MLTSA midterm: project proposal

Don't modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

Using a Hybrid LSTM-ARIMA Model to Forecast NDXT-100 Performance

Yuca Chen, github.com/yucachen, yucachen@udel.edu, Communication and Analysis Manager

Jacklyn Clauss, github.com/jclauss24, jclauss@udel.edu, Data Manager

Skyler Lindsey, github.com/skyler-ud, skyel@udel.edu, Literature and Analysis Manager

Dayanara Yanez, github.com/dayaYanez, dayalyz@udel.edu, Visualization Manager

Abstract: Much work has been done on using neural networks to predict stock data, and many models fail on short-term predictions. We plan to predict the NASDAQ-100 technology sector index, along with using Apple historic stock data, Google historic stock data, and Consumer Price Index data. The most successful models in past studies have been using a Long-Short Term Memory (LSTM) structure. To improve upon this model and to account for seasonality and periodicity in stock price, we plan to integrate an Auto Regressive Integrative Moving Average step into a LSTM model.

Introduction:

The average person's interest in the stock market has grown immensely over the last few decades [1]. Many people, outside of Wall Street, are now trying their hand in investing in the stock market in an attempt to take advantage of this avenue of making money – and who doesn't want to make money? If the behavior of the market can be forecasted accurately, then a more consistent profit can be made. This want of accurate forecasting motivates the use of machine learning in order to do these predictions [2].

Among the numerous methods that have been used for this forecasting, prominent algorithms that have been widely used in recent studies include artificial neural networks (ANN), support vector machines (SVM), and long short term memory recurrent neural networks (LSTM RNN) [3,4]. A range of different datasets have also been called upon to feed into these algorithms, including time-series data of available stock price information and macroeconomic information [2].

Data:

The primary data set being used in this project will be historical data on the NASDAQ-100 technology sector, also called NDXT-100. The NDXT-100 is an equal weighted

MLTSA midterm: project proposal

Don't modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

index, comprised of securities that are considered to be technological. The data columns of this data set include the value of the index, along with the high price, low price, and the net change, and is available in a daily time interval.

As supplemental data sets, we will also be using Apple and Google historical stock data. The columns of this data set include daily opening and closing prices, as well as the high and low prices over time. Consumer price index (CPI) data will also be used as another supplemental data set to bring in macroeconomic information. This data is an index of the price paid by consumers for all items and is an average over all United States cities. CPI can be used as a measure of the inflation rate over time. This data is available at monthly time intervals.

Dataset name	**URL**	**Number of rows**	**Number of columns**	**Number of relevant columns**	**Number of valid rows (not NaN on relevant columns)**	**Data type for each relevant column**
NDXT-100	Link to Data source Link to csv file in github	4803	5	5	4802	Date, Float
AAPL historic stock data	Link to data source Link to csv file on github	2517	6	6	2517	Date, Float
GOOGL historic stock data	Link to data source Link to csv file on github	2517	6	6	2517	Date, Float
US Consumer Price Index (CPI) Historic data*	Link to data source Link to csv file on github	37	13	13	37	Float

MLTSA midterm: project proposal

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

*Note that this data set is organized with the year as the row and the month as the column. Because of this, the 2025 row is “missing” data for March onward, as that data is not yet available.

```
merged_all_info.describe()
```

	CPI	IndexValue_NDXT	NetChange_NDXT	High_NDXT	Low_NDXT	Close/Last_GOOGL	Volume_GOOGL	Open_GOOGL	High_GOOGL	Low_GOOGL	Close/Last_AAPL	Volume_AAPL	Open_AAPL	High_AAPL	Low_AAPL
count	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	8.000000e+01	80.000000	80.000000	80.000000	80.000000	8.000000e+01	80.000000	80.000000	80.000000
mean	268.039275	5736.497621	9.292133	5790.580089	5675.568981	83.955922	3.695266e+07	97.739190	84.843744	83.043946	98.786505	1.225911e+08	97.739190	98.786505	96.694474
std	26.539729	2638.365152	109.248208	2670.481158	2608.219285	43.824001	1.473377e+07	64.850660	44.352052	43.340103	65.470142	7.056292e+07	64.850660	65.470142	64.104162
min	236.525000	2102.664689	-424.604916	2137.083486	2090.910392	27.165000	1.600613e+07	23.872500	27.290500	26.988000	24.116300	3.528143e+07	23.872500	24.116300	23.832500
25%	245.378000	3536.448809	-27.421121	3537.052936	3513.920414	47.531625	2.698290e+07	38.038125	47.868500	47.447500	38.138125	7.120070e+07	38.038125	38.138125	37.816875
50%	257.277000	5008.726820	10.142135	5066.038761	4969.873610	64.739250	3.302262e+07	66.973750	65.610750	64.234250	69.366250	1.064704e+08	66.973750	69.366250	66.876250
75%	296.284750	8108.494297	50.478459	8150.127249	7988.130917	122.763375	4.089553e+07	155.470000	123.045250	121.803625	156.192500	1.521625e+08	155.470000	156.192500	152.922500
max	315.664000	10832.367099	274.786761	10839.929986	10662.366556	182.990000	1.090090e+08	229.520000	183.880000	181.300000	229.650000	4.473492e+08	229.520000	229.650000	223.740000

