641	0.141	1.163	0.247	-0.115	0.443
Diff.L7	0.1585	0.120	1.322	0.188	-0.078	0.395
Diff.L8	-0.0358	0.085	-0.420	0.675	-0.204	0.133

Table 5: ADF Regression Summary for XLK

The selected number of lags started at 8 but ended up being 0 due to significance level. Since we are outside the critical value zone (lower than critical value) according to the ADF test on page 3 of Appendix, we can reject the null hypothesis that we do have a unit root, suggesting that the data is stationary and we don’t have a unit root.

SP500 Unit Root

	coef	std err	t	P>|t|	[0.025	0.975]
Level.L1	-1.2242	0.249	-4.925	0.000	-1.715	-0.733
Diff.L1	0.2660	0.230	1.157	0.249	-0.188	0.720
Diff.L2	0.2896	0.216	1.340	0.182	-0.137	0.716
Diff.L3	0.2577	0.200	1.286	0.200	-0.138	0.653
Diff.L4	0.0747	0.179	0.417	0.678	-0.279	0.429
Diff.L5	0.1260	0.154	0.816	0.416	-0.179	0.431
Diff.L6	0.2156	0.137	1.577	0.117	-0.054	0.486
Diff.L7	0.2256	0.118	1.919	0.057	-0.007	0.458
Diff.L8	-0.0023	0.085	-0.027	0.979	-0.169	0.165

Table 6: ADF Regression Summary for SP500

The selected number of lags started at 8 but ended up being 0 due to significance level. Since we are outside the critical value zone (lower than critical value) according to the ADF test on page 3-4 of Appendix, we can reject the null hypothesis that we do have a unit root, suggesting that the data is stationary and we don't have a unit root.

All ADF implementations can be found on pages 1-2 of Appendix.

5. Supporting Variables

	CDW	SP500	XLK
CDW	1.000000	0.782013	0.766448
SP500	0.782013	1.000000	0.973606
XLK	0.766448	0.973606	1.000000

Table 7: Correlation between supporting variables and CDW (on page 6 of Appendix)

SP500 prices help me to forecast CDW because CDW happens to be in the SP500 and this index captures the pulse of the American corporate economy. The two are connected in the sense that all of these stocks tend to move together despite their differences. XLK prices help me to forecast CDW because this index seeks to provide an effective representation of the technology sector of the S&P 500 Index. To which CDW would fall under that category. Table 7 shows a table showing correlations between the supporting variables and CDW, proving their strong relationship.

6. Causality Tests

	CDW_x	SP500_x	XLK_x
CDW_y	1.0000	0.0484	0.0482
SP500_y	0.0194	1.0000	0.0130
XLK_y	0.0275	0.0826	1.000

Table 8: Grangers Causation Matrix (p-values)

P-value of 0.0484 at (row 1, column 2) represents the p-value of the Granger's Causality test for SP500 causing CDW, which is less than the significance level of 0.05. So, you can reject the null hypothesis and conclude SP500 causes CDW. P-Value of 0.0482 at (row 1, column 3) represents the p-value of the Granger's Causality test for XLK causing CDW, which is less than the significance level of 0.05. So, you can reject the null hypothesis and conclude XLK causes CDW. In addition, the results say that these two variables affect CDW and CDW affects these two variables. (See implementation and test results on page 4 of Appendix)

7. Forecast

Forecasting Table Summary

Model	Residuals Q-stat	Q-stat p-value	Modulus of Highest Root	10th Day Forecast p^f_{t+10}	10th Day Interval	Loss	Diebold-Mariano (stat, p-value)
ARIMA(0, 1, 0)	[11.05]	[0.3533]	0.0	183.220	([165.97, 200.46])	-0.5	(nan, nan)
ARIMA(0, 1, 2)	[6.40]	[0.7804]	0.0	182.877	([164.04, 201.70])	-0.5	(29.53, 0.0)
ARIMA(0, 1, 3)	[5.54]	[0.8519]	0.0	182.955	([164.81, 201.09])	-0.5	(8.65, 0.0)
ARIMA(1, 1, 6)	[1.43]	[0.9991]	1.48	185.06	([167.60, 202.50])	88.9	(-13.5, 0.0)
ARIMA(1, 1, 7)	[1.05]	[0.9997]	1.72	185.093	([167.67, 202.51])	88.9	(-13.07, 0.0)
ARIMA(2, 1, 7)	[1.29]	[0.9994]	1.44	185.07	([167.60, 202.52])	88.9	(-13.82, 0.0)
ARIMA(4, 1, 3)	[8.10]	[0.6199]	38.26	184.501	([168.00, 200.99])	-0.5	(-4.45, 0.0016)
ARIMA(4, 1, 4)	[5.02]	[0.8899]	1.05	184.472	([166.36, 202.57])	-0.5	(-3.53, 0.0064)
VAR	[11.12]	[0.3480]	25.19	183.83	*	-0.5	(-2.12, 0.063)

