Mark Kipper

Final Report:

Cannabis Stock Price Prediction

Problem Statement

Cannabis used to be illegal in the United States. Then, in 1996, California voted to legalize it for medicinal use. In 2012, Colorado and Washington state were the first 2 states to legalize Cannabis for recreational use. Since then, the industry has thrived. There are several companies in the legal Cannabis market. Some of them invest in properties to grow cannabis, others invest in medical products, and still others invest in consumables infused with cannabis. Investors in these companies have seen their stock prices rise similar to the dot com bubble of the 1990s.

The question I intend to answer is: is this the right time to get into this market? The purpose of this project is to see if there is a reasonable way to predict the next day's stock price based on how the market has performed previously. This should, in no way, be considered sound financial advice and this model should not be used with the hope of realistically striking it rich in the market. Rather, this is an academic attempt to predict the next day's stock price on 10 of the top performing cannabis stocks in the market as of this writing.

Data Wrangling

Historical stock performance for all stocks in this project was downloaded from Yahoo! Finance. Since all of these stocks are relatively new, I was able to use the entire performance history since each stock's inception. I used 10 of the top performing cannabis stocks, 5 cannabis ETFs, and the Dow Jones Industrial average in different combinations to determine the best combination of features for predicting the next day's price.

For each stock, I set the timestamp as the index. I then added some new features to the dataframe. I shifted the data so that I could get the previous close price as well as the next day's closing price (my target variable). I also engineered a column for the difference in price for each day and the % change for each day. I then added 5 day and 10 day rolling averages as columns to give the model a little historical context in the features. Finally, I drop the "Adjusted Close" column. It contains redundant information.

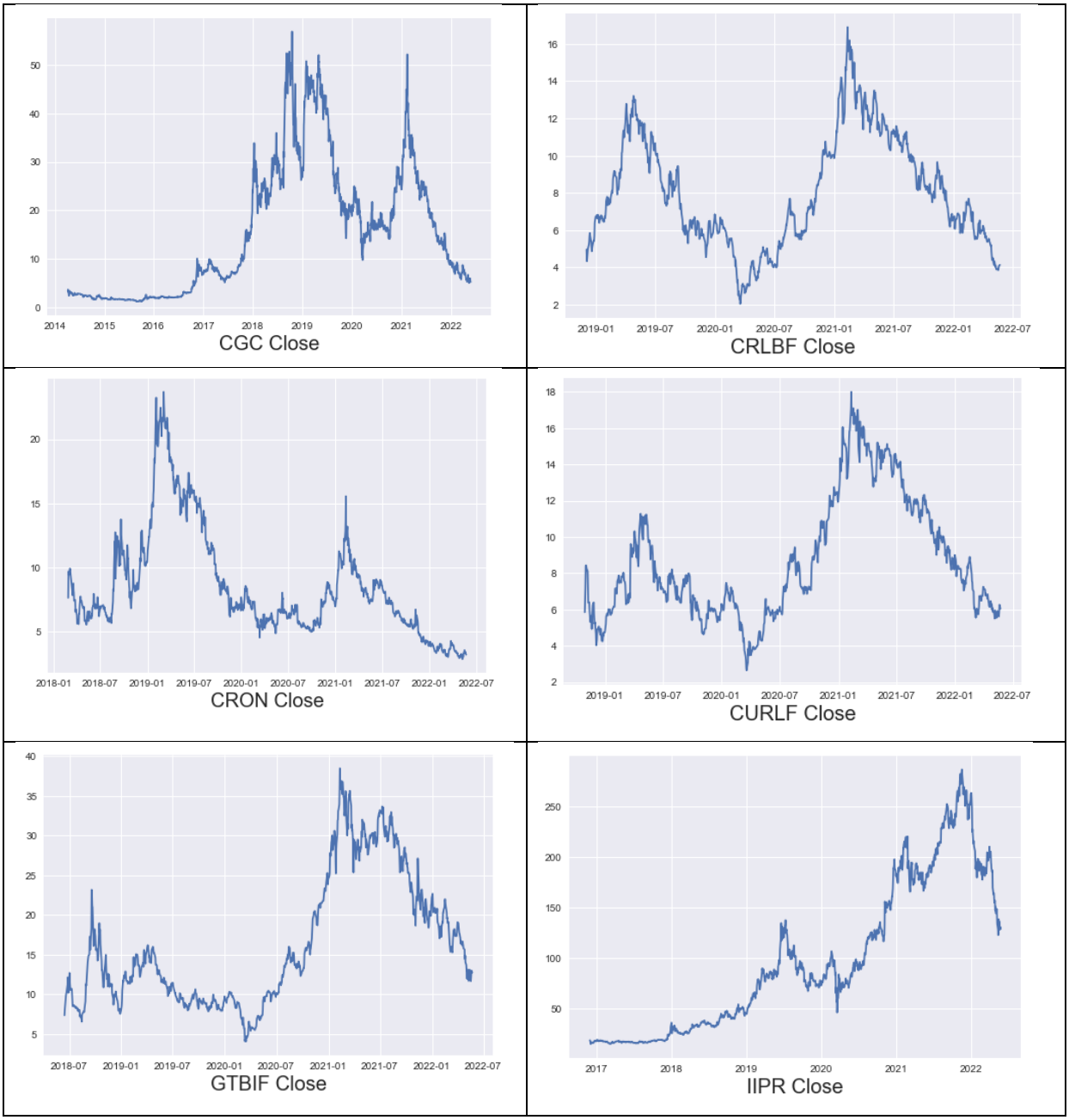
Because there are some stocks that are older than others, when I merge them together, I am often left with many rows of N/A values because some stocks do not have any price information for dates before they were publicly traded. In these cases, I create an isna() dataframe. This gives every row of data a True or False label depending on if there are any N/A values. This also duplicates all the column names,