Figure 1: A snapshot of the pandas command “df.describe()” on the main dataframe, “merged_all_info”, that combines all the data listed in the table above under an index of datetime. The total count of data points, mean, standard deviation, minimum value, 25th percentile, 50th percentile, 75th percentile, and maximum value for each data column in “merged_all_info” is printed.

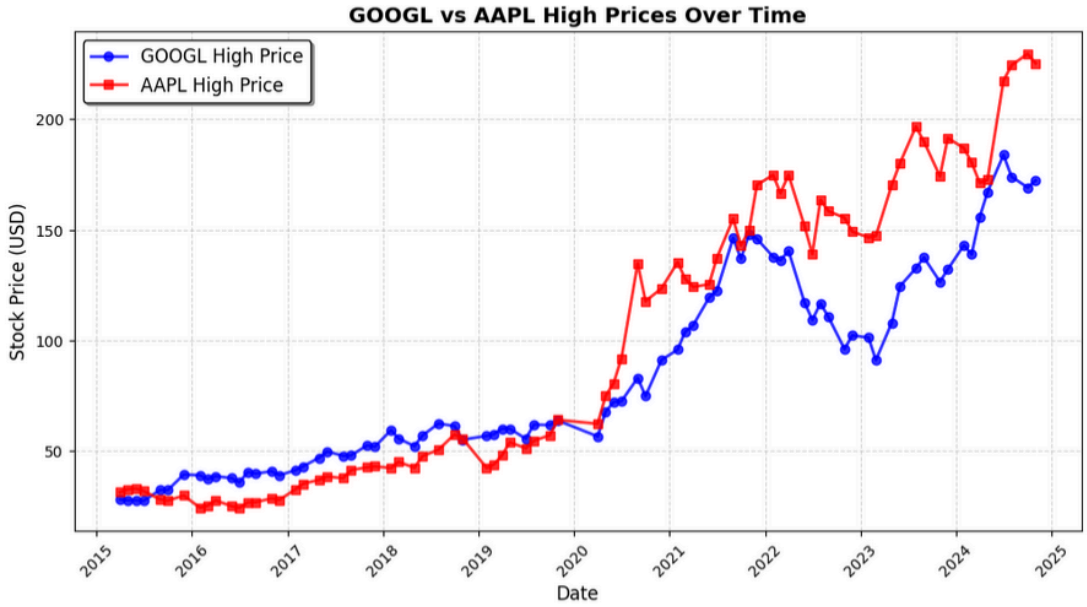

Figure 2: High stock prices of AAPL and GOOGL from 2015 to 2025. Although there are some noticeable swings, both show an overall rising trend: AAPL outperforms GOOGL in 2019–2021, is more volatile in 2022–2023, and continues to trend upward through 2024–2025.

