Predicting NASDAQ

Predicting NASDAQ stock market using base Economic Indicators, with ML Models

Submitted by:
Ohad Lavon Ben Moshe
Yehonatan Cohen
Gilad Erlichman

26/02/2023

Motivation

- Predicting stock market trends is crucial for investors to make efficient decisions.

The stock market is a vital component of the global economy, and its fluctuations can significantly impact individuals and businesses alike. Therefore, accurate predictions of market trends can be valuable for investors, businesses, and policymakers.

Our project aims to contribute to this field by using machine learning techniques to analyze and predict stock market trends based on various economic indicators. Through this project, we hope to provide insights and tools that can help individuals and businesses make informed decisions about their investments and financial strategies.

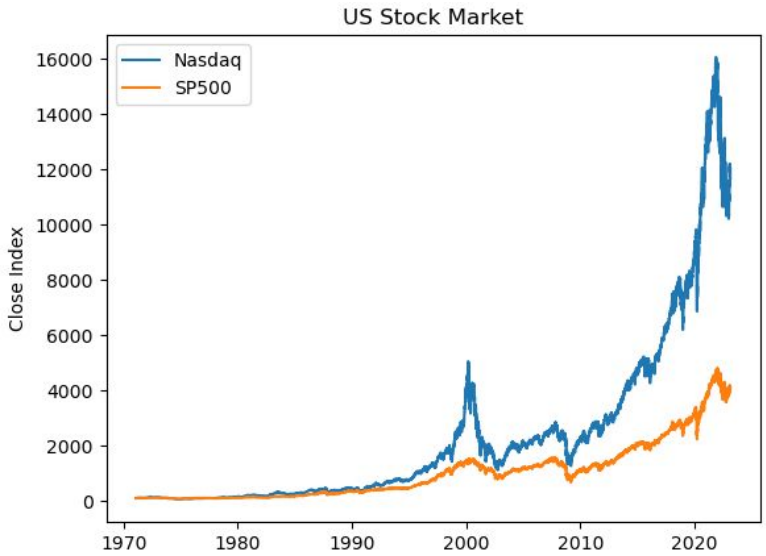
Project Objective

- The stock market is a complex and ever-changing system that is affected by a variety of factors, including economic indicators such as CPI, VIX, Interest, Governmental Bonds and other Macro Indices such as employment rate etc .
- The objective of this research project is to determine whether it is possible to predict NASDAQ value based on historical data and the values of these economic indicators.
- The research question is:
    - Can we accurately predict NASDAQ value using economic indicators.

Dataset

- The dataset will include historical values for the NASDAQ stock indices, as well as economic indicators such as CPI, VIX, and Bonds.
- We obtained the economic indicators data from two sources:
    - Yahoo Finance API via yfinance Python library.
    - Federal Reserve Economic Data (FRED) API via fredapi Python library.
- The following economic indicators were evaluated:
    - Closing data and turnover of Nasdaq
    - Closing data and transaction turnover of 7-year and 20-year government bonds
    - VIX fear index
    - Interest rates on loans, short-term deposits, and long-term deposits
    - Inflation forecast for the year, 3 years, and 5 years
    - Day of the week and holidays in the US.
    - unemployment rate

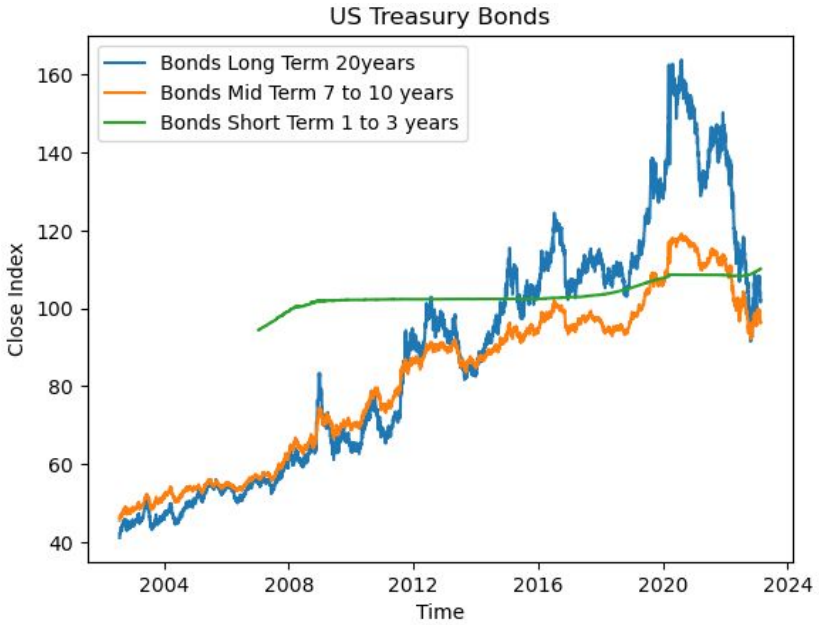
Data Explanation - Nasdaq



We choose to focus on Nasdaq over S&P 500

- **Industry Focus**: Nasdaq is known for being the go-to exchange for technology companies, while S&P 500 is a broader index that covers a range of industries.
- **Growth Potential**: Historically, Nasdaq has tended to outperform S&P 500 in terms of growth..
- **Volatility**: Because Nasdaq is more focused on technology companies, it may be more volatile than S&P 500.

Data Explanation - 20-year and 7-year government bonds



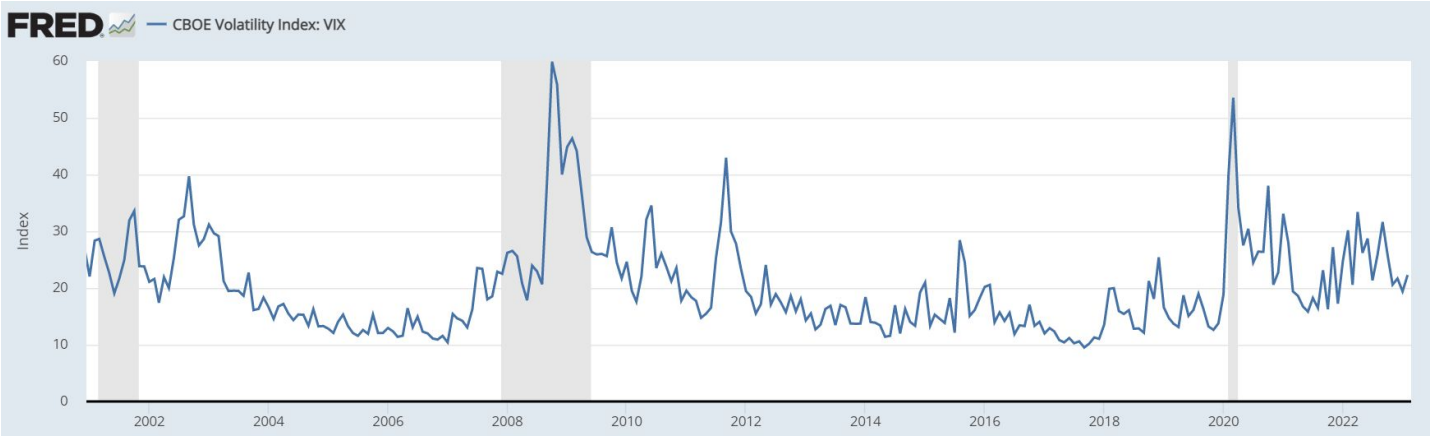
We choose to included 7-year and 20-year government bonds as predictors

- The bond market is closely related to the stock market, and changes in bond prices and yields can impact the stock market.
- We selected government bonds as one of the predictors because they are considered to be the safest and most reliable investment options and are used as a benchmark to evaluate other investments, including the stock market.
- Among various types of government bonds, we chose the 7-year and 20-year bonds because they are considered to be medium to long-term bonds and have a higher sensitivity to interest rate changes.
- The yield on a bond is inversely related to its price, and higher bond yields can indicate a stronger economy and the potential for higher interest rates. This can lead to a decrease in stock prices as investors move away from stocks to invest in bonds with higher yields.
- We hypothesized that changes in the yields of 7-year and 20-year bonds could indicate a potential shift in investor sentiment towards the stock market, and therefore, they could be used as predictors for the NASDAQ index price.

To test this hypothesis, we included the yields of the 7-year and 20-year government bonds as predictors in our machine learning model, and our results showed that these predictors were statistically significant in predicting the NASDAQ index price.

Data Explanation - VIX fear index

- We chose to add the VIX index to the model because it is commonly known as the "fear index" and is used as a measure of investor sentiment and market volatility. When the VIX index is high, it indicates that investors are fearful and uncertain about the market, which may lead to decreased trading volume and lower stock prices. Therefore, including the VIX index in our model can help us to better understand the relationship between market sentiment and stock prices and improve the accuracy of our predictions.