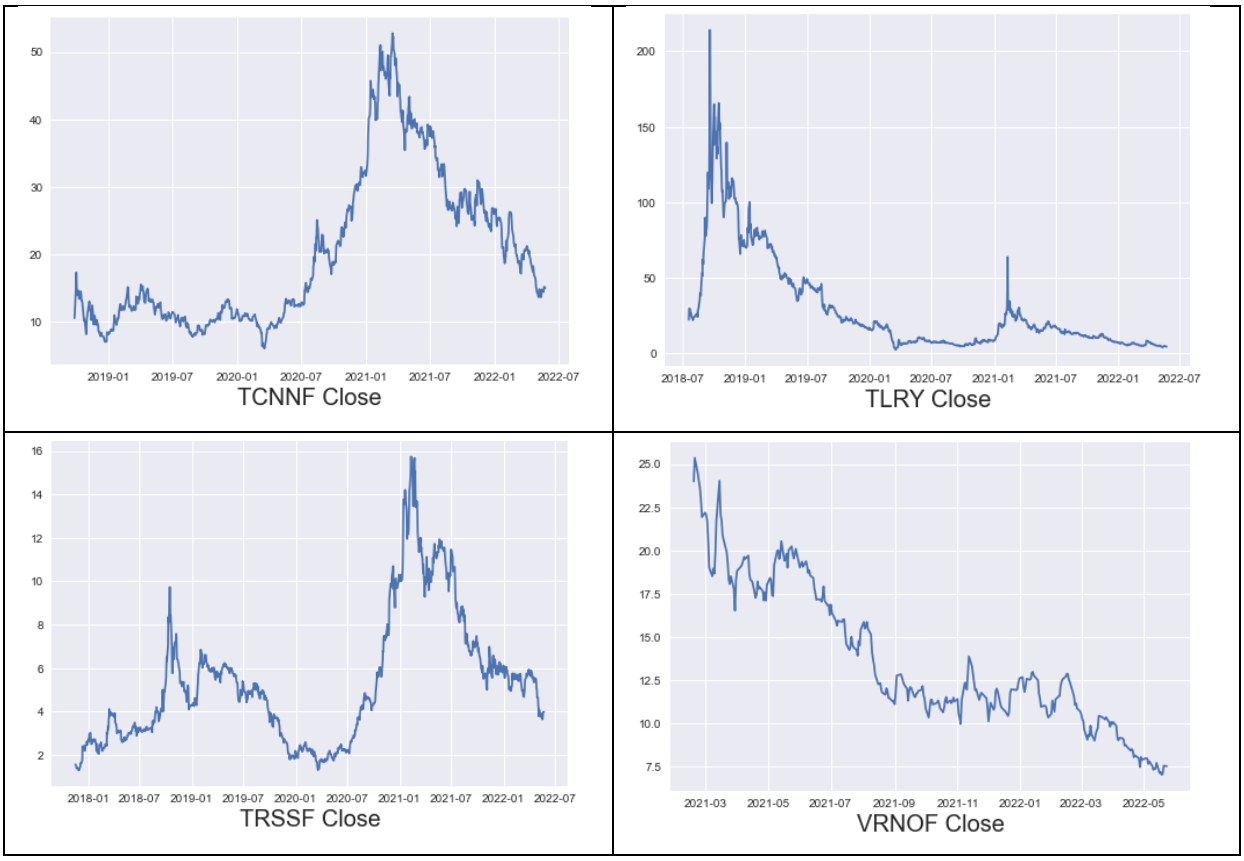
so I add a suffix to the duplicates. I fill the N/As from the original dataframe with zeroes and concatenate them both.