MLTSA midterm: project proposal

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

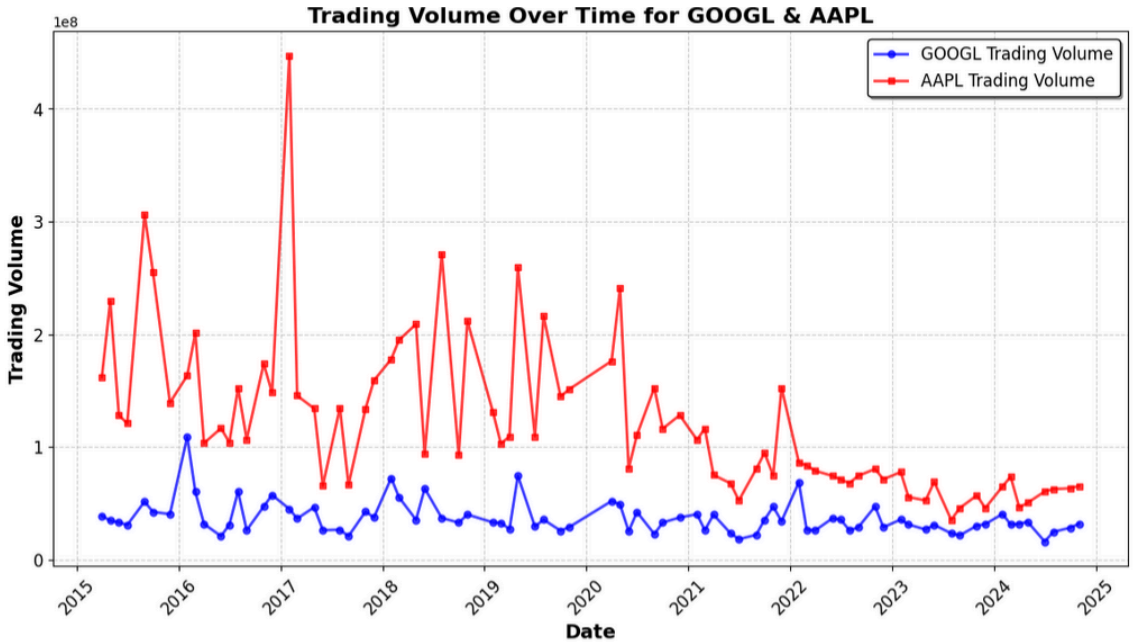

Figure 3: Trading volumes of GOOGL (blue) and AAPL (red) from 2015 to 2025. While GOOGL's trading volume remains relatively stable with minor fluctuations, AAPL exhibits higher volatility with notable spikes around 2017 and 2020. The visualization underscores differences in market activity, showing that AAPL generally experiences higher and more volatile trading volumes than GOOGL.

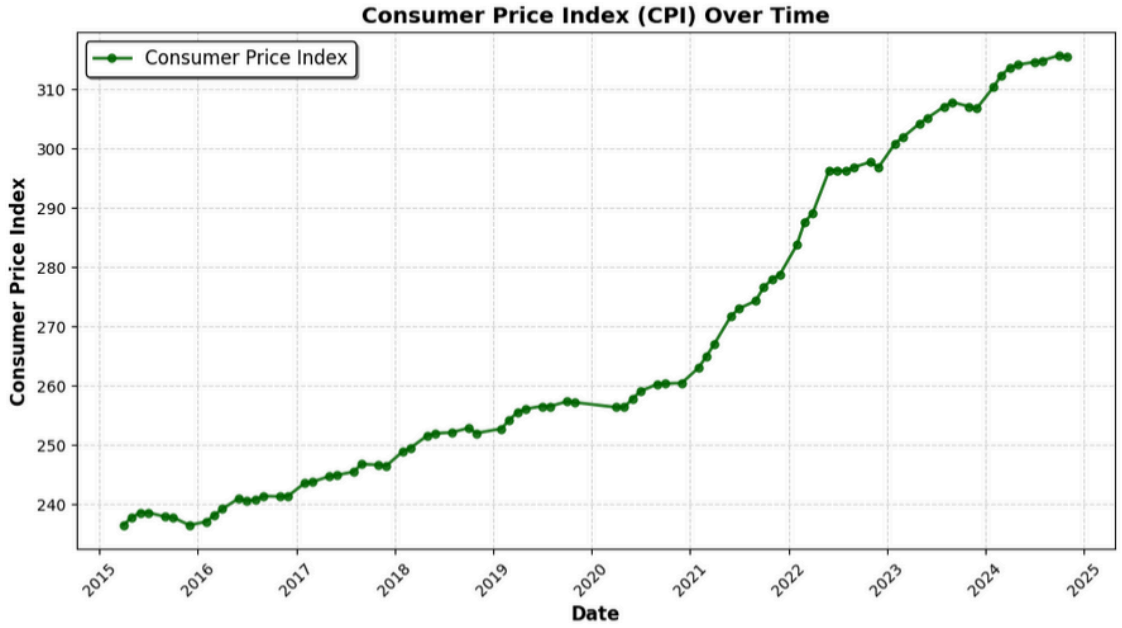

Figure 4: The Consumer Price Index (CPI) from 2015 to 2025. The Consumer Price Index (CPI) trend from 2015 to 2025 is displayed in this line plot, which illustrates how the price of goods

and services has changed over time. The years are shown on the x-axis, while the CPI values are shown on the y-axis. A time of growing inflation is shown by the data's consistent upward trend, which noticeably accelerates from 2021 onward.

