Project 2 – Whitepaper Draft
Jerica Tripp – DSC680
July 15, 2022

Business Problem

The goal of my project is to try and predict closing stock prices of select banks. It would be beneficial for stock traders to be able to utilize historical data to make decisions on which bank stock to buy or sell. The data is very large and is not an easy task for an individual to pour through and so having an automated way to making buying decisions would be beneficial for brokers and other investors.

For my analysis, I will be focusing on the following bank's stock data: J.P. Morgan, Discover Financial Services, American Express, Citi Bank, Bank of America, US Bank, and Wells Fargo. The idea is that users may only have the option to buy one stock, and this model would help them to make that decision.

Background/History

Historically, predicting stock closing prices has been a difficult task. Mitchell (2021) of Investopedia, states: "Predicting the market is challenging because the future is inherently unpredictable". Issues that arise, such as the 2008 housing market crash or COVID cannot be predicted, so all predictions around stock should be taken with a grain of salt. Investors can and should leverage trend information to make good buying decisions as best they can.

Common techniques to predict stock prices include moving average, exponential moving average, long and short-term memory models (Li, 2022). Aspects that these methodologies don't account for well are seasonality and holidays. Facebook's forecasting tool, Prophet, offers a solution to bake in seasonality and other date trends that are difficult to add to other models. For my analysis, I will be using Prophet to make my predictions.

Data Explanation

Data Gathering

AlphaVantage is an open-source API that allows users to utilize stock data. They offer intra-day data down to 5-minute intervals. For my analysis, I will be using the version of this data that provides the last two years of daily closing data. The data is accessed using one stock at a time. The ticker is passed to the API and the following fields in the below data dictionary section are returned.

Data Dictionary

Main Data

1. Timestamp – date value of record
2. Open – opening stock price
3. High – highest stock price for the day
4. Low- lowest stock price for the day
5. Close- closing stock price for the day
6. Volume- volume of stock for the day
7. Ticker- this stock identifier (will be a code like DFS, JPM, etc.)

Data Prep

The data required a few manipulations to make it work for Prophet. Prophet only accepts a dataframe with two columns: y: a numeric value and ds: a date. My first step was to convert the date field provided into a timestamp because it was of 'object' type originally. I then created a dataframe that only contained the ticker value, the closing price, and the date.

I was able to do most of my statistical analysis of the data using the above dataframe. For the model, I had to feed one ticker's data at a time. To accomplish this, I created a list of dataframes and looped over each item in the list when running the model.

Prophet accepts a 'holiday' argument. This allows the model to consider holidays in its algorithm. To do this, I had to create a dictionary for each holiday that the US stock market observes.

Analysis

My first step in my exploratory data analysis was to ensure that every stock had similar date ranges and record counts. I found that all stocks had the same date range (November 1, 1999 to July 26, 2022) ,except for Discover, which had a date range of June 14, 2007 to July 26, 2022. This is accurate, as Discover did not start trading on the stock market until June of 2007 (Discover, n.d.). Discover has 3,806 records and all others have 5,720 records.

Using the describe function, I was able to assess the descriptive statistics of all the stocks combined closing price, as well as each stock individual closing price. Figure 1, below, shows these stats. We can see that JPM has the highest mean, Wells Fargo has the lowest, and the average closing price of all are around $49.14. The standard deviation is interesting, with USB & WFC having similar deviations while the rest range from low twenties to high thirties. The minimums show that there have been extreme lows, although this could be initial offering prices. A plot of the closing prices should tell us more.

Closing Stock Price Summary Statistics

	all	DFS	JPM	AXP	C	BAC	USB	WFC
count	38126.00000	3806.00000	5720.00000	5720.00000	5720.00000	5720.00000	5720.00000	5720.00000
mean	49.14313	51.81464	62.96439	71.42865	45.09183	34.45394	35.16919	43.97322
std	28.54789	30.36424	35.01464	37.00227	18.97208	20.86484	11.86619	11.89401
min	1.02000	4.89000	15.45000	10.26000	1.02000	3.14000	8.82000	8.12000
25%	30.97250	23.41000	38.49000	45.80000	37.98750	15.75000	25.03750	33.76000
50%	45.81000	54.27000	47.36000	57.71500	48.27000	29.92000	32.44000	46.25500
75%	57.74000	70.57500	83.91750	89.76250	54.88000	48.27250	43.67000	53.17500
max	198.38000	135.38000	171.78000	198.38000	81.91000	89.01000	63.25000	73.00000

Figure 1

Looking at the distribution of closing prices of all the bank stocks we've selected, we can see in Figure 2 that the distribution is left-skewed. Closing prices tend to have more outliers in the higher range, with the distribution heavier in the lower price range. Around 50, the distribution peaks.