I decomposed the date time index and added month, day, year, and quarter columns.

Exploratory Data Analysis

I plotted the close of each of the 10 stocks to visualize the data.





This also gives me an idea of what I should expect from my models as I used the last 10% of entries as my test set.

It can be seen from the plots that all stocks are trending downwards. Also notable is that there are many peaks in the plot for each stock. Is this the end of the cannabis stock bubble or are we about to see a rally?

Model Selection

I am using a linear regression model. My focus rather, is the feature combinations which will give me the best result. I have outlined 7 different linear regression models:

1. Target stock + Dow Jones Industrial Average (Dow)
2. Target stock + all 5 ETFs
3. Target stock + ETFs + Dow
4. Target stock + remaining 9 stocks
5. Target stock + remaining 9 stocks + Dow
6. Target stock + remaining 9 stocks + ETFs
7. Target stock + remaining 9 stocks + ETFs + Dow

Because I am looking at the top 10 performing cannabis stocks, I ran these models 10 times in the above combinations.

For each model I used only as many dates as were available in the target stock. I used a 90/10 split for training and testing. With stock price prediction analysis, it is important that the training and testing sets are sequential. I scored the models using root mean squared error (rmse). Each stock was run through the seven combinations and the rmse was recorded for each in a hyper table. The lowest rmse is understood as the best performing model in this analysis. I sorted the hyper table by rmse and saved only the top performer. After all ten stocks were tested with all seven combinations, I gathered the first entry of each hyper table into a final hyper table to see which of the seven combinations performed best for each target stock.

Results

The hyper table details which combination performs best.

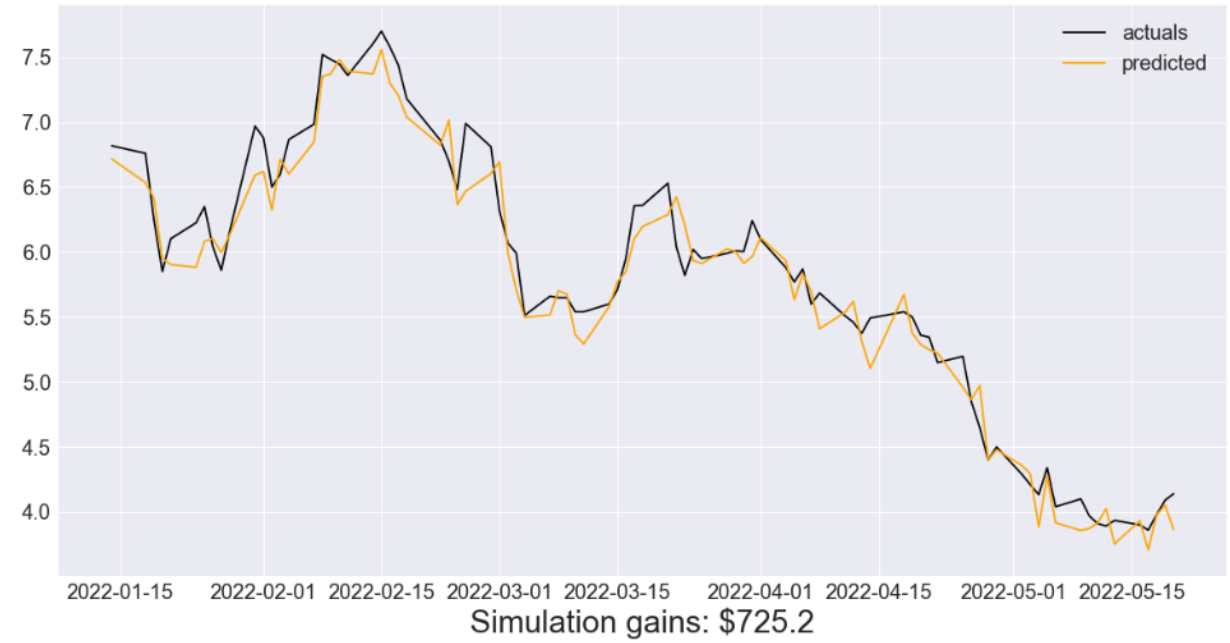
	version run	stock	version	model	rmse	mae	actual_price	predicted_price	rows	train_rows	test_rows	columns
0	1	cgc	1+ETFs	LinearRegression()	0.374605	0.258421	5.190000	6.002075	2047	1842	205	160
1	1	crlbf	1+ETFs	LinearRegression()	0.187848	0.149452	4.140000	3.864823	873	785	88	160
2	3	cron	All stocks	LinearRegression()	0.141724	0.113036	3.180000	3.525383	1067	960	107	264
3	1	curlf	1+ETFs	LinearRegression()	0.182203	0.142273	6.050000	6.231867	897	807	90	160
4	2	gtbif	1+ETFs+Dow	LinearRegression()	0.355154	0.273537	12.750000	13.132380	992	892	100	186
5	0	iipr	1+Dow	LinearRegression()	5.107221	4.160234	128.990005	134.334104	1377	1239	138	56
6	1	tcnnf	1+ETFs	LinearRegression()	0.386340	0.305602	15.000000	14.620530	919	827	92	160
7	0	tlry	1+Dow	LinearRegression()	1.289206	1.054205	4.490000	4.286098	968	871	97	56
8	1	trssf	1+ETFs	LinearRegression()	0.162104	0.126405	3.990000	3.773157	1134	1020	114	160
9	1	vrnof	1+ETFs	LinearRegression()	0.249591	0.202206	7.500000	7.297171	318	286	32	160