Table 9: Results (Pages 5-7 Appendix)

The table above shows a record of each model, its Q-statistic, Q-stat p-value, modulus of the highest root, 10-day ahead out of sample forecast price, 10-day ahead forecast confidence interval, custom loss ($p_t = 183.22, , p_{t+10} = 165.44$), and the Diebold Mariano statistic (being compared to ARIMA(0,1,0)) (Appendix pages 18-21). Technically these are ARIMA models because the differences (I) are all 1 which corresponds to the first order differencing. I probably did not need to estimate ARIMA models with high p and q because observational evidence strongly suggests White Noise, but I've been doing the same models all semester and wanted to see what would happen with different results. There is a * for the VAR model's confidence interval, because I was not sure how the inverse transformation worked for the confidence interval values.

Detailed summary of best model

Six models tie for being the best in terms of the loss, and logically it might make sense to choose the VAR model (VAR summary on page 8-9 on Appendix) as the best one because of how connected the variables are; however, all tests of residuals do not reject the null of White Noise suggesting that residuals of all models are White Noise. Due to this it seems to be a much simpler route to choose ARIMA(0,1,0) as the best model. The results of the Diebold Mariano test shows that we can reject the null hypothesis stating that at a 5% significance level these forecasts are the same, meaning the results are actually significantly better from the ARIMA(0,1,0) model, except for the VAR model.

Reasons for choosing this one: 1) it has White Noise residuals according to WN test; 2) it is the simplest of all models; 3) its loss is tied for the lowest, 4) the confidence interval is the most precise.

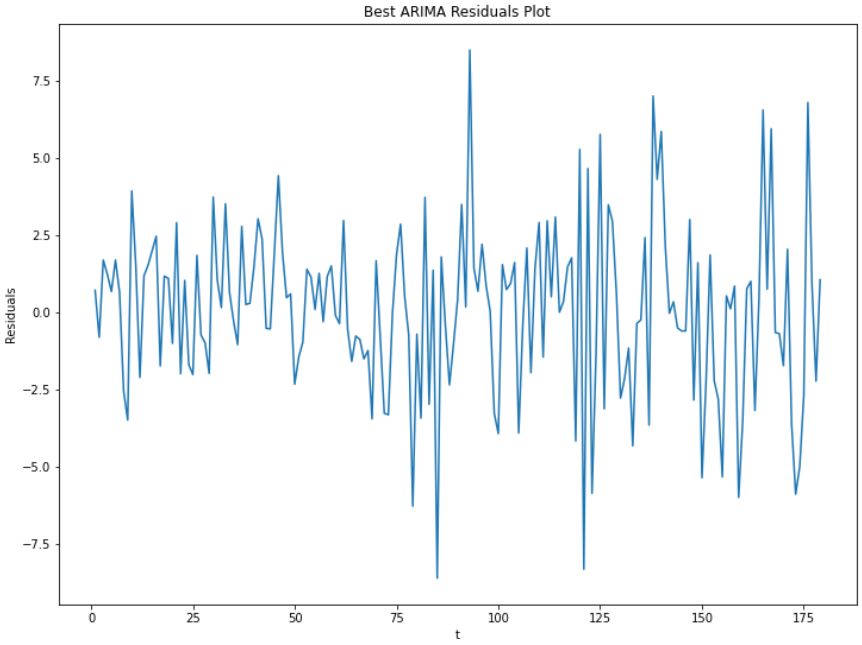

Figure 4: Residuals of best chosen model (Page 8 Appendix)

Figure 4 shows residuals of the best chosen model are around zero, cross zero often, move quickly, have no trend.

Since our best chosen model is ARIMA(0,1,0), we can compare the prediction interval to the one that had the closest difference to it (ARMA(1,1,7). ARMA(0,1,0) had a confidence interval of (165.98, 200.46); whereas, ARMA(1,1,7) had a confidence interval of (167.67, 202.51). Which shows that the ARMA(0,1,0) shrunk the interval by a difference of .36 (difference of first minus the difference of second).

In conclusion, none of these models were actually any good despite the fact that I did not lose money. The reason being is because the actual 10-day ahead price was way below what any of them predicted, and so in the future it might be better to factor either the confidence intervals in the loss function, or penalize the model's loss by adding in the differences between the actual and predicted price.