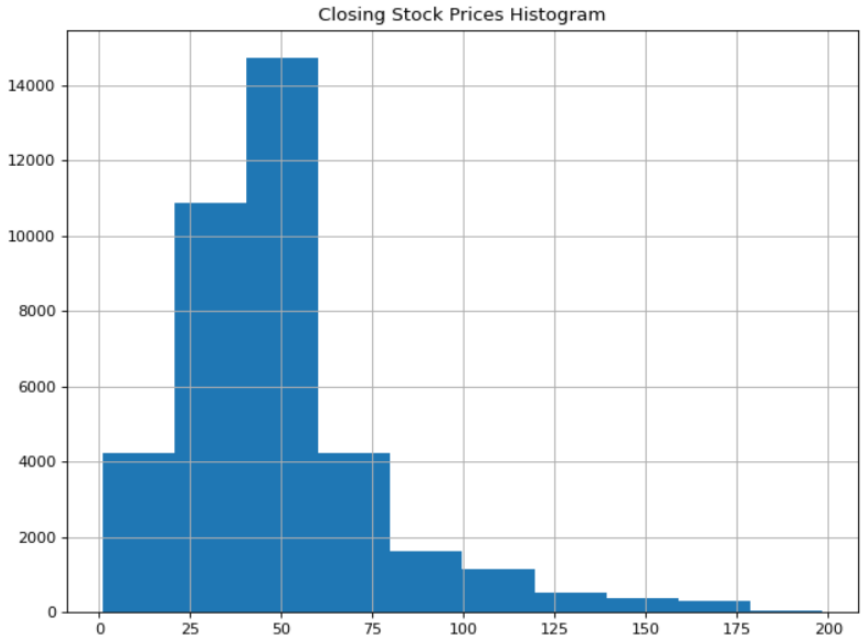

Figure 2

The below histograms in Figure 3, show the closing price distributions of each bank stock we're reviewing. We can see that the distribution of Discover, JP Morgan, American Express, Bank of America, and US Bank are all left-skewed, like the overall distribution in figure 2. However, Wells Fargo and Citi take on more of a normal distribution.

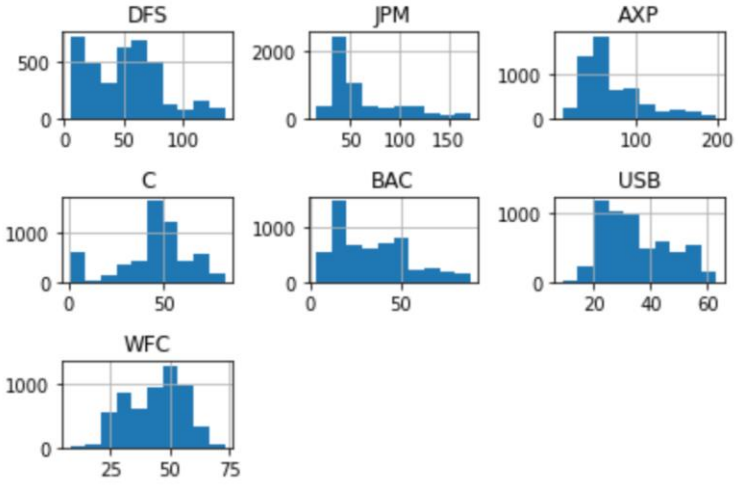

Figure 3

When plotting the closing price for each day, we see similar trends for the combined and the individual stocks. The 2008 recession and the impact of COVID on the stock market can clearly be seen in both charts in Figure 4.