Using the results of the hyper table, I ran a basic simulation of hypothetical trades using my model on the best performing combination of explanatory variables for each target stock. Assuming a portfolio that contains these investments in unlimited quantity, the premise is: if the predicted closing price is greater than the actual Open price, then we sell 100 shares at the actual Closing price. After the test set has been processed in this way the sum of all the trades made, based on these criteria, is given. I have included the plot for the best performing model run for each stock along with the simulation gains based on the model.

CGC


actuals
predicted
20
18
16
14
12
10
8
6
2021-08
2021-09
2021-10
2021-11
2021-12
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
Simulation gains: $1655.3

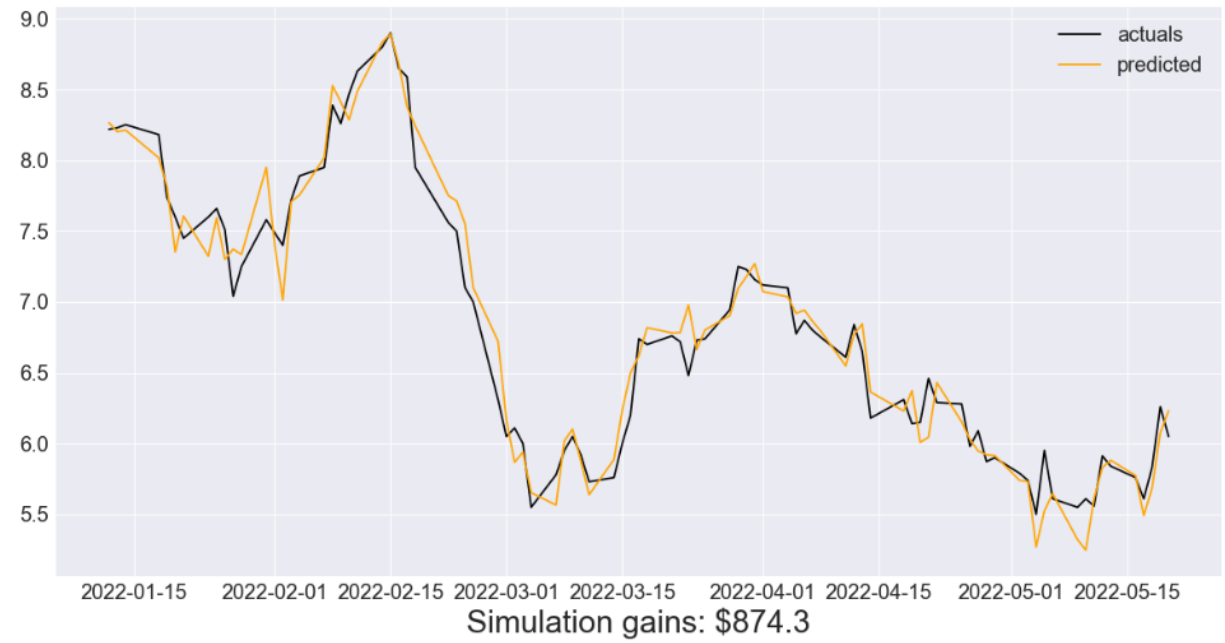
CRLBF


actuals
predicted
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
2022-01-15
2022-02-01
2022-02-15
2022-03-01
2022-03-15
2022-04-01
2022-04-15
2022-05-01
2022-05-15
Simulation gains: $725.2

