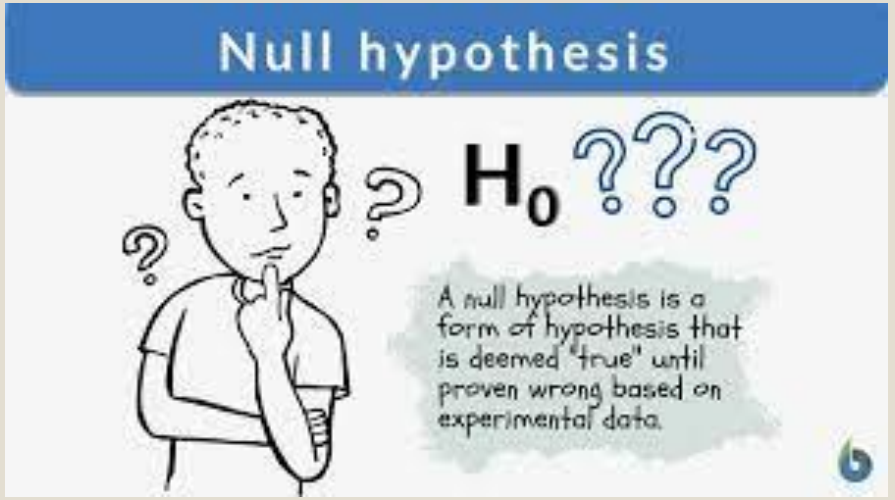
PCDSI0121
DS105-MACHINE LEARNING PROJECT

IS STOCK PRICE PREDICTABLE USING LSTM?

KANG BOON KIAT

Hypothesis

- Null Hypothesis: Many people believe that stock closing price is predictable and forecastable using general LSTM model with stock closing price.
- Alternate Hypothesis: I believe that LSTM not able to predict and forecast stock price closing price by using stock closing price.
- To test the hypothesis:
  - Implement the LSTM model to a stock data (Closing price)
  - Visualise the prediction and forecast result.

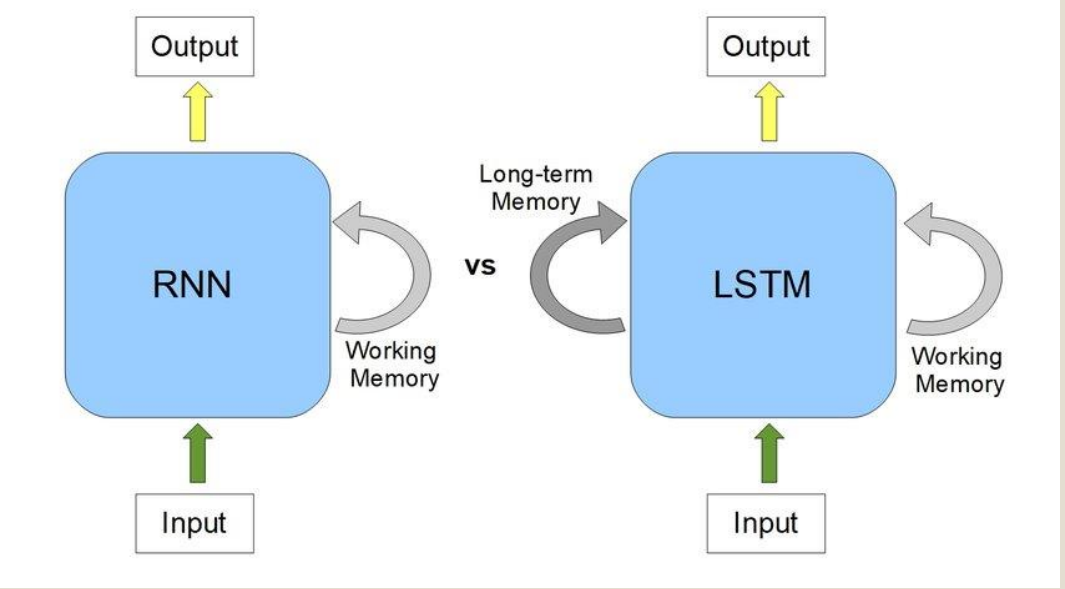


Content

- Hypothesis
- What is LSTM?
- Hypothesis Testing
- Conclusion to the Hypothesis Testing
- Why Stock data is/not Predictable?
- Difficulties and Take-away

What is LSTM?

- Long Short Term Memory Network is type of Recurrent neural network.
- Unlike standard RNN, it is capable of learning long-term dependencies, especially in sequence prediction problems.
- LSTM has feedback connections, i.e., it is capable of processing the entire sequential data, apart from single data points such as images.



Hypothesis Testing

LSTM Model Testing

1. Download dataset
2. Exploratory Data Analysis (EDA)
3. Building Model
4. Evaluation Test Model
5. Visualisation of the result