Data Explanation - Day of the week and Holidays

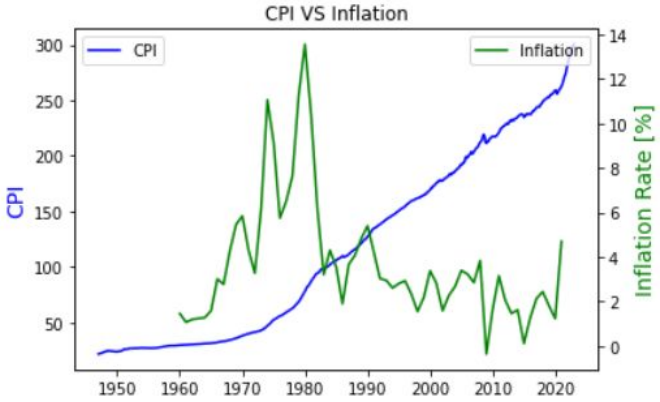
- The day of the week and holidays in the US calendar as variables in our model because they may have an impact on stock prices and trading volume. For example, some studies have shown that stock prices tend to be higher on Fridays compared to other weekdays, while others have found that stock prices tend to be lower on Mondays. Additionally, holidays may affect trading volume because the stock market may be closed on those days or investors may be less active due to the holiday. By including these variables in our model, we can account for these factors and potentially improve the accuracy of our predictions.
- We notice the lack of correlation between the week day and other parameters and decided not to include it in the final models.
- Holiday days which has no trades were not included in the final model.

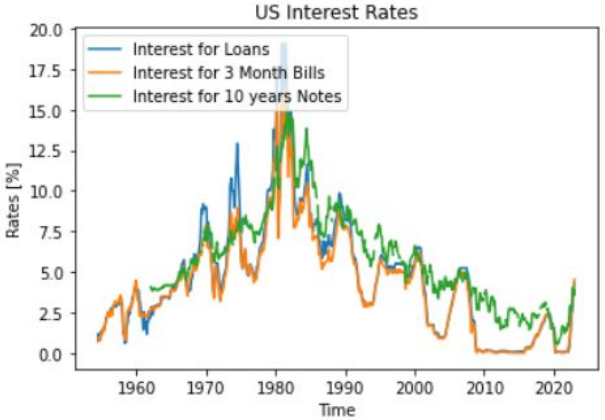
Data Explanation - CPI, inflation, interest rates, and unemployment rates

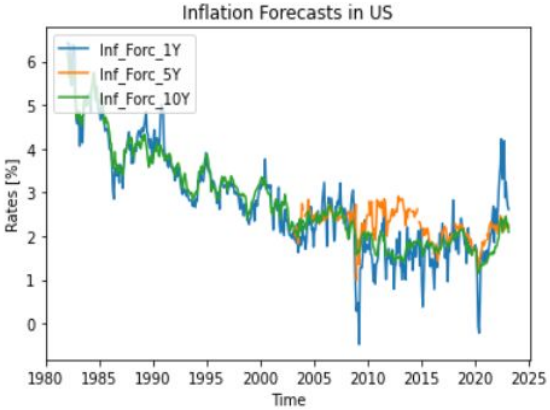
- We chose not to use CPI, inflation, interest rates, and unemployment rates in our model mainly because the data for these indicators is updated on a monthly basis, while we are basing our analysis on daily data. Additionally, these indicators may not have a direct impact on the stock market, and their relationship with the market may be more complex and indirect. For the sake of simplicity and relevance, we focused on using indicators that are updated on a daily basis and have a more direct relationship with the stock market, such as government bond yields and the VIX index.

```
Int64Index: 5179 entries, 7948 to 13126
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Date                5179 non-null   datetime64[ns, America/New_York]
 1   WeekDay             5179 non-null   int64
 2   is_holiday          5179 non-null   bool
 3   Close_Nasdaq        5179 non-null   float64
 4   Volume_Nasdaq       5179 non-null   int64
 5   Close_Bonds_20more  5179 non-null   float64
 6   Close_Bonds_7to10   5179 non-null   float64
 7   VIX                 5179 non-null   float64
```

Data Explanation - CPI, inflation, interest rates, and unemployment rates

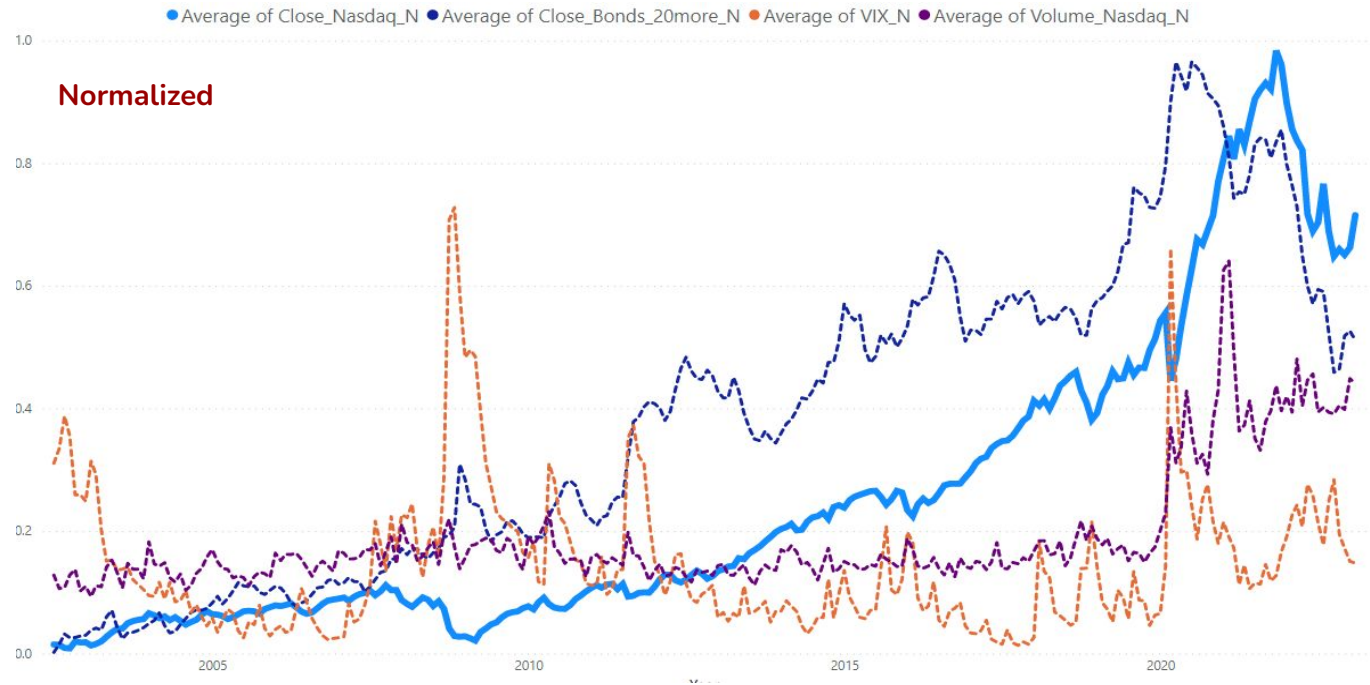






Graph plots of the CPI, inflation, interest rates, and unemployment rates

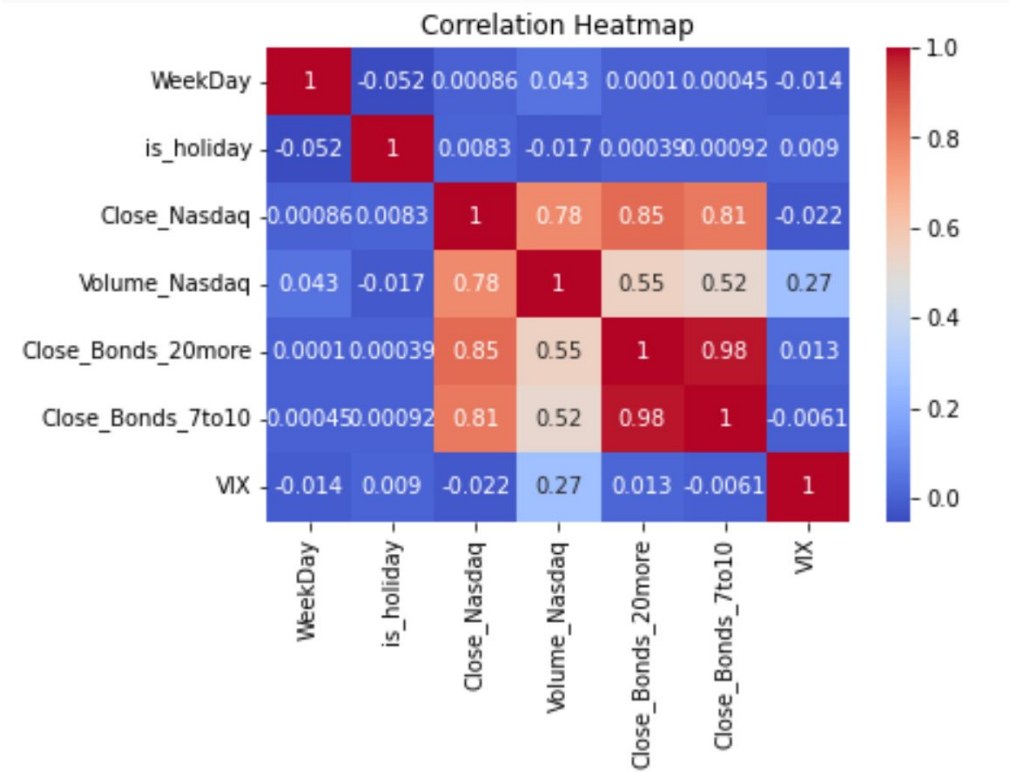
Summary Statistics



- We normalized the data to plot it on the same graph
- For example, here we can see the market response to the volatility index (VIX) and the lack of correlation of the market response to the volume