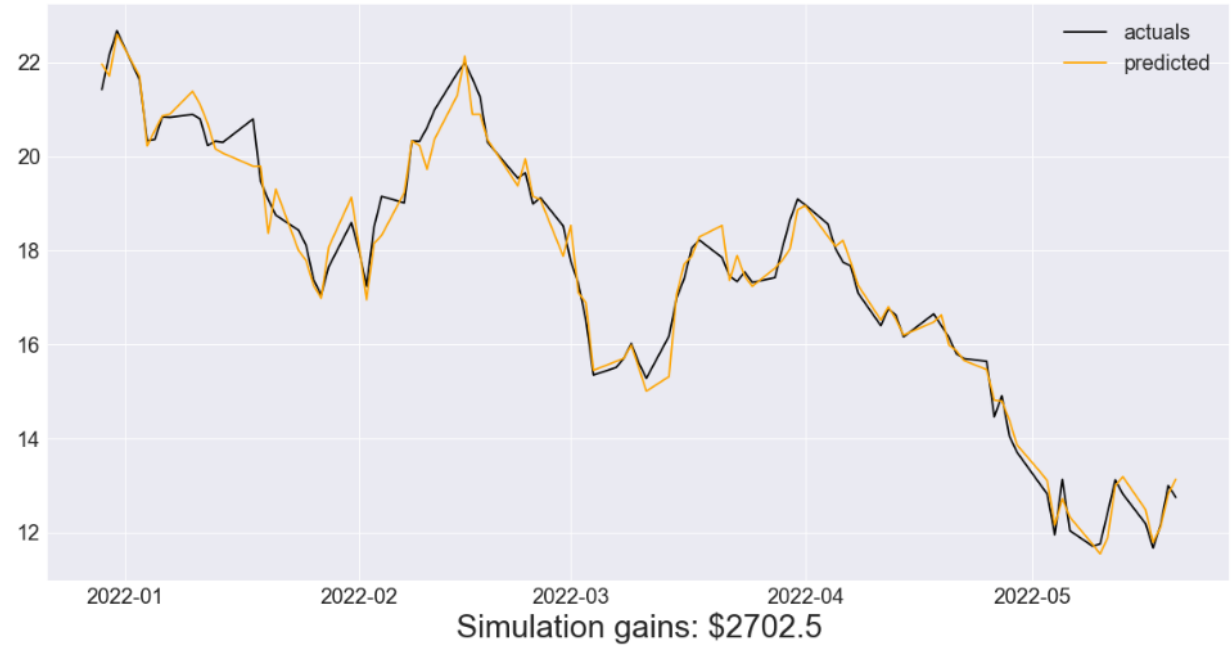
CRON


actuals
predicted
4.50
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2022-01
2022-02
2022-03
2022-04
2022-05
Simulation gains: $241.5