Download dataset

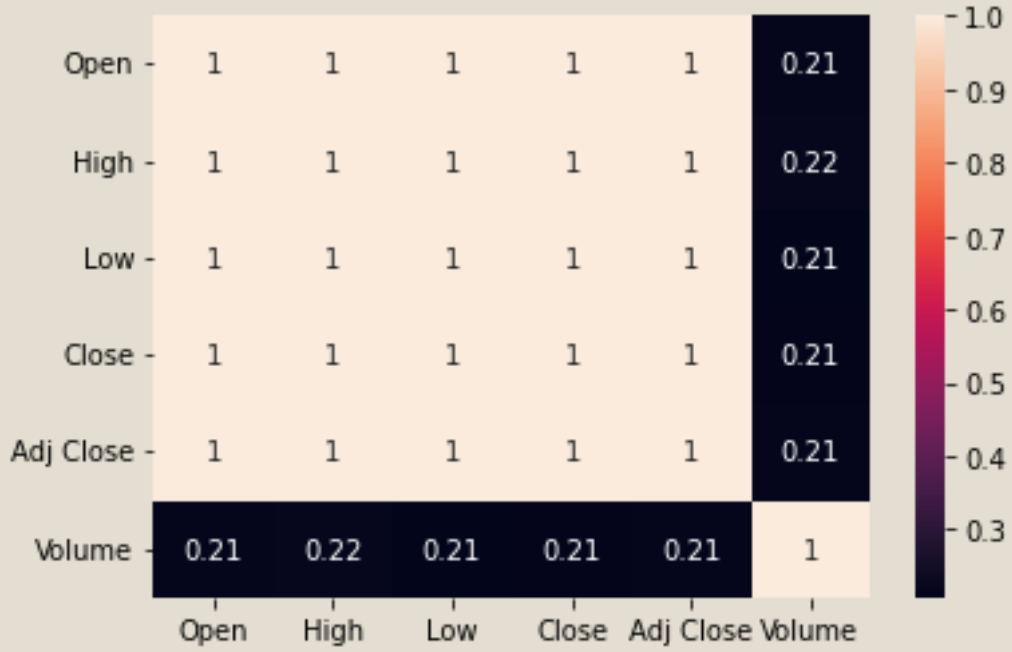
- Stock data of Vanguard Information Technology Index Fund ETF (VGT)
  - Dated 2013-12-31 – 2022-04-14

Date	Open	High	Low	Close	Adj Close	Volume
2013-12-31	89.099998	89.580002	89.080002	89.540001	82.474739	428500
2014-01-02	89.160004	89.160004	88.430000	88.620003	81.627304	358100
2014-01-03	88.760002	88.779999	88.199997	88.320000	81.350998	829500
2014-01-06	88.379997	88.430000	87.820000	88.070000	81.120728	386300
2014-01-07	88.500000	89.099998	88.300003	88.959999	81.940506	317000
...	...	...	...	...	...	...
2022-04-08	401.489990	402.250000	396.660004	397.649994	397.649994	444300
2022-04-11	392.540009	392.880005	387.890015	388.230011	388.230011	499800
2022-04-12	393.980011	396.850006	385.540009	387.170013	387.170013	2719400
2022-04-13	387.260010	395.260010	386.089996	393.920013	393.920013	592000
2022-04-14	394.269989	395.140015	384.000000	384.179993	384.179993	392700

2088 rows × 6 columns

Exploratory Data Analysis (EDA)