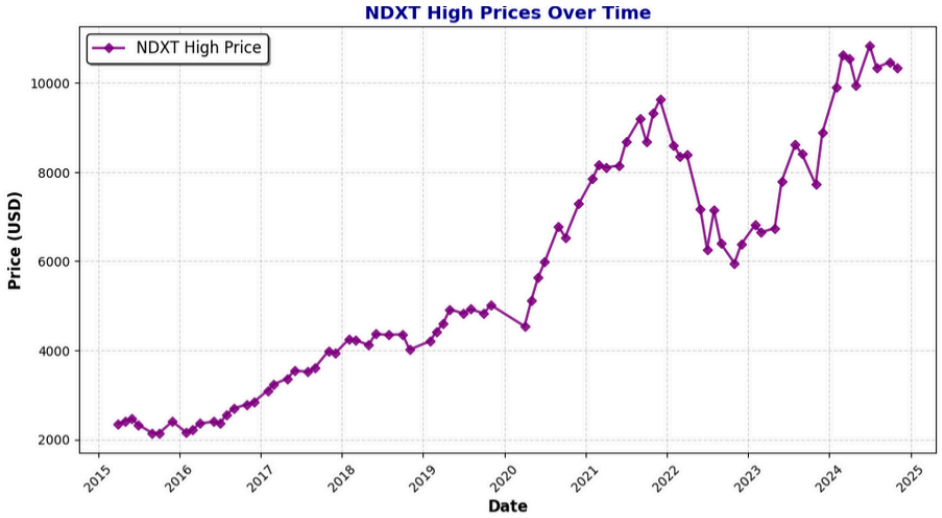

Figure 5: NDXT Prices from 2015 to 2025. The NASDAQ Tech Index's (NDXT) rising prices throughout time are depicted in this line graphic. The years 2015–2025 are represented by the x-axis, and the index price in USD is represented by the y-axis. The variations in the high price of NDXT are depicted by the purple line with diamond markings.

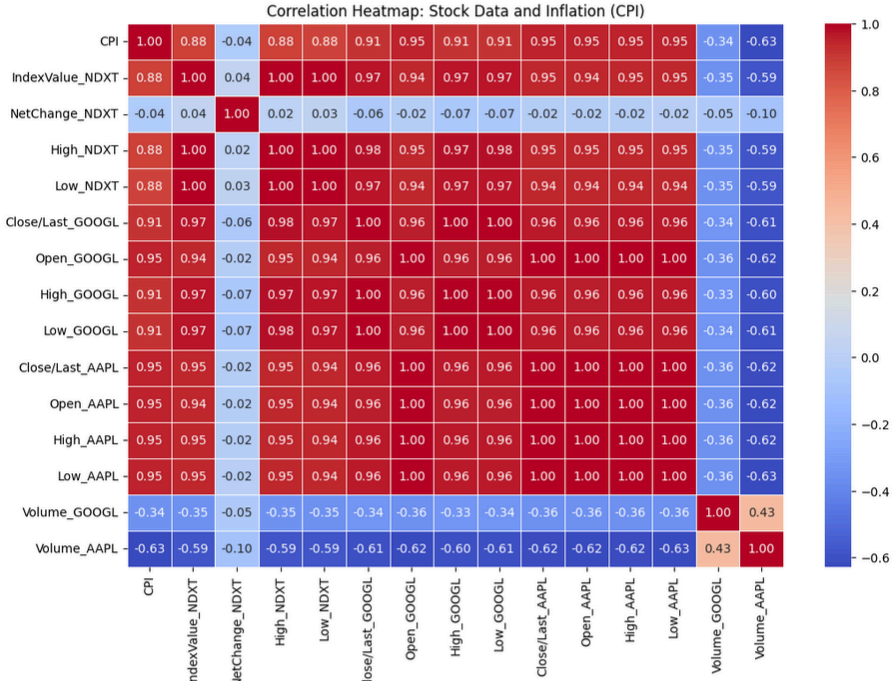

Figure 6: Correlation coefficients between trade volumes, the NASDAQ Tech Index (NDXT),

stock prices (GOOGL & AAPL), and the Consumer Price Index (CPI). Stronger correlations are shown by darker hues on the color scale, which goes from -1 (blue, negative correlation) to +1 (red, strong positive correlation). The CPI and the stock prices of GOOGL and AAPL show a strong positive correlation (>0.90), indicating that changes in inflation are in line with changes in stock prices. Higher trading volumes frequently accompany price drops, as evidenced by the negative correlations between trading volumes (GOOGL & AAPL) and stock prices.