*Done with PBI

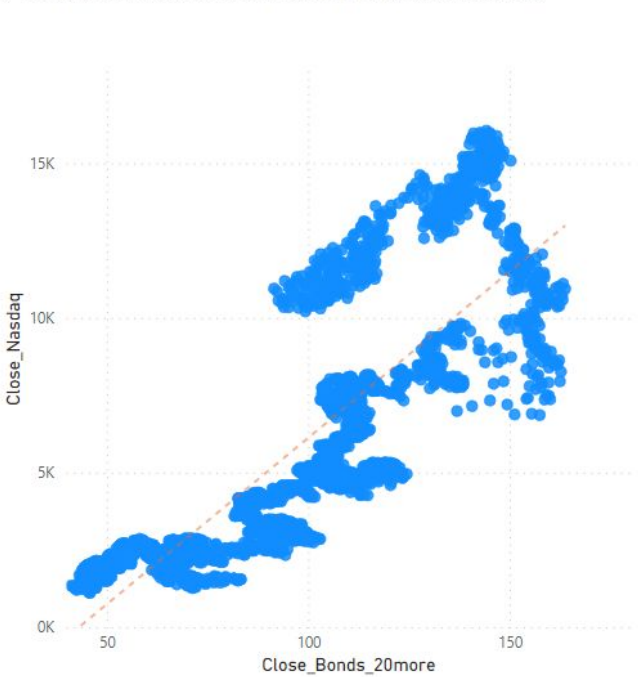
Exploratory Data Analysis



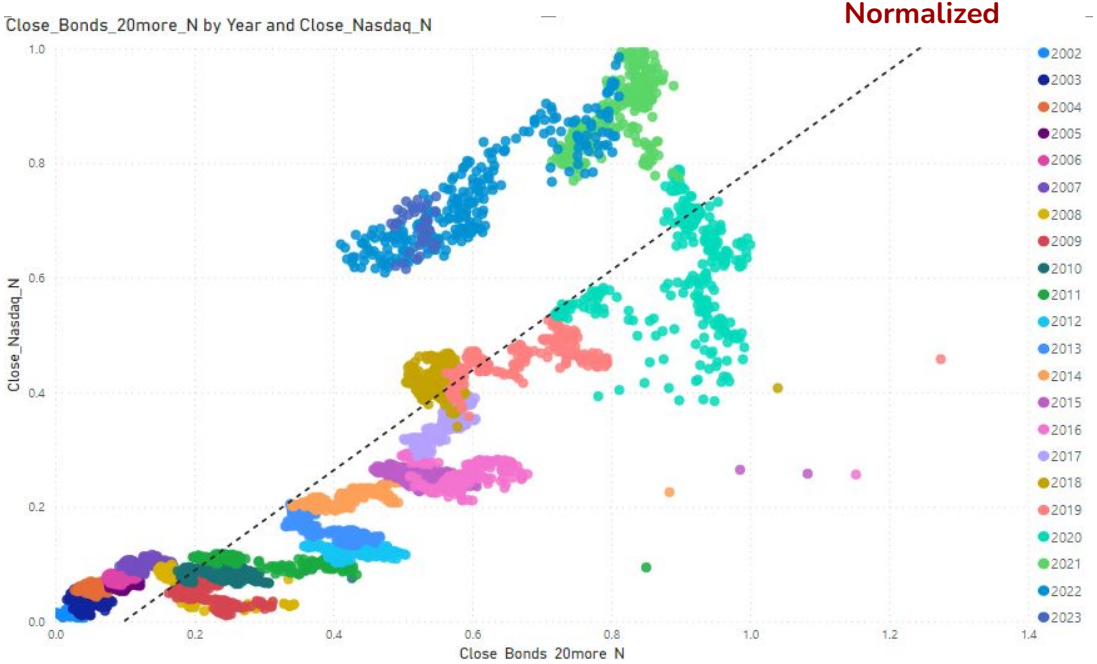
- we can see that there is a strong positive correlation between Close_Nasdaq and Volume_Nasdaq (0.776), indicating that as the closing price of Nasdaq increases, the trading volume tends to increase as well.
- We also see that there is a strong positive correlation between Close_Bonds_20more and Close_Bonds_7to10 (0.982), which is not surprising since both of these variables represent bond prices with similar maturity dates.
- Interestingly, we can see that there is a weak negative correlation between WeekDay and is_holiday (-0.052), which suggests that there may be slightly fewer holidays during the workweek compared to weekends.
- Overall, understanding the correlations between variables can help us identify which variables are most relevant to our analysis and which ones may have redundant information.

Exploratory Data Analysis

On average When bonds increases nasdaq also increases.



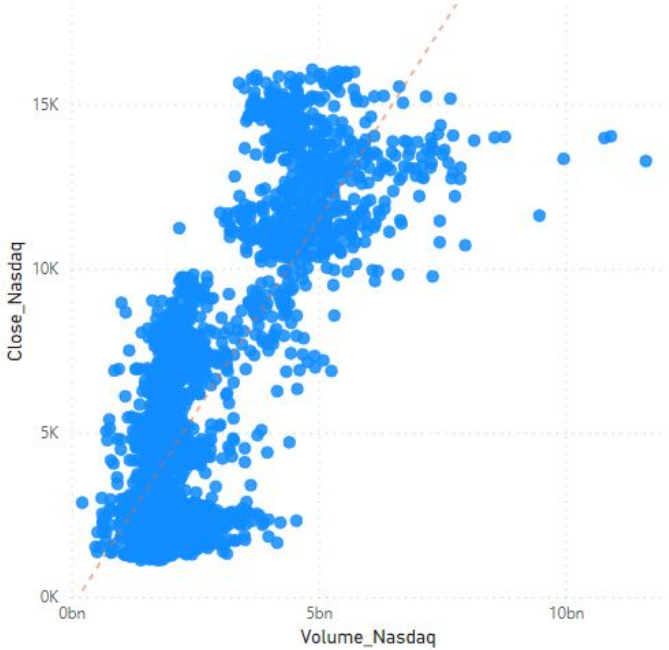
Normalized



*Done with PBI

Exploratory Data Analysis

On average when volume increases nasdaq also increases

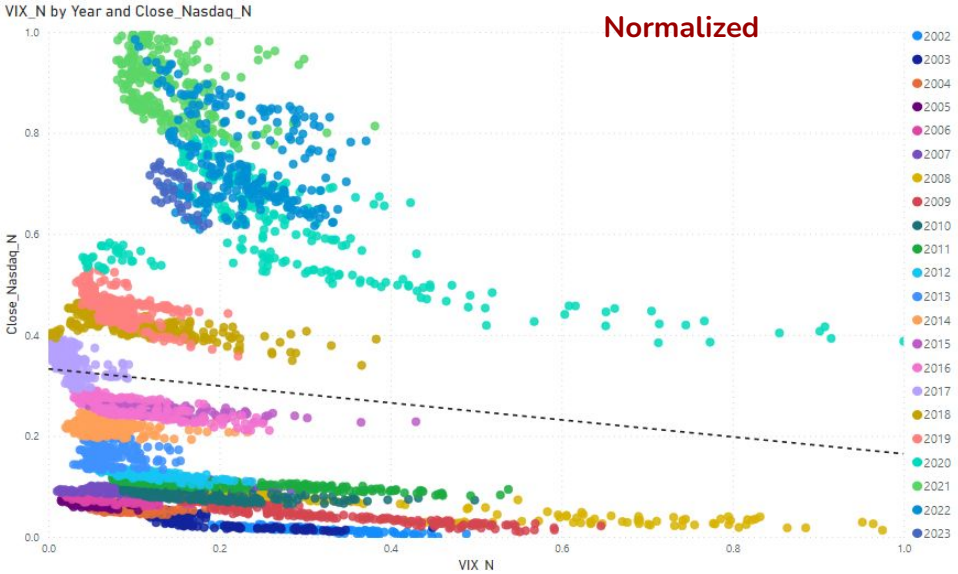


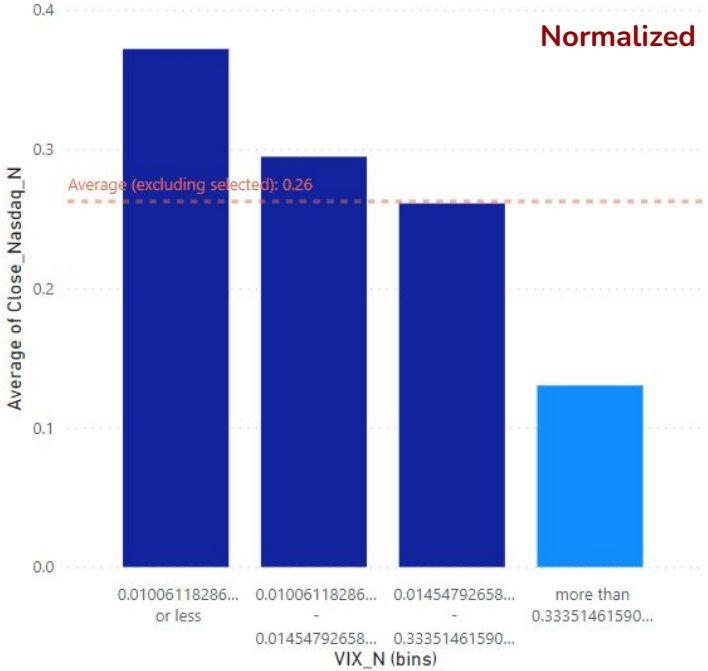
Normalized

*Done with PBI

Exploratory Data Analysis

Nasdaq is more likely to decrease when VIX is more than 0.33 (normal) (on average).