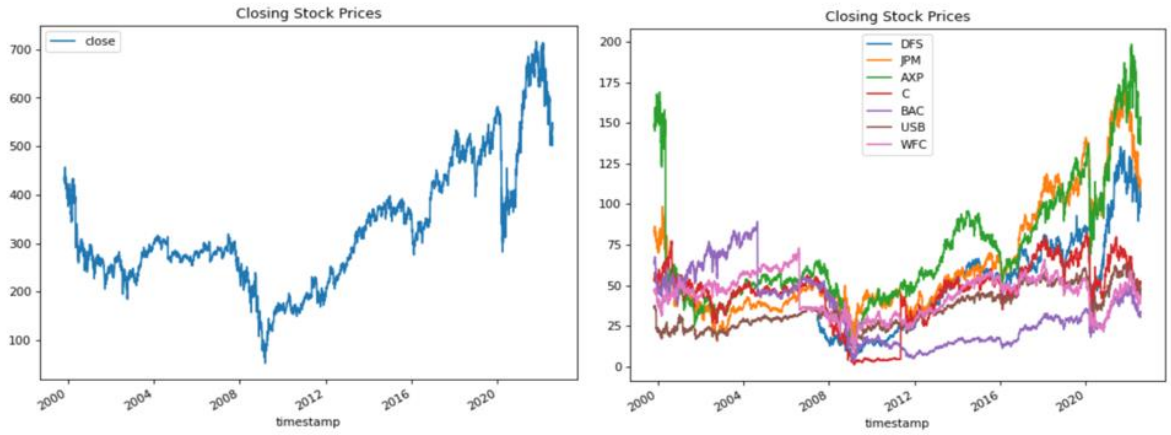

Figure 4

Methods

For my model, I will be using a training and validation population. I'd like to be able to advise the users as to the confidence they should have in the recommendation. I will be using MAPE and RMSE to gauge the accuracy of the predictions.

I did not limit the data supplied to the model, as Prophet is meant to be able to incorporate daily weekly, quarterly, yearly, etc. seasonality trends. I am using a linear growth type. I also am passing my list of holidays to the model to allow the model to consider these in its predictions. For each stock, I will be plotting the results of the predictions to the actual and will be updating a dataframe to output the results of each run.

Conclusion & Implementation

To implement the model on my data, I wrote a function that accepted a data, number of days to predict in the future, and what day to start with. This way, using a loop, I am able to run the function for each stock.

In the appendix, I've included the visual plot of each run for each stock. Figure 6 is the result for Discover. The first plot shows the prediction (blue line), confidence interval (blue area), and actual values (black dots). The chart below that shows the values for the future predictions.

The final output of this model is a table which is shown in figure 5 below. Each row is a given stock. The columns include the RMSE, MAPE, the actual value and the date where the last actual value was obtained, the first forecast (forecast_1), the first forecast date(forecast_1_date), the same for forecast 2 & 3, the difference between the first forecast and the last actual day. This value will indicate if the price is expected to be higher or lower. FC_2_diff and FC_3_diff do the same thing. The last two columns are meant to provide an aggregate of the 3 day forecasts expected performance. 3_day_gain_avg shows the

average of the price differences between the actual and the three predicted future values. Users can use this combined with the RMSE and MAPE to make an informed decision. For example, Citi Bank is predicted to have an average closing price increase of $25.5, but the RMSE is quite high at ~24%. Wells Fargo and DFS have the lowest RMSE and MAPE scores and they both are predicted to have an increase of over $1. They may be a safer investment.

	stock	RMSE	MAPE	actual_dt	actual	forecast_1_dt	forecast_1	forecast_2_dt	forecast_2	forecast_3_dt	forecast_3	FC_1_diff	FC_2_diff	FC_3_diff	3_day_gain	3_day_gain_avg
0	DFS	4.45558	0.03345	2022-07-26	98.36000	2022-07-27	99.97981	2022-07-28	99.68464	2022-07-29	99.55641	1.61981	1.32464	1.19641	4.14087	1.38029
0	JPM	40.69932	0.26367	2022-07-26	113.42000	2022-07-27	146.53384	2022-07-28	146.85188	2022-07-29	146.54431	33.11384	33.43188	33.12431	99.67003	33.22334
0	AXP	10.92461	0.05475	2022-07-26	149.83000	2022-07-27	148.03265	2022-07-28	148.25940	2022-07-29	148.37087	-1.79735	-1.57060	-1.45913	-4.82707	-1.60902
0	C	23.58177	0.31773	2022-07-26	51.40000	2022-07-27	74.92827	2022-07-28	74.96955	2022-07-29	74.83525	23.52827	23.56955	23.43525	70.53306	23.51102
0	BAC	6.27102	0.15691	2022-07-26	33.05000	2022-07-27	39.35414	2022-07-28	39.49575	2022-07-29	39.66155	6.30414	6.44575	6.61155	19.36144	6.45381
0	USB	10.81914	0.18643	2022-07-26	47.42000	2022-07-27	55.56950	2022-07-28	55.75201	2022-07-29	55.83640	8.14950	8.33201	8.41640	24.89791	8.29930
0	WFC	2.22804	0.03514	2022-07-26	42.90000	2022-07-27	44.39417	2022-07-28	44.51920	2022-07-29	44.61750	1.49417	1.61920	1.71750	4.83087	1.61029

Figure 5 - Final Output

Ethical Assessment

There is an ethical consideration to not necessarily tell a user to buy a certain stock. Instead, I plan to only offer the information available and let the user make the decision. There are many factors that effect the stock market and users should take as much data into account as possible.