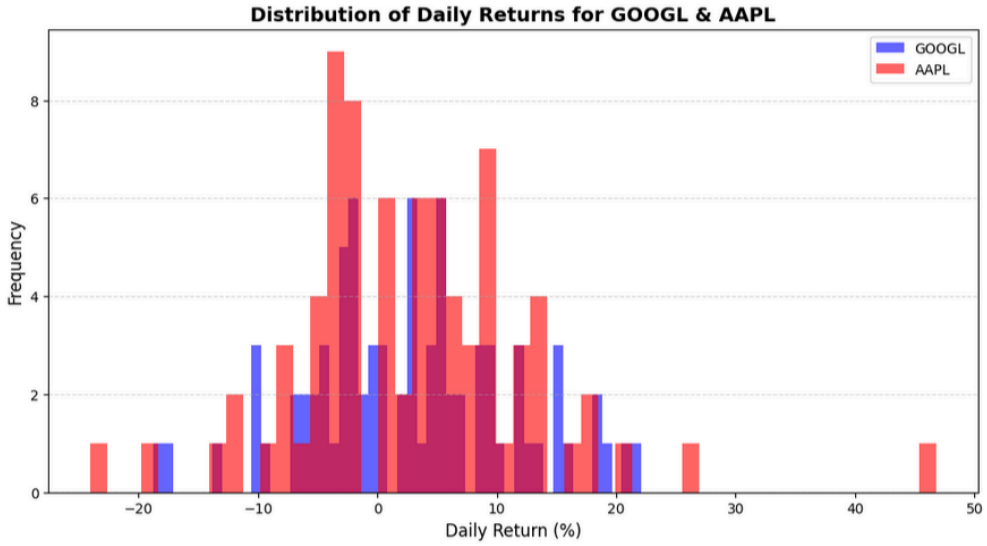

Figure 7: GOOGL & AAPL's Daily Return Distribution. The distribution of daily returns (%) for AAPL (red) and GOOGL (blue) over a specified time period is shown in this histogram. The bell-shaped distribution of both stocks indicates that the majority of daily returns are grouped around 0%, with sporadic significant positive and negative fluctuations.

Methodology:

This project will attempt to model the high price behavior of NDXT-100 using Apple and Google stock data as supplemental information, as well as consumer price index data. We plan to utilize a long short term memory model (LSTM) along with a autoregressive integrative moving average model (ARIMA) in order to accomplish this task. LSTM is a kind of recurrent neural network which is designed especially for use with time series data to model its non-linearity. In particular, these neural networks are useful at recognizing long-term dependencies within time series data. ARIMA is a type of model that is useful in recognizing non-stationary time series data.

Combining the two of these together, the predictions made from LSTM can be corrected using ARIMA’s error correction, along with capturing both local non-linear and overall linear trends. Thus, having these two together can help make the model more robust against the volatility associated with stock and financial series data, lending to a higher chance of being able to capture and predict the behavior of said data more accurately. An example of this combination of models has been studied in Kashif and Ślepaczuk 2024, where this combined LSTM-ARIMA model was tested on the three indices of S&P 500, FTSE 100, and CAC 40 [5].

MLTSA midterm: project proposal

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

Deliverable:

Our Deliverable will be a conclusion on whether these supplemental data sources are a significant predictor of the performance of the NDXT-100, an analysis of the impact of the significance of each supplemental dataset on the main dataset, as well as a projection of our model into the future in order to attempt to predict the performance of the NDXT-100. We will determine a reasonable time frame for our prediction based on the performance of our model against test sets at different intervals of time. We will then give a final prediction of the performance of our main dataset in the best appropriate time interval.

Link to GitHub repo: The GitHub repository where this project will be developed can be found at: https://github.com/jclauss24/MTLSA_Project/tree/main.

MLTSA midterm: project proposal

Don’t modify above this line. Everything in blue below this line has to be updated. Everything in gray should be read carefully, followed carefully, and removed from the template

---

Bibliography:

[1] Badolia, Lokesh. *How can i get started investing in the stock market.* Educreation Publishing, 2016.

[2] Kumbure, Mahinda Mailagaha, Christoph Lohrmann, Pasi Luukka, and Jari Porras. "Machine learning techniques and data for stock market forecasting: A literature review." *Expert Systems with Applications* 197 (2022): 116659.

[3] Mintarya, Latrisha N., Jeta NM Halim, Callista Angie, Said Achmad, and Aditya Kurniawan. "Machine learning approaches in stock market prediction: A systematic literature review." *Procedia Computer Science* 216 (2023): 96-102.

[4] Liu, Yang. "Novel volatility forecasting using deep learning–long short term memory recurrent neural networks." *Expert Systems with Applications* 132 (2019): 99-109.

[5] Kashif, Kamil, and Robert Ślepaczuk. "Lstm-arima as a hybrid approach in algorithmic investment strategies." arXiv preprint arXiv:2406.18206 (2024).