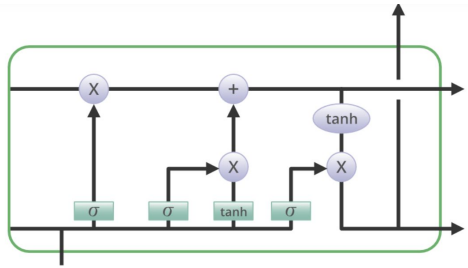


*Done with PBI

Modeling

We decided to examine models that expertise in series models:

- **Moving Average** - Simple method, easy to perform, and a common analyzing tool among traders and investors. However, this method relates to the history of the data, and ignores other variables that may affect that data.
- **Random Forest** - This model can work with time series very well. Regression, k-nearest neighbors and other models that are based on distance measurements, enforce Normalization of Data. Unlike those Models, the Random Forest is an ensemble learning algorithm that operates by constructing multiple decision trees and combining their predictions to produce a final output. Hence, **normalization is not required.**
- **Long Short Term Memory (LSTM)** - is a type of Recurrent Neural Network (RNN) that is specifically designed to handle sequential data, such as time series, speech, and text



Modeling - Data Set and Data Preparation

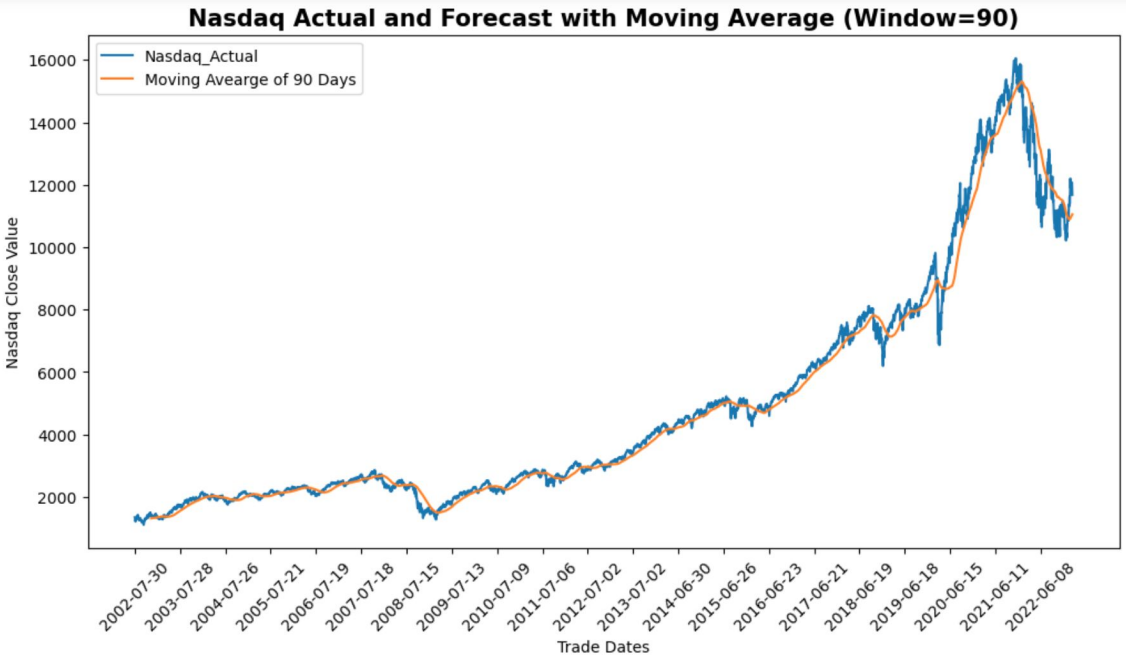
	Date	WeekDay	is_holiday	Close_Nasdaq	Volume_Nasdaq	Close_Bonds_20more	Close_Bonds_7to10	VIX
0	2002-07-30	1	0	1344.189941	1728270000	41.130577	45.574963	31.92
1	2002-07-31	2	0	1328.260010	1633300000	41.640152	45.992958	32.03
2	2002-08-01	3	0	1280.000000	1548860000	41.877327	46.182476	36.95
3	2002-08-02	4	0	1247.920044	1419790000	42.306160	46.539185	41.29
4	2002-08-05	0	0	1206.010010	1336720000	42.492863	46.773273	45.08

```
1 Consolidated_Data.shape
```

```
(5176, 8)
```

- **WeekDay** - Since the market share is in USA, then Monday = 0, Tuesday = 3, etc. This variable was converted to Dummies.
- **Is_holiday -** Indicator for holidays according to holidays in US. Usually the stock market is idle during holidays, abd in our Data Set only 46 occurrences found. Henc, this column was omitted.
- **Rest of the columns** - the figures are continues and Normalized with SKLEARN **MinMaxScaler.**

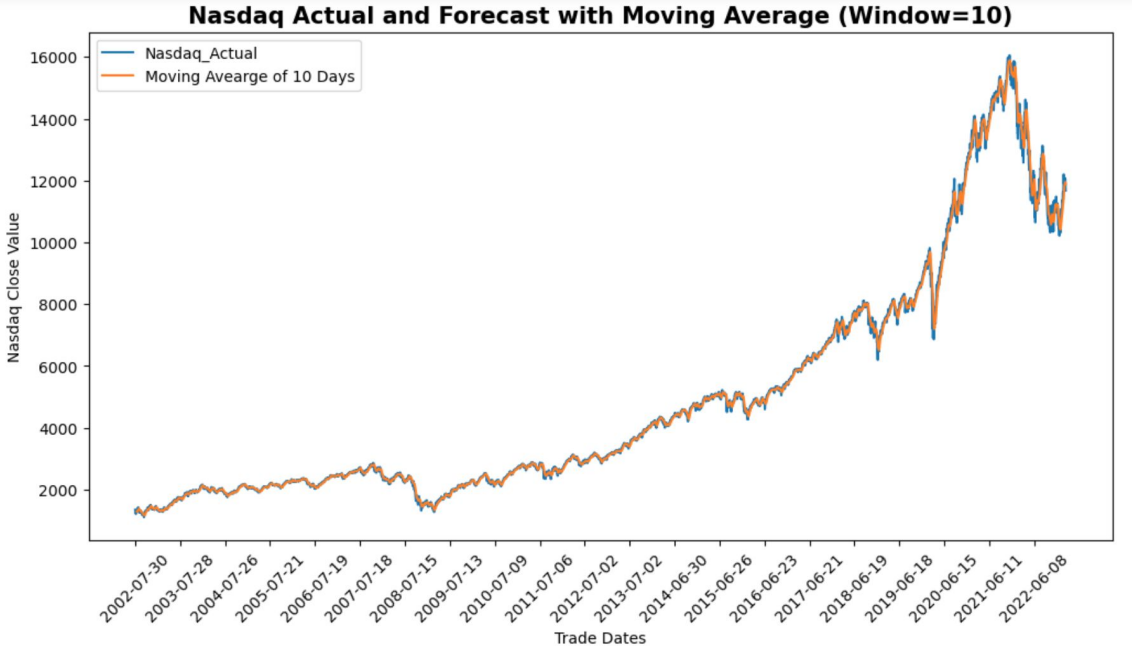
Modeling - Moving Average




```
SQRT MSE (Window=90 days):                295.51
MSE is/Nasdaq average closing price         6.02 %
Pearson Correlation:                        0.99
```

Will smaller Window improve accuracy?

Modeling - Moving Average



SQRT MSE (Window=10 days):	91.87
MSE is/Nasdaq average closing price	1.87 %
Pearson Correlation:	1.00

- Accuracy increases as ‘Window’ is narrower.
- Model is based on history only. Dramatic and rapid changes will not be reflected on time.
- **Will explanatory data predict rapid changes on time?**

Modeling - Random Forest

Data Preparation:

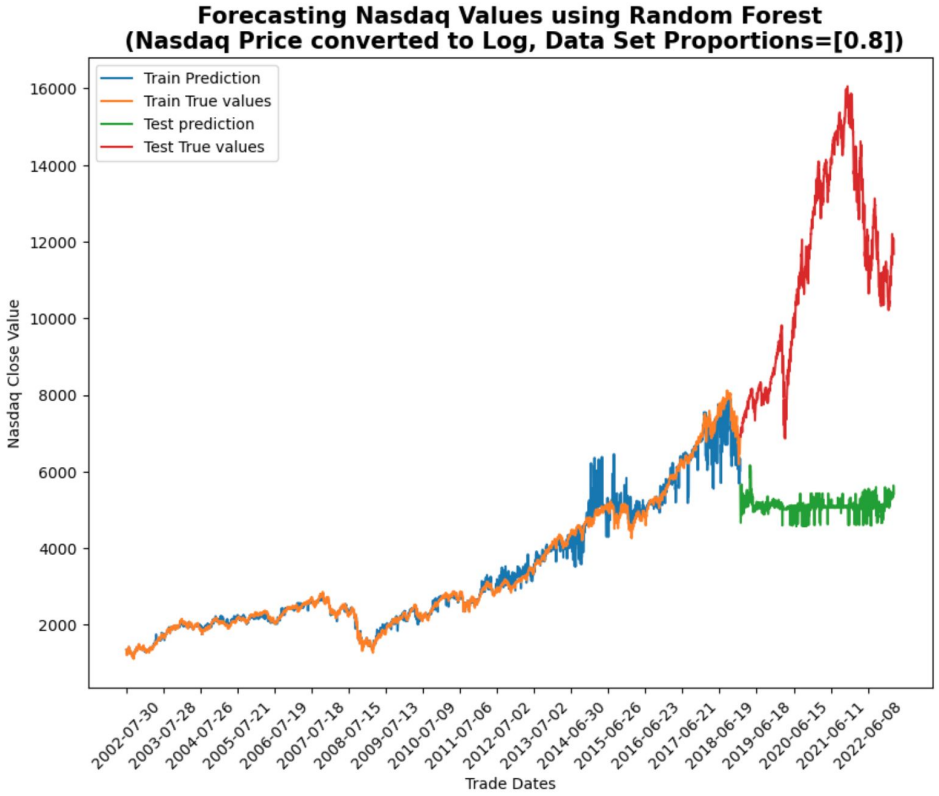
- **Splitting** - Train - 80% , Test-20%
- **Normalization -** Not required
- Nasdaq Prices converted to Log[Nasdaq Prices]
- **Modeling:**

```
from sklearn.ensemble import RandomForestRegressor
Random_Forest = RandomForestRegressor(n_estimators=500, random_state=42, min_samples_split=2,
                                      min_samples_leaf=1, max_depth=10, bootstrap=True)
Random_Forest.fit(X_train, y_train.Close_Nasdaq_Log)
```

```
RandomForestRegressor(max_depth=10, n_estimators=500, random_state=42)
```

- **After Model Run:** Prediction converted back to Stock Prices with Anti-log.

Modeling - Random Forest



Train Results:

Mean Absolute Error:	100.76	**Percent of MAE from Mean:**	3.03 %
Root Mean Squared Error:	200.08	**Percent of RMSE from Mean:**	6.01 %
Pearson Correlation:	0.9932		
Random Forest Score for Test:	0.9923		

Test Results

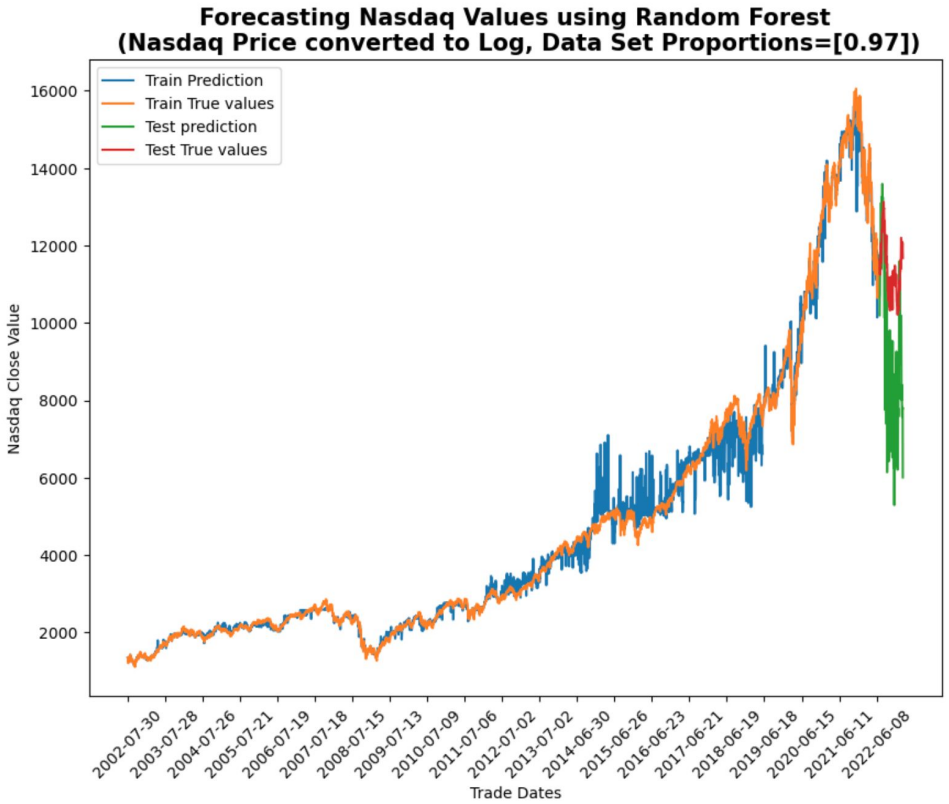
Mean Absolute Error:	6,113.41	**Percent of MAE from Mean:**	54.53 %
Root Mean Squared Error:	6,641.35	**Percent of RMSE from Mean:**	59.24 %
Pearson Correlation:	-0.0514		
Random Forest Score for Train:	-10.3422		

Model Manages to predict the Train Set (Where the Dependent variable is available), however, when predicting the Test Set it seems that it start with the recent trend found in the test set, and then, simply run an average.

There is a substantial **Over Fitting**

Question: Will smaller Test Set can improve Prediction?

Modeling - Random Forest



It seems, that just like moving average, when prediction is executed for short term, the results improved. However one can see that there is huge lag between the actual and predicted prices - **Over Fitting** still exist.

Train Results

Mean Absolute Error:	151.66	**Percent of MAE from Mean:**	3.22 %
Root Mean Squared Error:	282.66	**Percent of RMSE from Mean:**	6.01 %
Pearson Correlation:	0.9969		
Random Forest Score for Test:	0.9943		

Test Results

Mean Absolute Error:	2,354.69	**Percent of MAE from Mean:**	20.69 %
Root Mean Squared Error:	2,793.21	**Percent of RMSE from Mean:**	24.54 %
Pearson Correlation:	0.6447		
Random Forest Score for Train:	-24.7618		

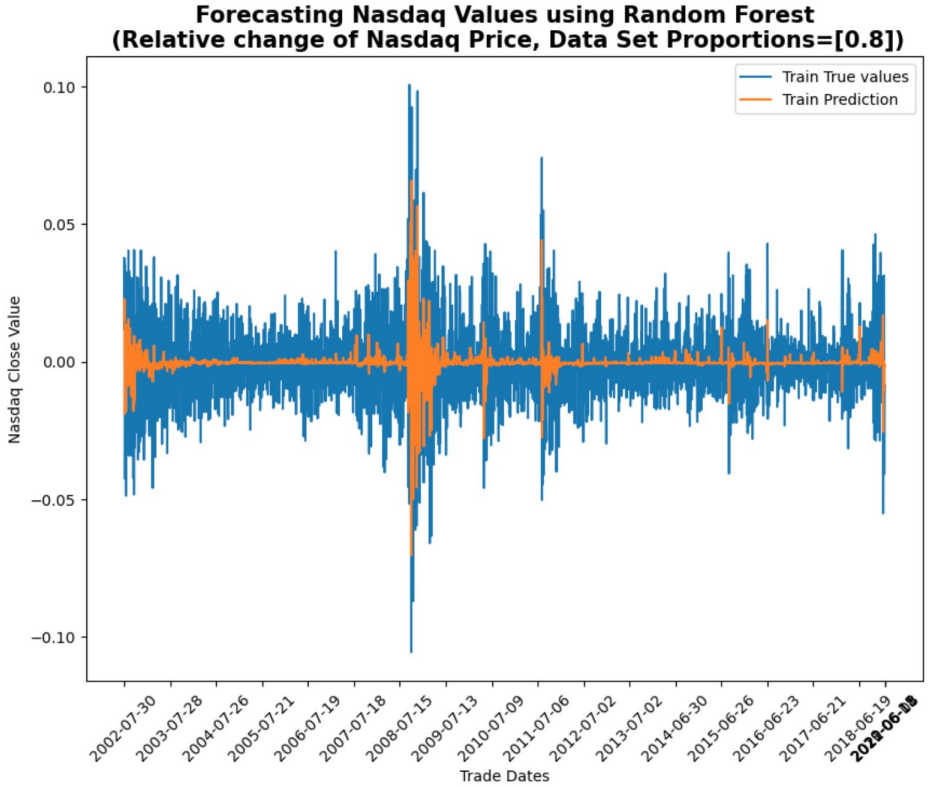
Modeling - Attempts to improve the Model

Forecasting with Relative change of the Nasdaq Price: **P(t)=[P(t+1)-P(t)]/P(t)**

As can be seen, the results are not of the best. Usually the model provides average results. However, rapid changes - up or down, are reflected promptly.

Hypothesis: Long term history may lead to high variances in the prediction.

The rational: Data from 2022 us not relevant to the current state of the market.