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 2088 entries, 2013-12-31 to 2022-04-14
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Open       2088 non-null   float64
 1   High       2088 non-null   float64
 2   Low        2088 non-null   float64
 3   Close      2088 non-null   float64
 4   Adj Close  2088 non-null   float64
 5   Volume     2088 non-null   int64
dtypes: float64(5), int64(1)
memory usage: 114.2 KB
```

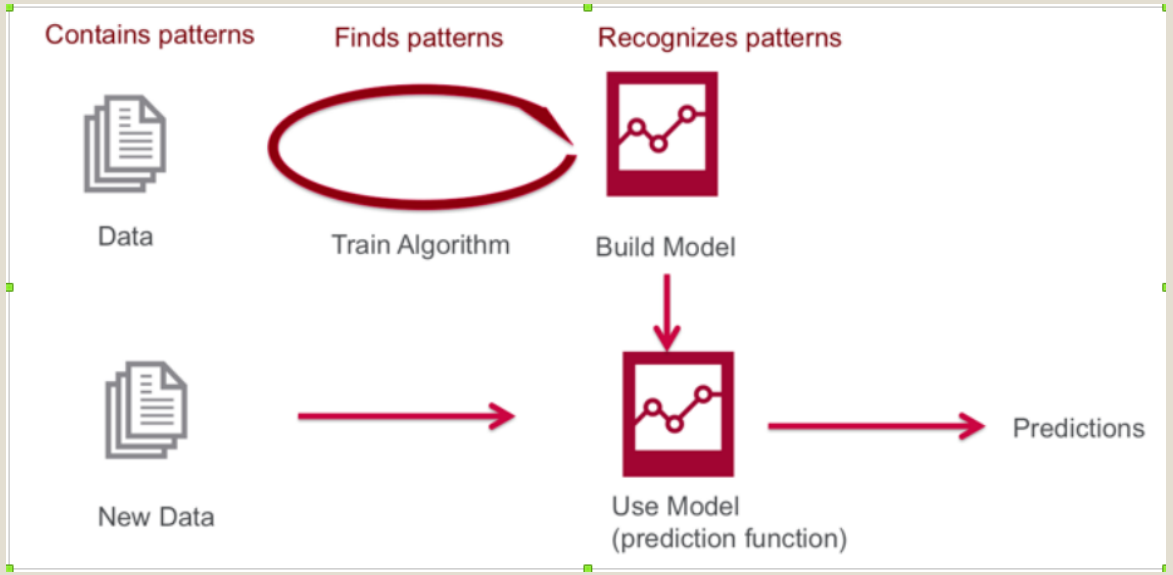


All data are non-object and correlations all are ‘1’. It is a standard stock dataframe.


VGT STOCK
Price
Volume
450
400
350
300
250
200
150
100
2014
2016
2018
2020
2022

Building Model

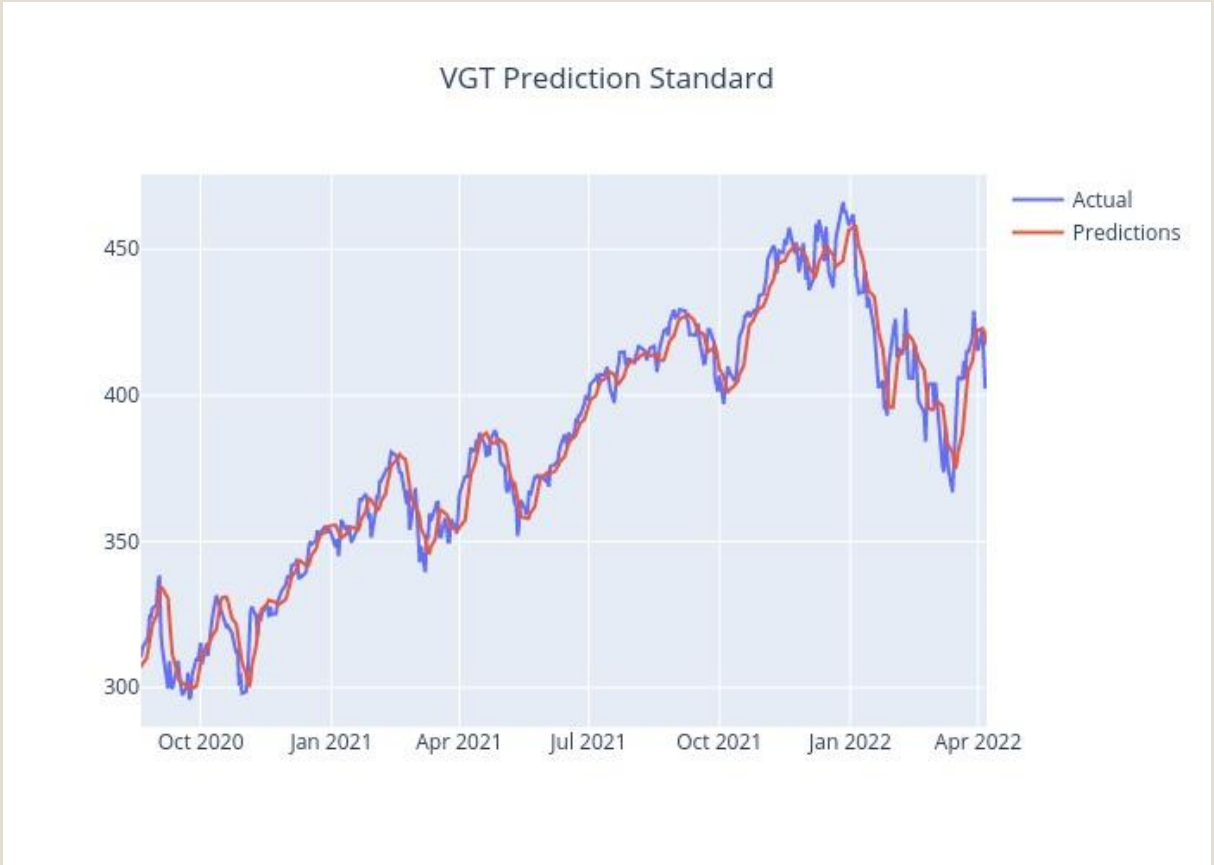
- Slicing 80% for train and 20% for test
- Using past-60 days to predict the next day
- Standard Scaling method (Whole dataset):