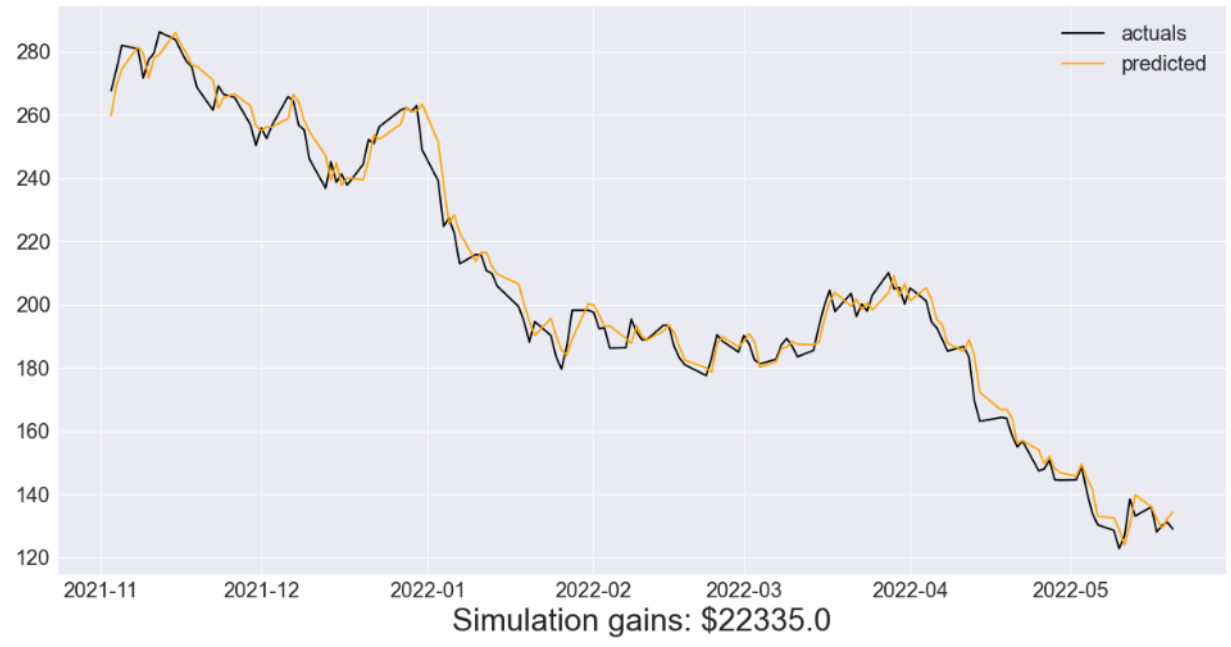
CURLF


actuals
predicted
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
2022-01-15
2022-02-01
2022-02-15
2022-03-01
2022-03-15
2022-04-01
2022-04-15
2022-05-01
2022-05-15
Simulation gains: $874.3

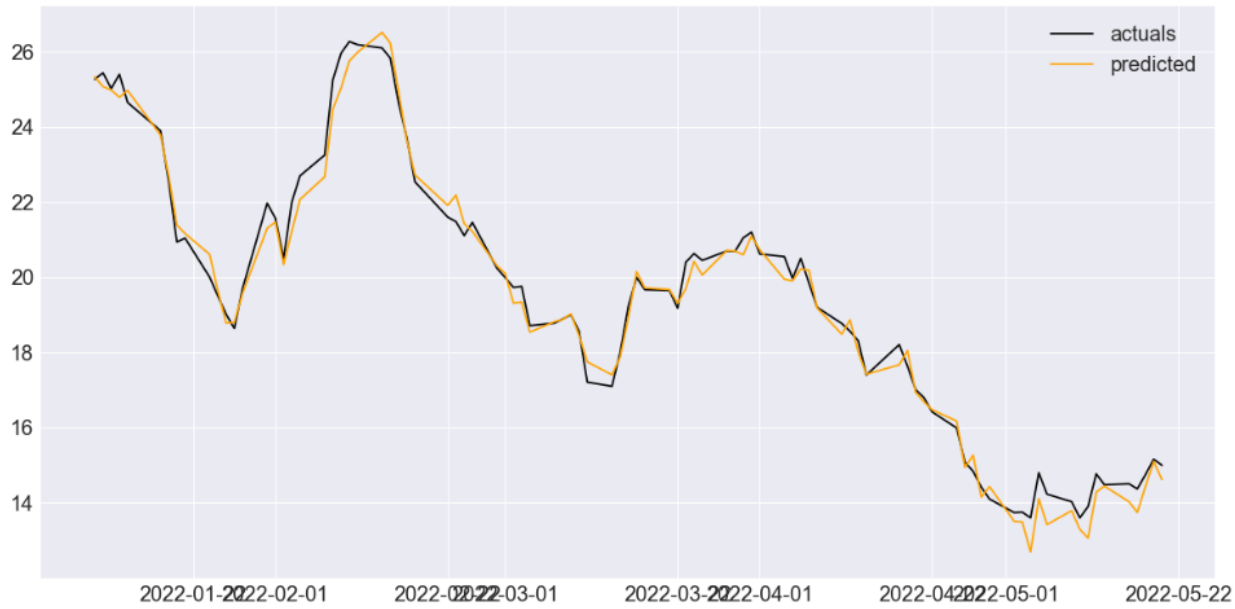
GTBIF


actuals
predicted
22
20
18
16
14
12
2022-01
2022-02
2022-03
2022-04
2022-05
Simulation gains: $2702.5