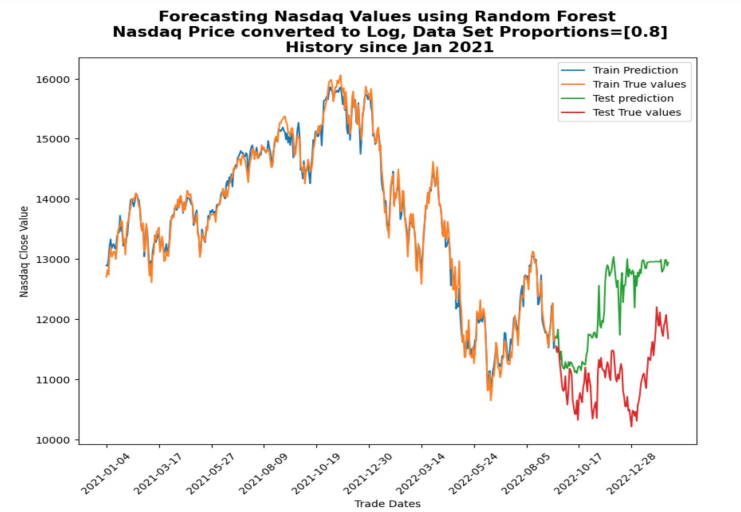


Modeling - Attempts to improve the Model

Using Data from 04/Jan/2021 until 17/Feb/2023 (536 rows in data set)

Converting Nasdaq Price to Log(Nasdaq Price). Converting back to Nasdaq Price with Anti_log

- The model successfully Learns the History.
- When splitting, 80:20 the model The prediction is ABOVE Actual the actual Nasdaq Prices.
- Accuracy pretty low (Pearson = 0.41)
- Obvious **over fitting**



Train Results

Mean Absolute Error:	94.29	**Percent of MAE from Mean:**	22.03 %
Root Mean Squared Error:	121.02	**Percent of RMSE from Mean**	28.27 %
Pearson Corelation:	1.00		
Random Forest Score for Test:	0.99		

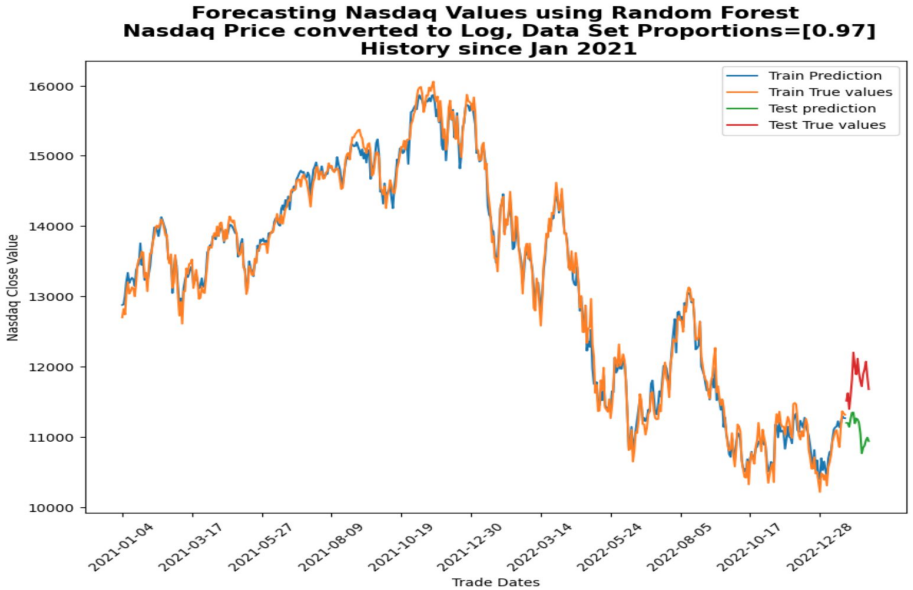
Test Results

Mean Absolute Error:	1,175.00	**Percent of MAE from Mean:**	10.64 %
Root Mean Squared Error:	1,348.59	**Percent of RMSE from Mean:**	12.21 %
Pearson Corelation:	0.41		
Random Forest Score for Train:	-6.55		

Modeling - Attempts to improve the Model

Splitting 97:03

- The model, again, successfully Learns the History.
- The prediction is again BELOW the Actual the actual Nasdaq Prices.
- Accuracy pretty low and worse (Pearson = 0.06)



Train Results

Mean Absolute Error:	96.64	**Percent of MAE from Mean:**	18.62 %
Root Mean Squared Error:	123.57	**Percent of RMSE from Mean**	23.81 %
Pearson Corelation:	1.00		
Random Forest Score for Test:	0.99		

Test Results

Mean Absolute Error:	719.17	**Percent of MAE from Mean:**	6.08 %
Root Mean Squared Error:	767.41	**Percent of RMSE from Mean:**	6.49 %
Pearson Corelation:	0.06		
Random Forest Score for Train:	-12.76		

Modeling - LSTM

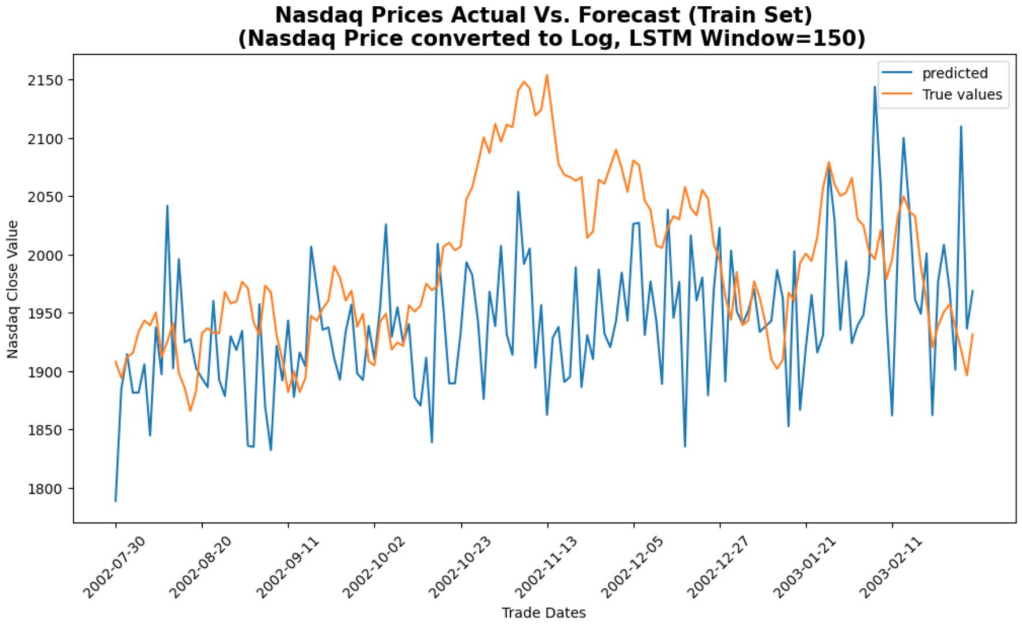
Data Preparation:

- **Normalizing Data.**
- **Nasdaq Prices converted to Log**
- **Splitting Data 70:30**
- **Defining model's parameters:**
- **Samples=300, Forecast Horizon = 150 days**
- **Constructing Model:**

```
Model: "sequential_8"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_8 (LSTM)               (None, 100)               42000

 dense_24 (Dense)            (None, 300)               30300

 dense_25 (Dense)            (None, 100)               30100

 dense_26 (Dense)            (None, 150)               15150

 reshape_8 (Reshape)         (None, 150, 1)            0

=================================================================
Total params: 117,550
Trainable params: 117,550
Non-trainable params: 0
_________________________________________________________________
```

- **After Model Run:** Prediction converted back to Stock Prices with inverse-transform and Anti-Log.

Modeling - LSTM



Train Results:

Train Results

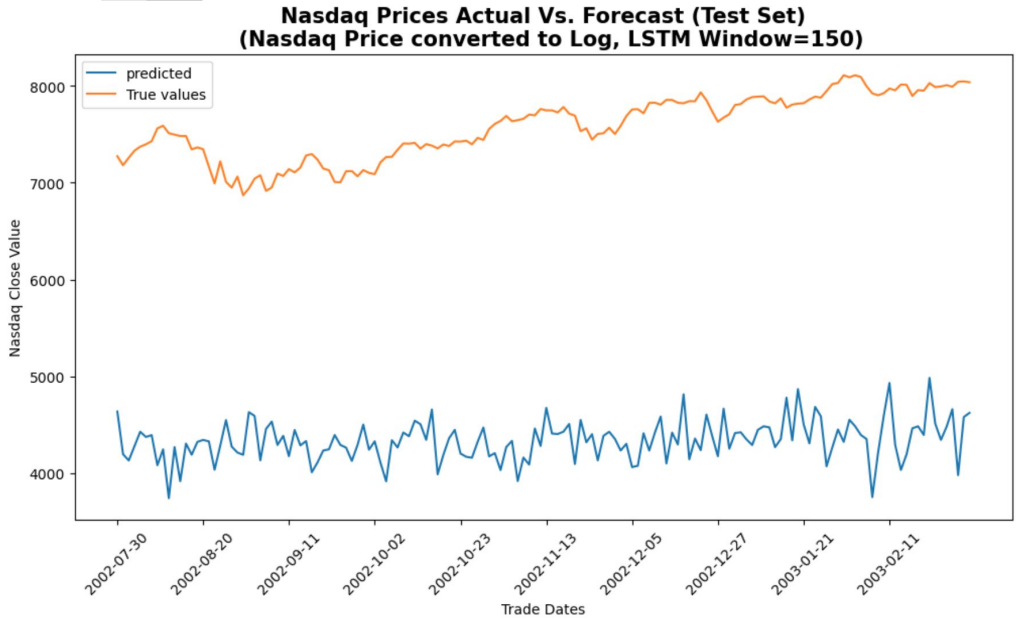
Mean Absolute Error:	74.06	**Percent of MAE from Mean:**	3.7189 %
Root Mean Squared Error:	93.21	**Percent of MAE from Mean:**	4.6804 %
Pearson Correlation:	0.23		

Prediction is consistently below the actual Market Prices.