  - StandardScaler
  - MinMaxScaler
- LSTM stacking layer of 2 and 50 units each
- Epochs = 20 and batch size = 16

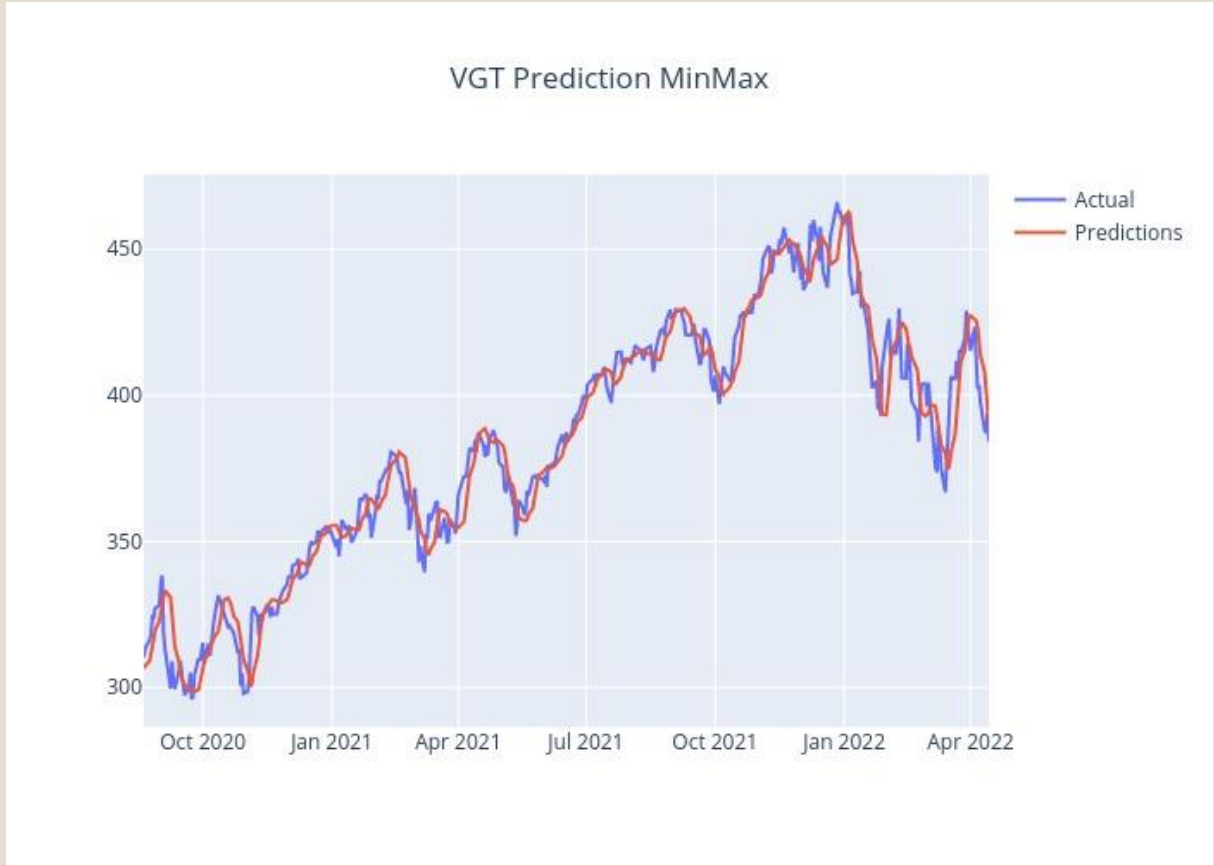


Evaluation of model

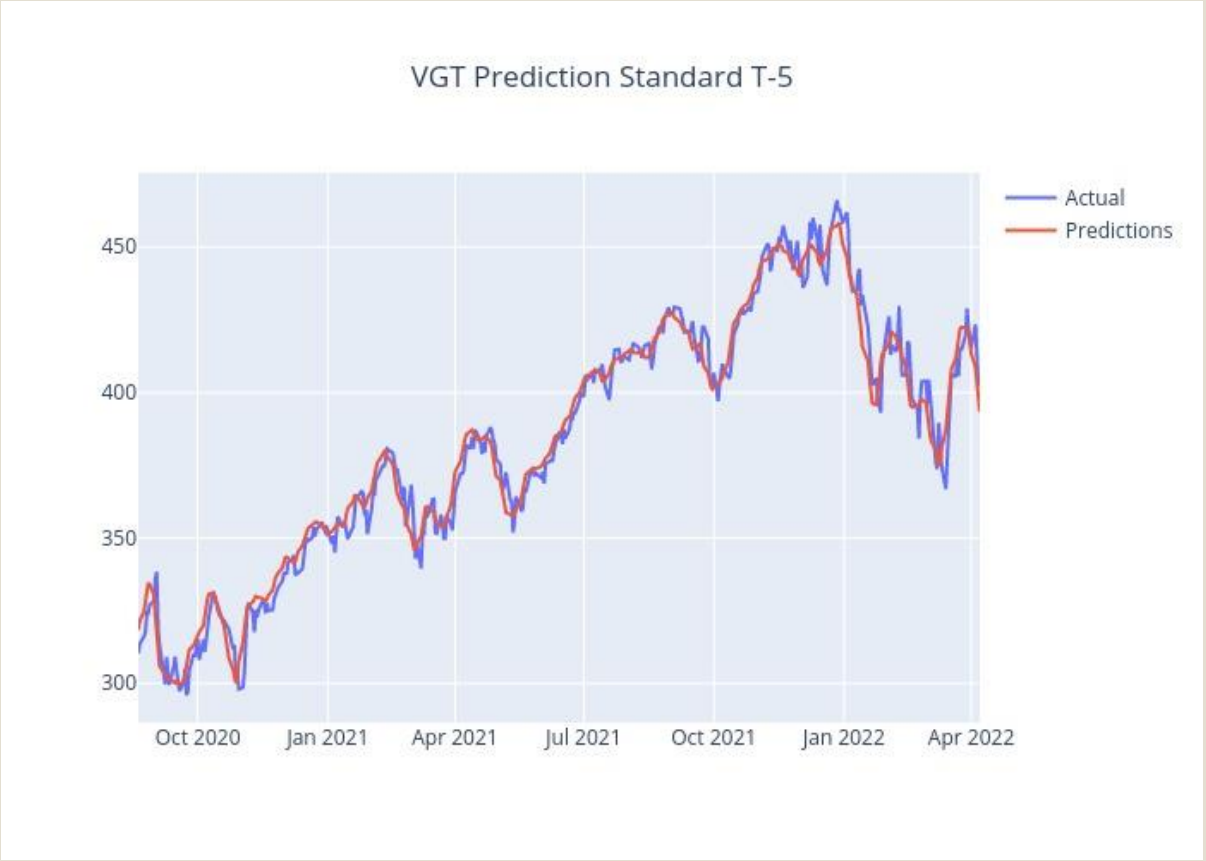
	MinMaxScaler	StandardScaler
Adjusted-R-Squared	0.95	0.95
Mean Absolute Percentage Error	0.017	0.017
Max Absolute Percentage Error	0.1	0.1
Min Absolute Percentage Error	3.09e-05	0.00016

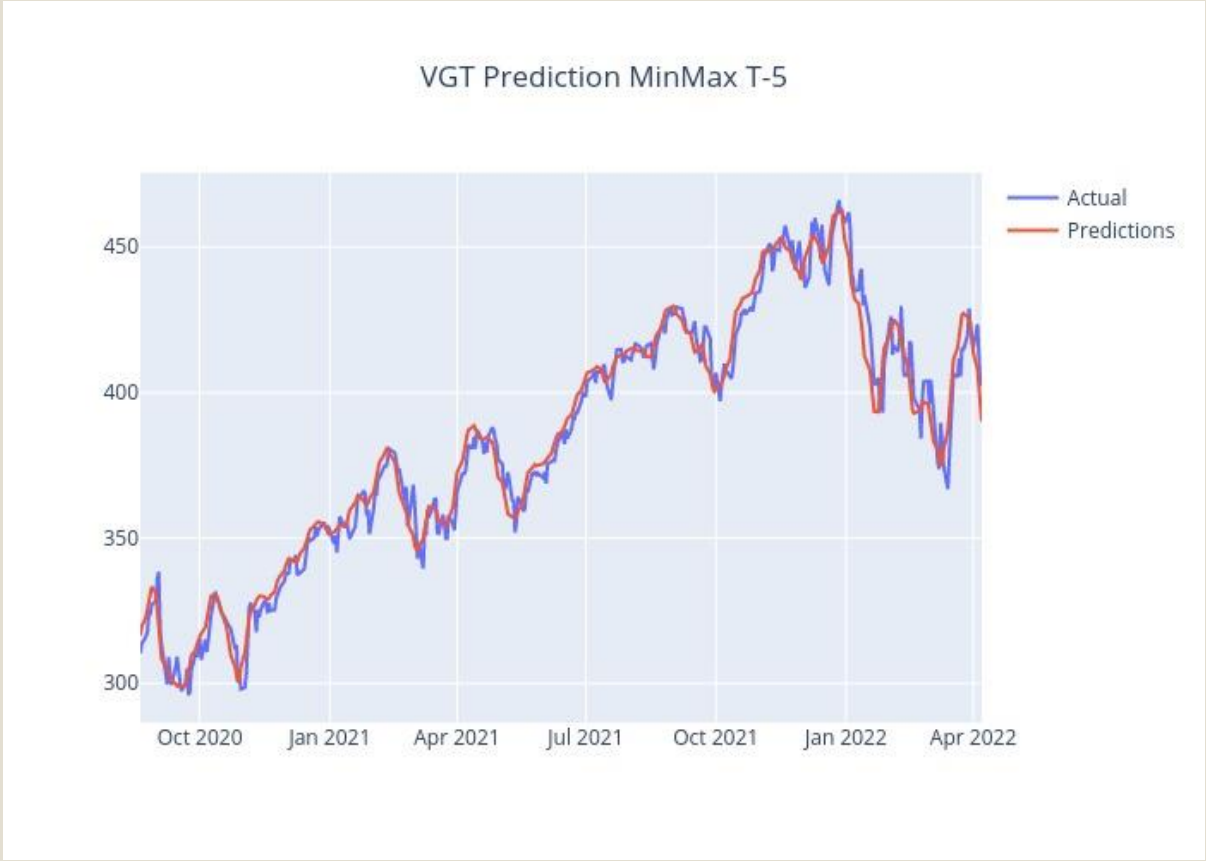
Prediction against Actual Plot





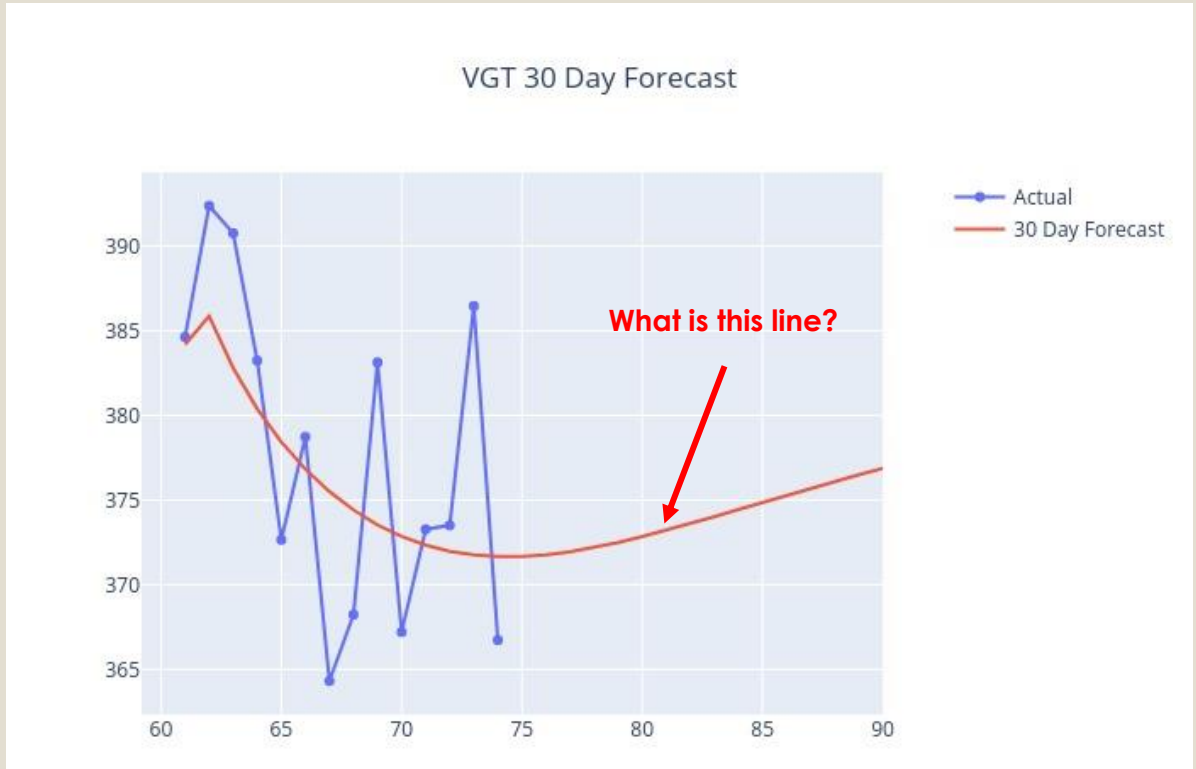
Plot of Shifting the prediction (T-5)



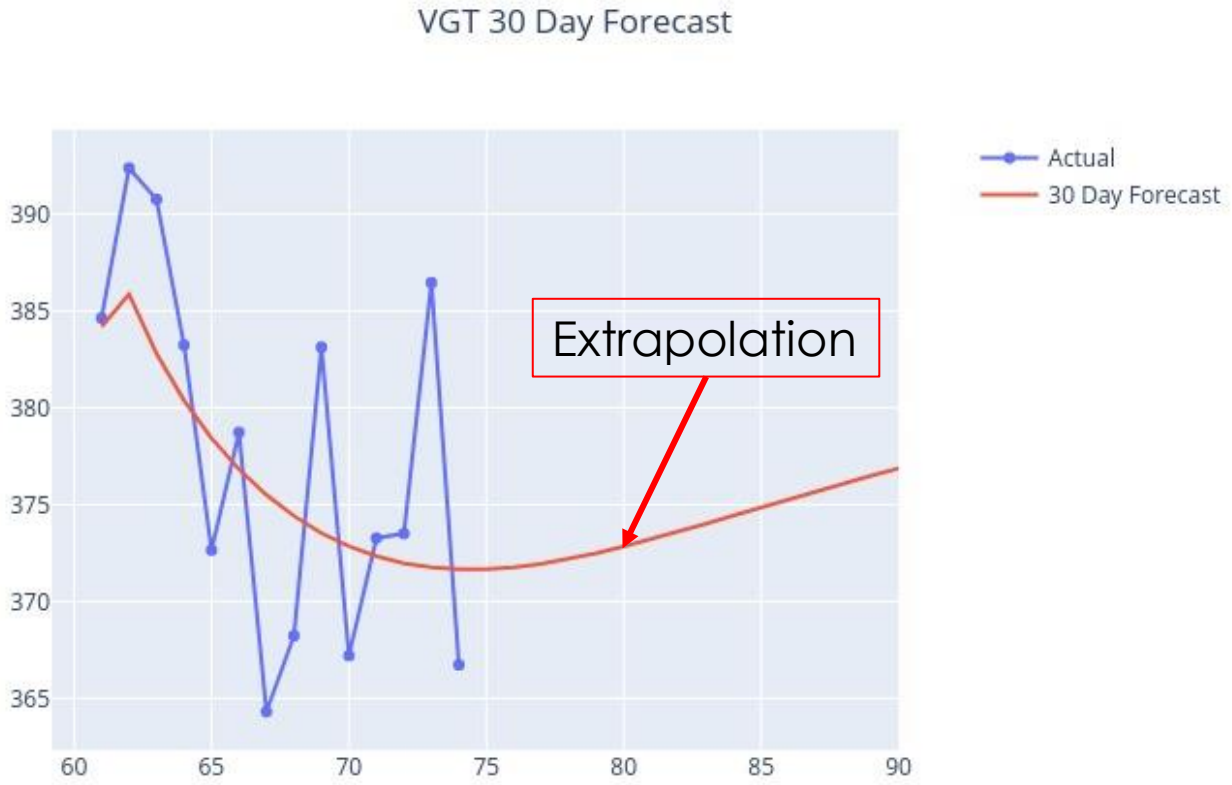


'Predicted values' are based past price movement

Forecasting Plot



Extrapolation



Extrapolation is defined as an estimation of a value based on extending the known series or factors beyond the area that is certainly known.