IIPR

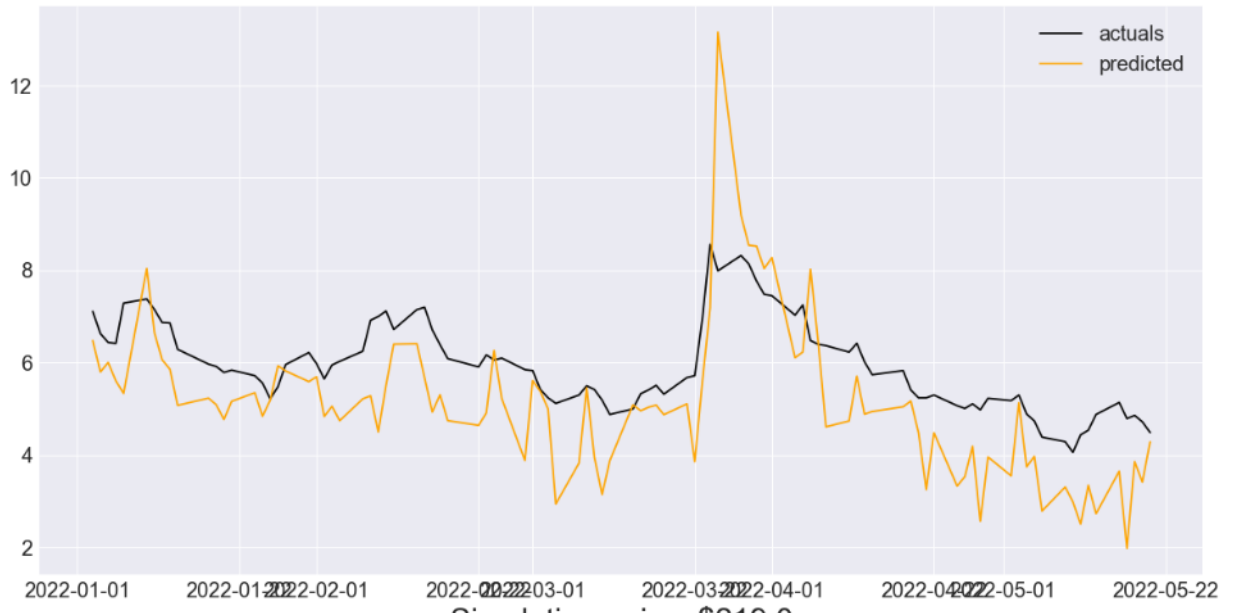

actuals
predicted
280
260
240
220
200
180
160
140
120
2021-11
2021-12
2022-01
2022-02
2022-03
2022-04
2022-05
Simulation gains: $22335.0