The Model didn't succeed to reflect the trends of the market.

Model Indicators are low.

Modeling - LSTM



Test Results:

Test Results

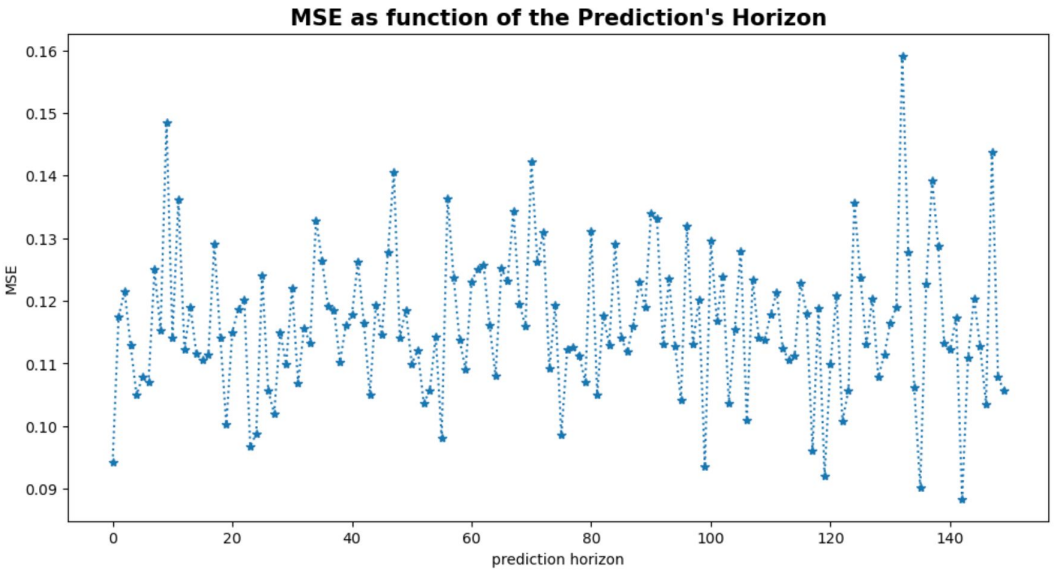
Mean Absolute Error:	3,234.12	**Percent of MAE from Mean:**	42.7142 %
Root Mean Squared Error:	3,253.45	**Percent of MAE from Mean:**	42.9695 %
Pearson Correlation:	0.19		

Obviously the Model didn't succeed in predicting nor follow the trends.

Prediction seems conservative as reflex some average and consistently below the actual Market price.

The gap between actual values and prediction is of -42% averagely.

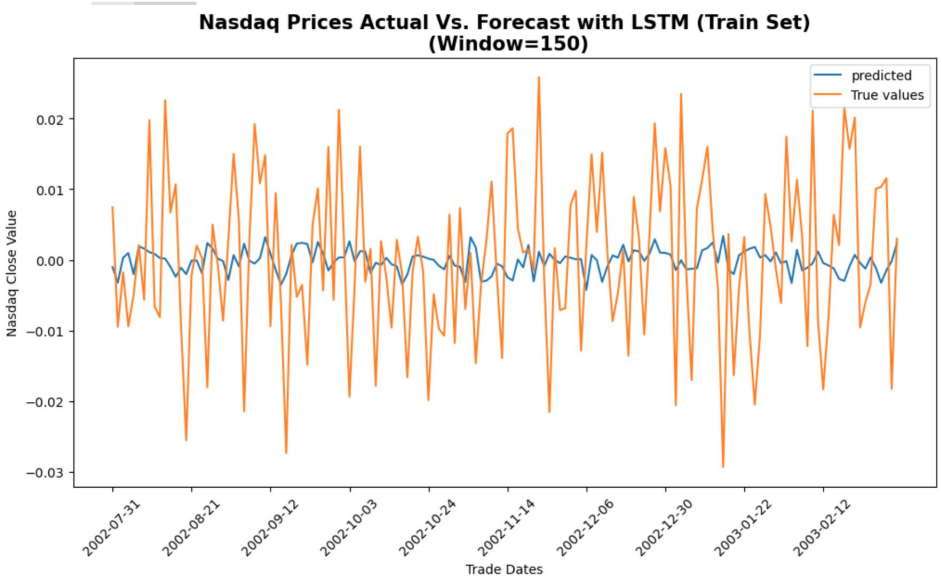
Modeling - LSTM



MSE as function of Prediction horizon:

No consistency found. And we can conclude that the model with the current parameters is not succeeding in predicting Nasdaq precises.

Modeling - LSTM



Pearson Correlation (Train): -0.03

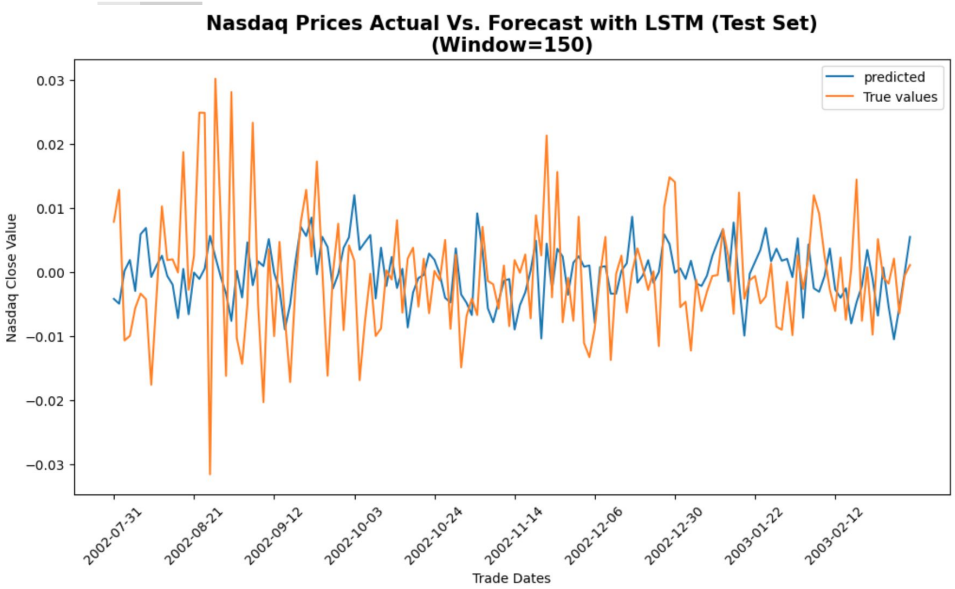
<u>Conversion of Nasdaq Prices to relative change:</u>

Forecasting with relative change of the Nasdaq Price: **P(t)=[P(t+1)-P(t)]/P(t)**

The relative Market prices are fluctuating in range of +/-0.025.

The Model's forecast is steady around the average with low correlation..

Modeling - LSTM



Pearson Correlation (Test): 0.0047

Conversion of Nasdaq Prices to relative change:

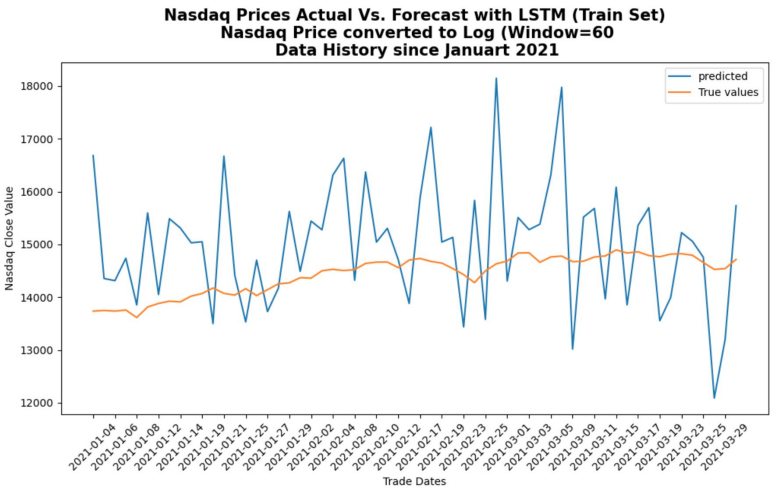
The Pearson correlation is lower than the one received with Train Set.

Modeling - LSTM

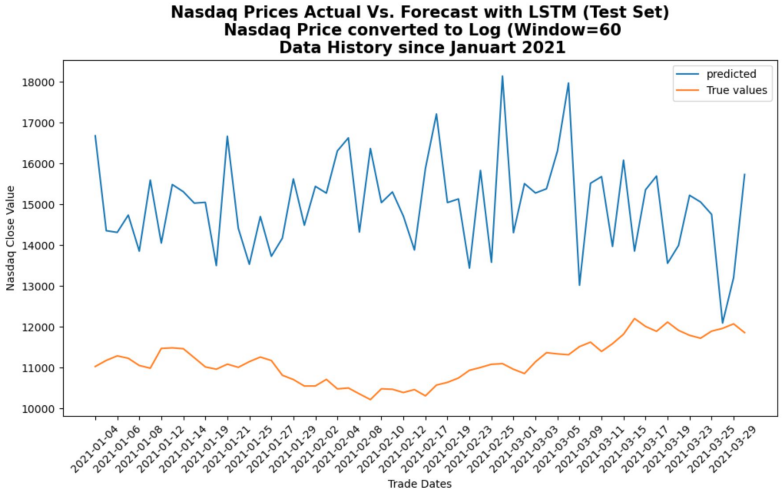
Using Data from 04/Jan/2021 until 17/Feb/2023 (536 rows in data set), Prediction Window = 60 days

Converting Nasdaq price to log(Nasdaq Price). Returning back to Nasdaq price with Anti-Log(Nasdaq Price)

- Better prediction achieved at Train Set (Relative MAE=6.5%)
- At Test Set inaccurate prediction, higher than the actual prices and unsteady.
- When splitting, 80:20 the model manage to predict the train however miss the actual results.
- When splitting 97:03, Prediction improved.