Also, because I am using a free API, I can only offer updates on this data 5 times a day. I would need to make this clear if this were to become an application. This would be to ensure no one is making decisions using outdated information.

Sources

Li, K. (Y. (2022, May 13). *Predicting stock prices using machine learning*. neptune.ai. Retrieved July 15, 2022, from https://neptune.ai/blog/predicting-stock-prices-using-machine-learning#:~:text=Predicting%20stock%20price%20with%20Moving%20Average%20(MA)%20technique&text=Commonly%20used%20periods%20are%2020,Simple%20MA%20and%20Exponential%20MA.

Mitchell, C. (2021, September 8). *Profit without predicting the market*. Investopedia. Retrieved July 15, 2022, from https://www.investopedia.com/articles/trading/10/profit-without-predicting.asp#:~:text=Predicting%20the%20market%20is%20challenging,will%20happen%20in%20the%20future.

Welcome to discover. Discover. (n.d.). Retrieved July 26, 2022, from https://www.discover.com/company/our-company/#:~:text=Discover%20Financial%20Services%20becomes%20an,first%20time%20on%20July%202.

Appendix

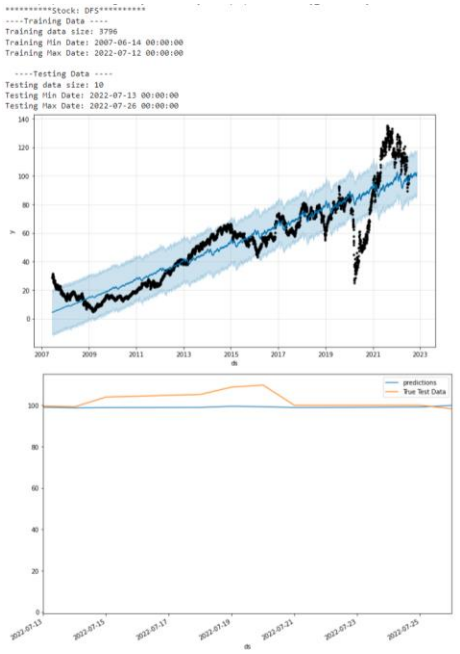

Figure 6 - DFS

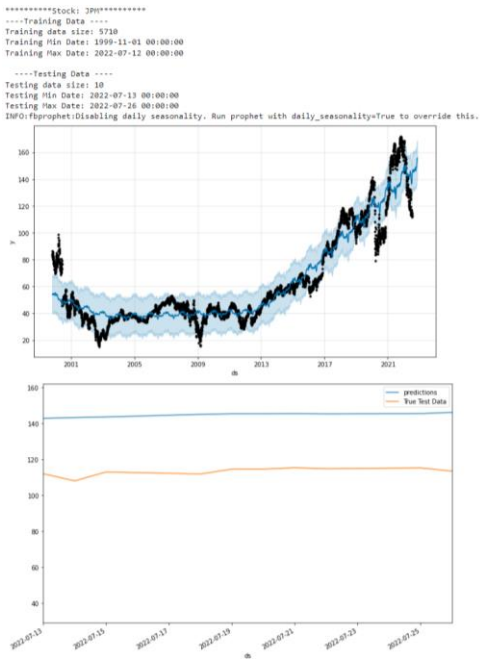

Figure 7 – JPM

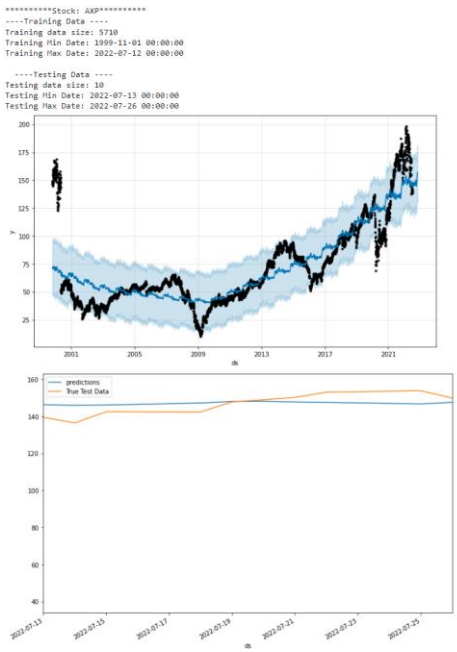

Figure 8 - American Express

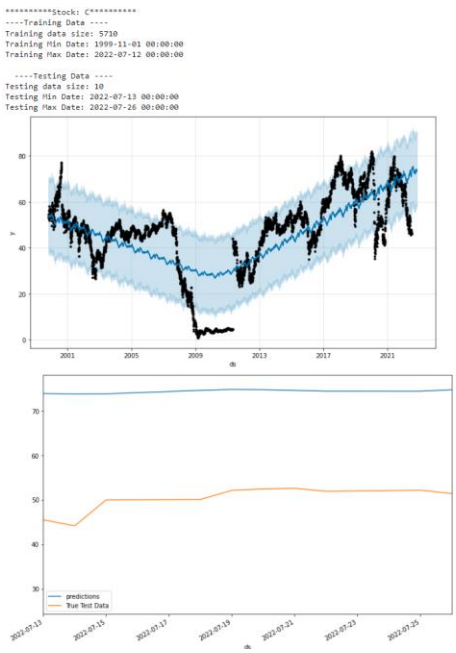

Figure 9 - Citi Bank

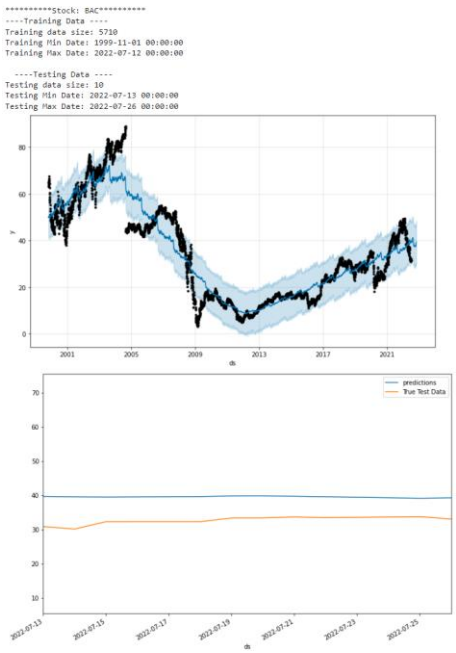

Figure 10 - Bank of America

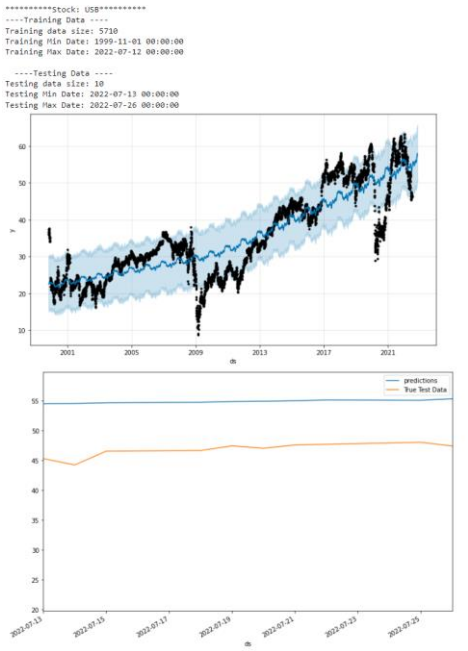

Figure 11 - US Bank

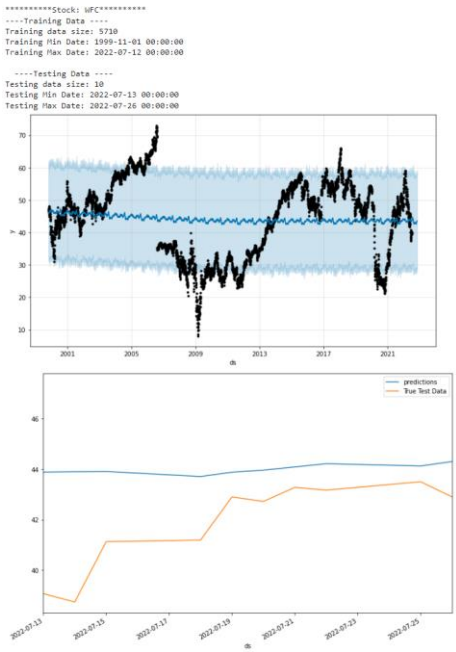

Figure 12 - Wells Fargo