TCNNF



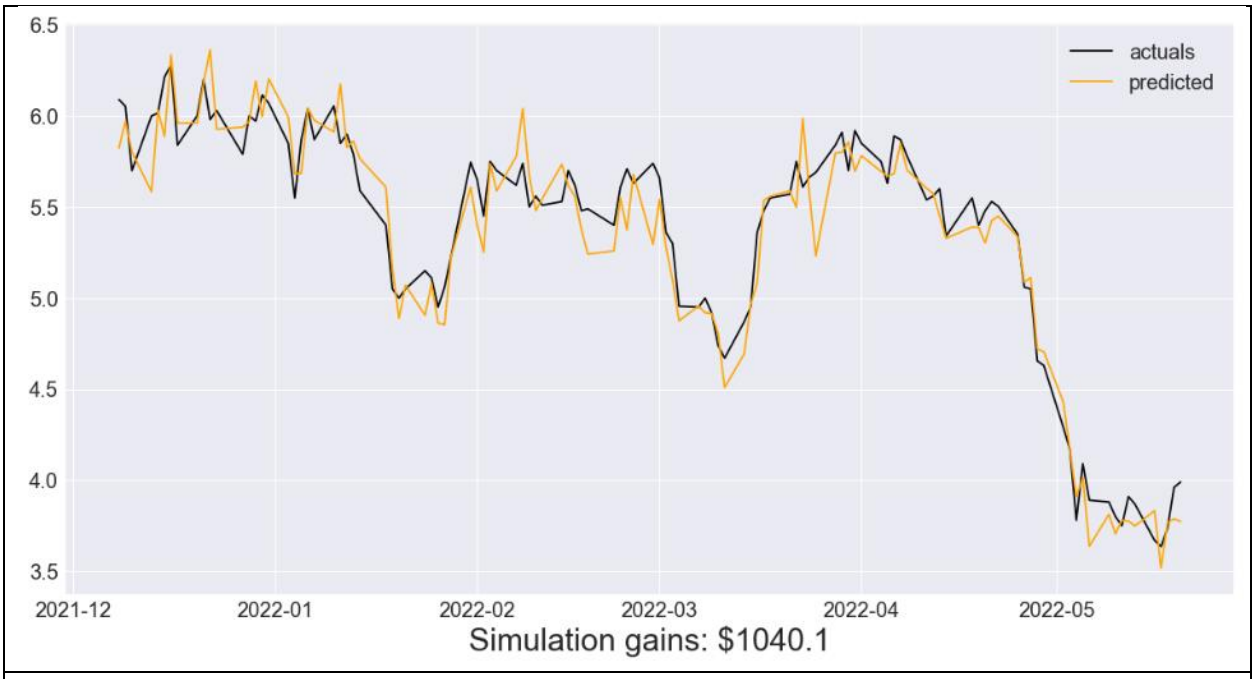
Simulation gains: $1700.6

TLRY

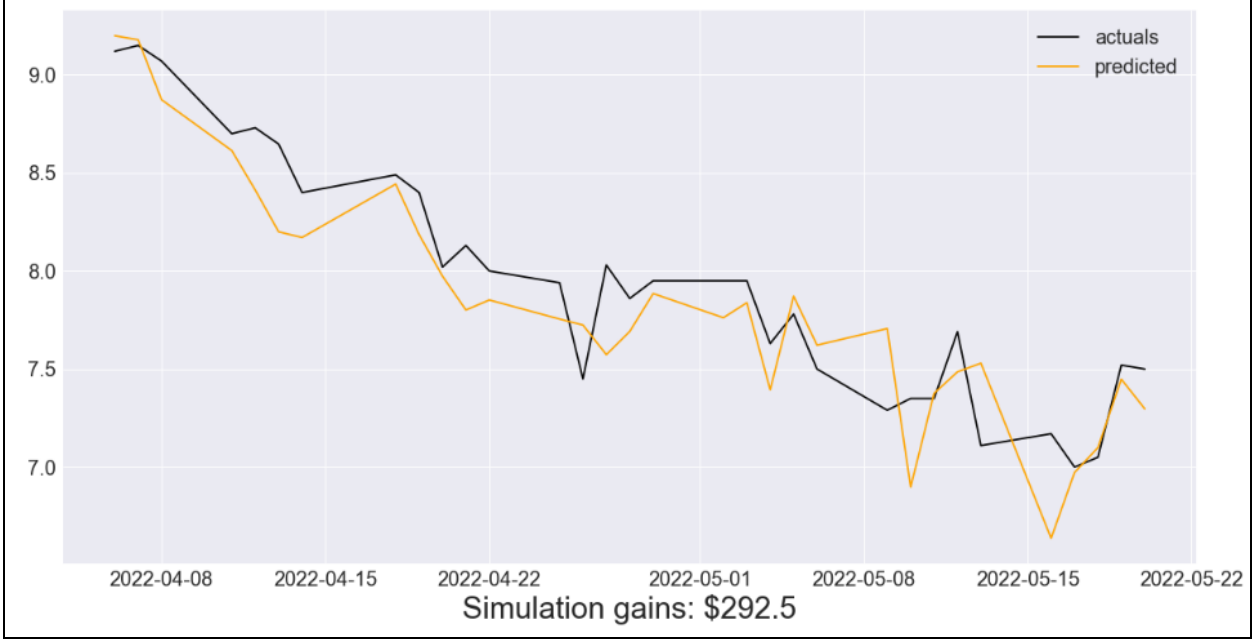


Simulation gains: $219.0

TRSSF



VRNOF



The total gains using just these 10 cannabis stocks and my basic simulation yields $31,786.00.

Future Work

There are many ways to perform a stock price prediction. ARIMA models and LSTM neural networks have received notable results. I would like to perform these predictions using one or both methods to compare the results.

There are also many types of regression models that can be used. My focus here was the combination of explanatory variables that can be included in the multiple regression model to yield the best results.

I would have like to include a ‘Buy’ simulation, but to be consistent with the ‘Sell’ simulation, I would have to re-run the models using the ‘next_Open’ target variable for all stocks.