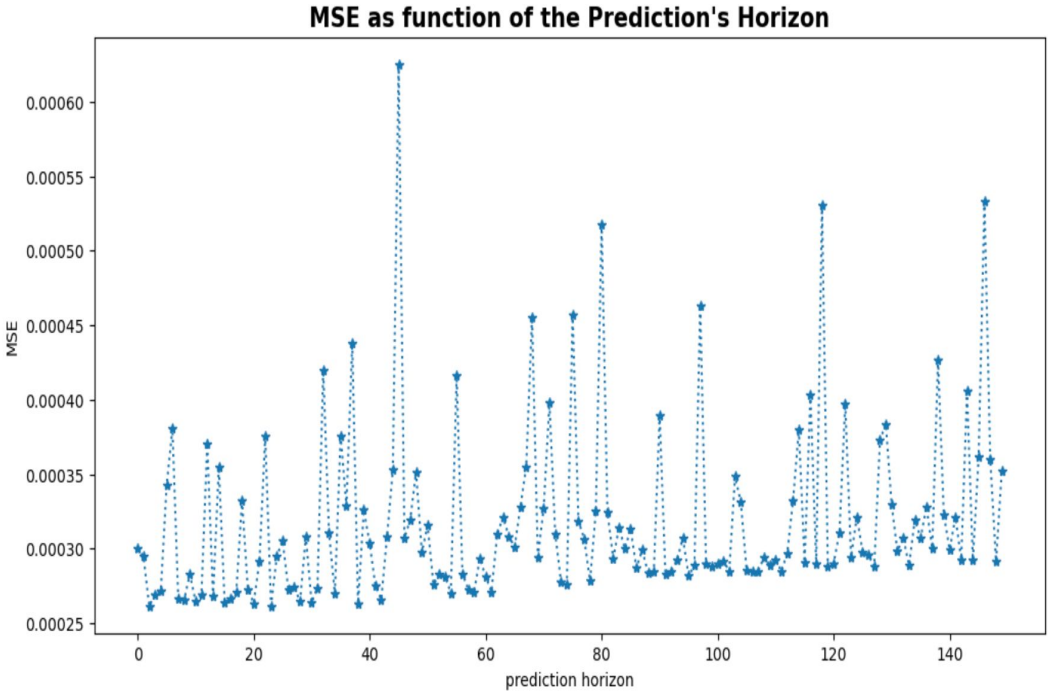


Pearson Correlation (Train): 0.14



Pearson Correlation (Test): -0.26

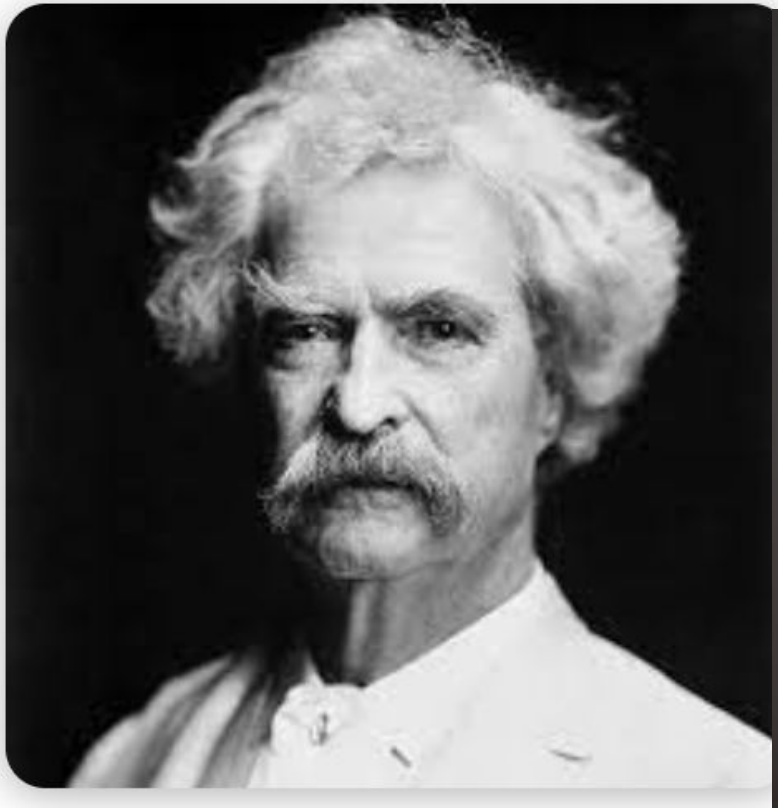
Modeling - LSTM

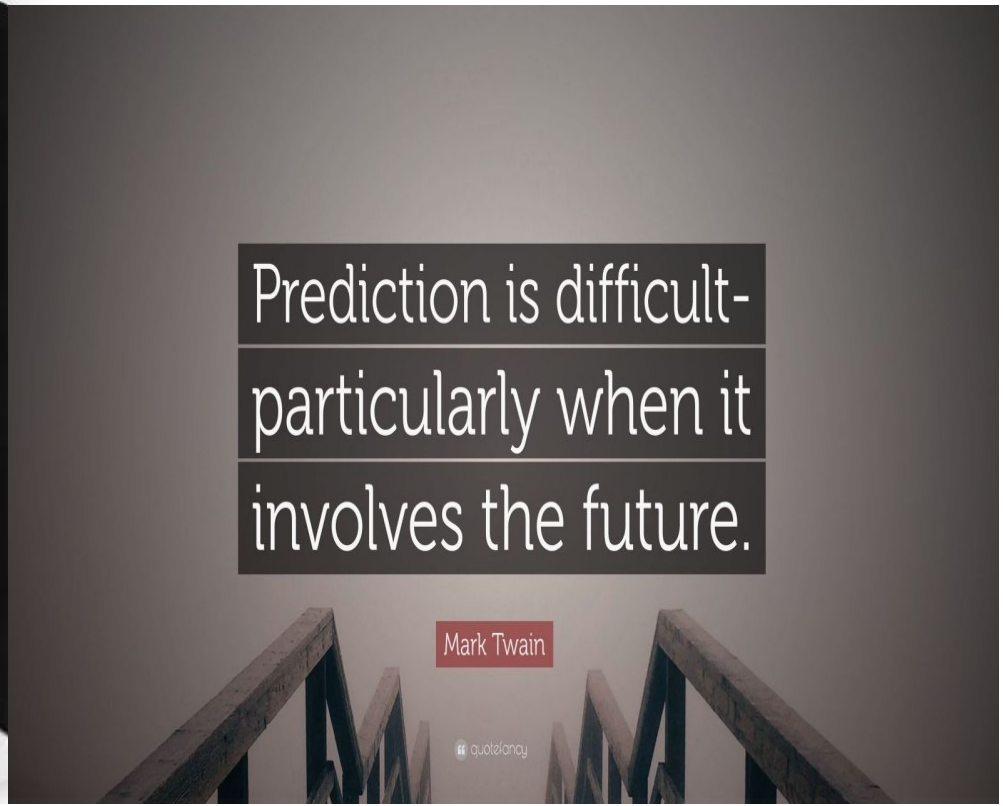


MSE as function of Prediction horizon:

No consistency found. And we can conclude that the model with the current parameters is not succeeding in predicting Nasdaq precises.

Summary and Conclusions





Summary and Conclusions

- Although **LSTM and Random Forest are known as models** who can learn and predict **Time Series**, we achieved un-satisfying results.
- The explanatory **variables we choose are known as very relevant** to the security market. We assume that the variables we worked with, is not necessarily the reason for the low results we achieved.
- There could be Bug in the Models Construction that we didn't reveal, however, the scripts were thoroughly tested and examined. We believe that **Bug in models is less reasonable**.
- **Data History**: We were using data that started at 2002. The market share today is not the same is it was by then. Perhaps **giving weight to old data, can distort the learning of ML models and it forecasts**.

Summary and Conclusions

- **Forecasting Horizon** - we tried to predict 150 days ahead (30 Working weeks of the Nasdaq). Fluctuations and changes of market's direction can degenerate the learning stage and prevent accurate prediction (We saw excellent fit on Train set with RandomForest, however very disappointing forecast on Test Set.
- **Rapid Changes:** We believe that as long as the market is steady and incline/decline moderately, there is fair chance to predict the future. Dramatic change of direction like the one occurred in 2020, is a harsh obstacle for predictions.

Summary and Conclusions

Recommendations:

- Examine more models.
- Instead of predicting Nasdaq value, with continuous variable, the model will use a Binary variable where ‘1’ indicates that within X days the market will go up, and ‘0’ indicates that market will go down.
- Working on short periods of history time, and run predictions to up to 5 days (a working week).
- **Future work** - Find more variables. Using NLP tools to measure investors’ sentiments and other ‘unformatted’ data (analyst opinions, news etc’) can give strong foundations for good prediction.


Thank you!