Under LSTM, the extrapolation tends to move linearly.

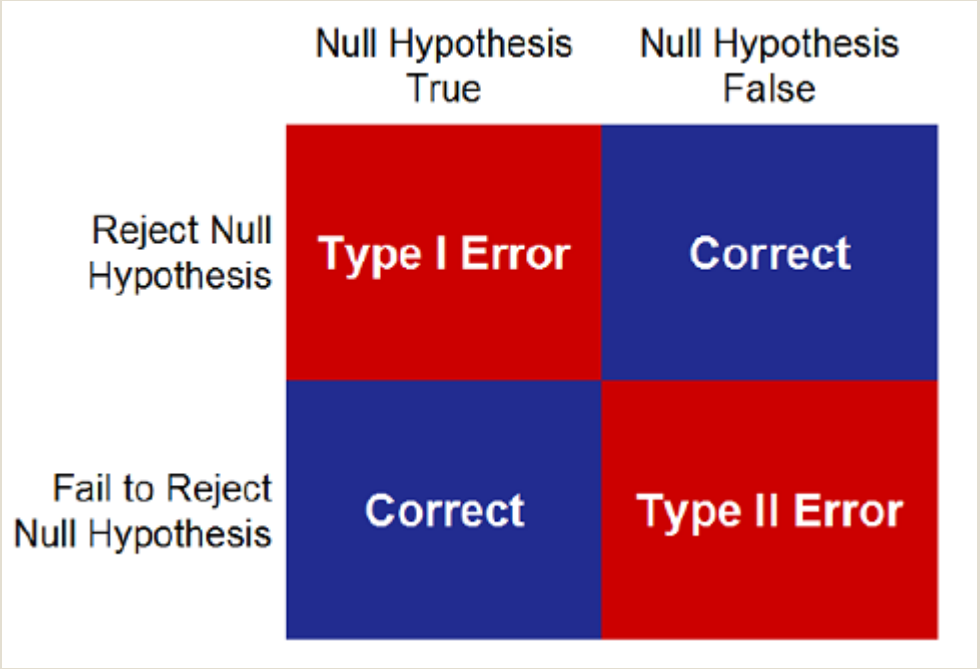
The extrapolation depends on the model not the data.

Conclusion to the Hypothesis Testing

Reject Null Hypothesis.

Reason:

- The predicted stock price is based on past price movement
- Stock price cannot be forecasted due to the forming of extrapolation line.



Why Stock data ~~is~~/not Predictable?

For Time series to be predictable, the data set must be stationary and normal distribution.

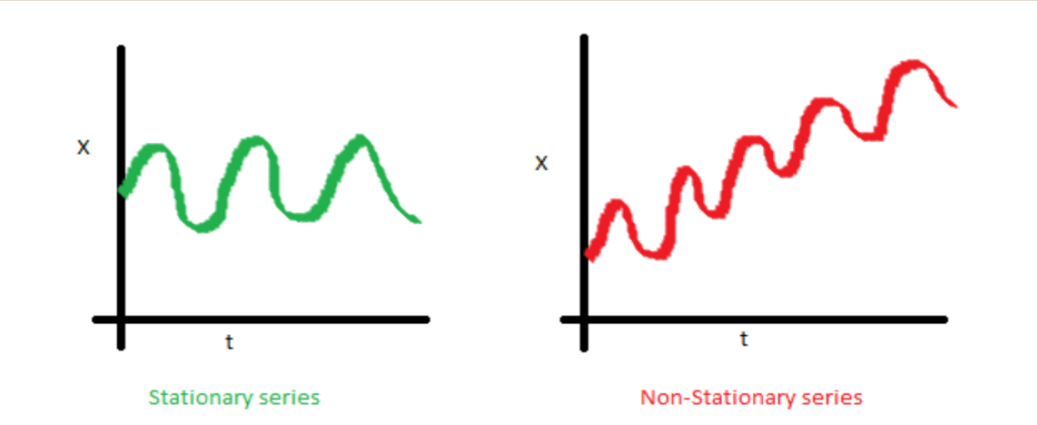
- A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times.

- Normal Distribution is an important concept in statistics and the backbone of Machine Learning.
  - is a continuous probability distribution. It has a bell-shaped curve that is symmetrical from the mean point to both halves of the curve.
  - Properties:
    - The mean, mode and median are all equal.
    - The curve is symmetric at the center (i.e. around the mean, μ).
    - Exactly half of the values are to the left of center and exactly half the values are to the right.
    - The total area under the curve is 1.

Test for Stationarily

- Augmented Dickey-Fuller Test (ADF Test)
  - A common statistical test used to test whether a given Time series is stationary or not
  - For non-stationary, statistic test value > critical value and p-value > significant value(0.05)
  - For stationary, statistic test value and p-value are very low than critical value and significant value (0.05)

- Note: The significant value can be any values (0 – 1) depends on the confident level



Test for Stationarily

Input the VGT's Close values into the ADF:

```
ADF Statisitic: 0.287946
p-value: 0.976785
Critical Values:
        1%: -3.434
        5%: -2.863
        10%: -2.568
Time Series is Non-Stationary
```

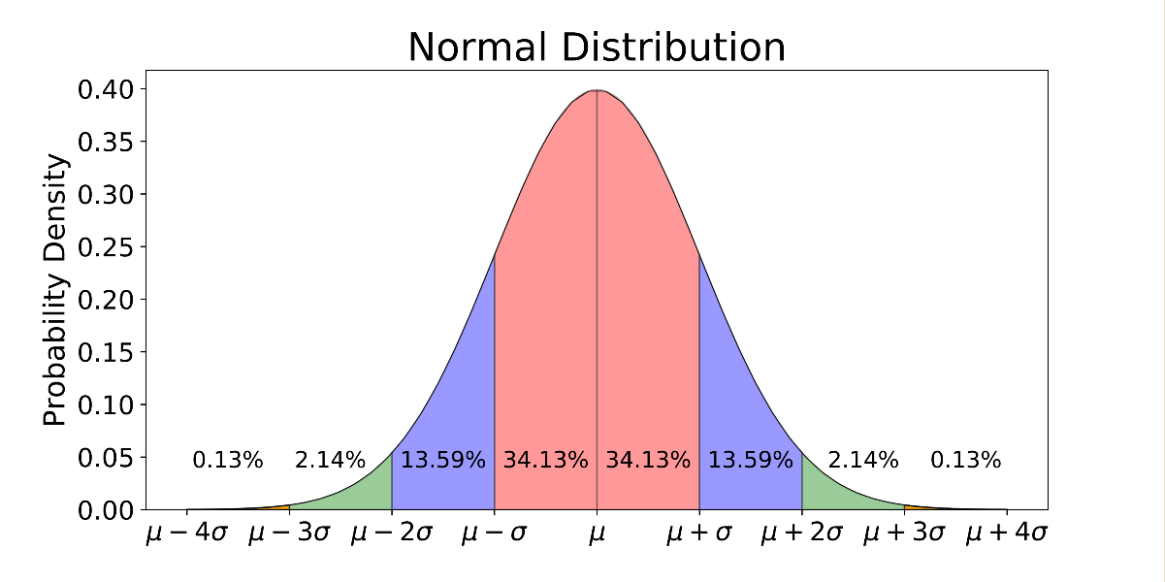
Possible to convert non-stationary to stationary? Yes. Using Differencing.

Result:

```
ADF Statisitic: -11.794485
p-value: 0.000000
Critical Values:
        1%: -3.434
        5%: -2.863
        10%: -2.568
Time Series is Stationary
```

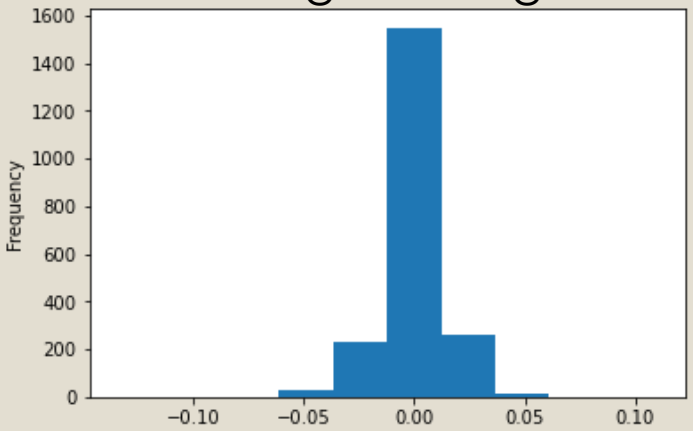
Test for Normal Distribution

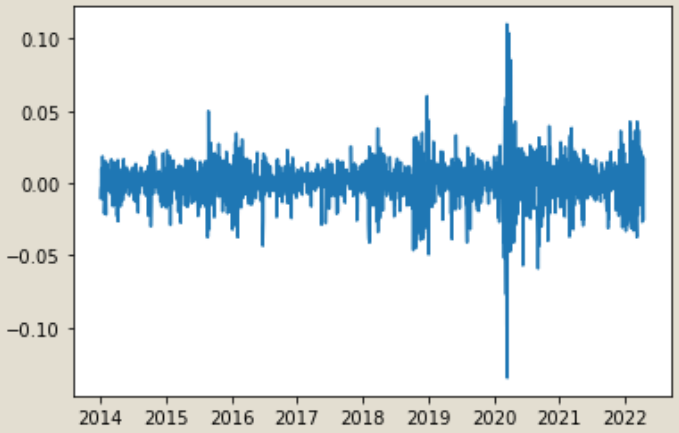
- Percentage Change
- Log Transformation
- Hypothesis Testing Library
  - Kolmogorov Smirnov test (ks-test)
  - Shapiro Wilk test

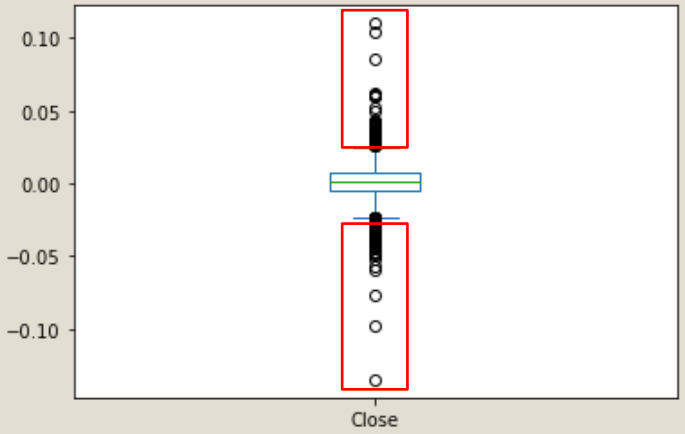


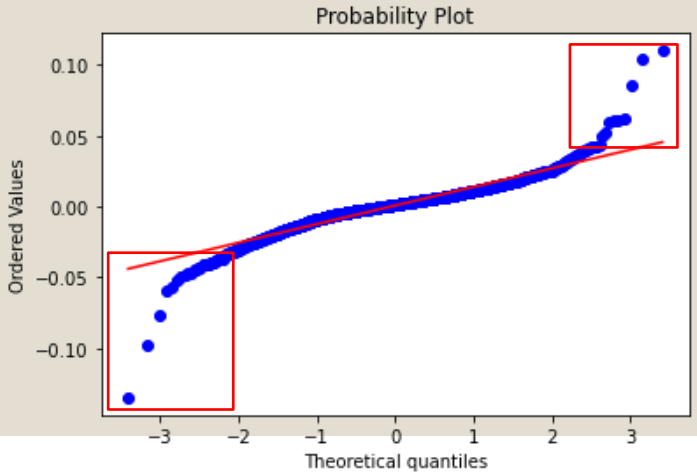
Test for Normal Distribution

Percentage Change:





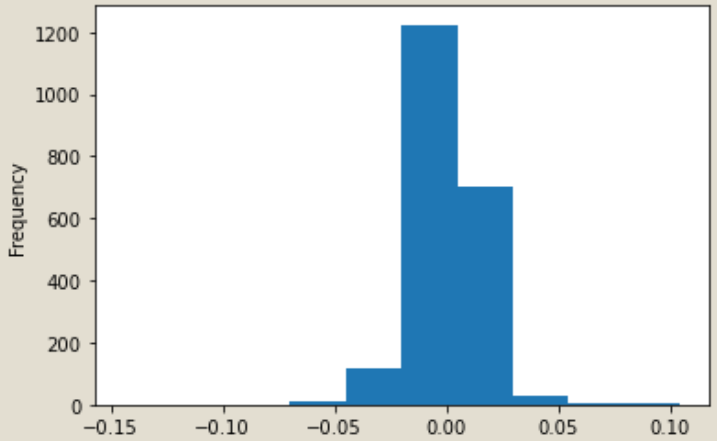


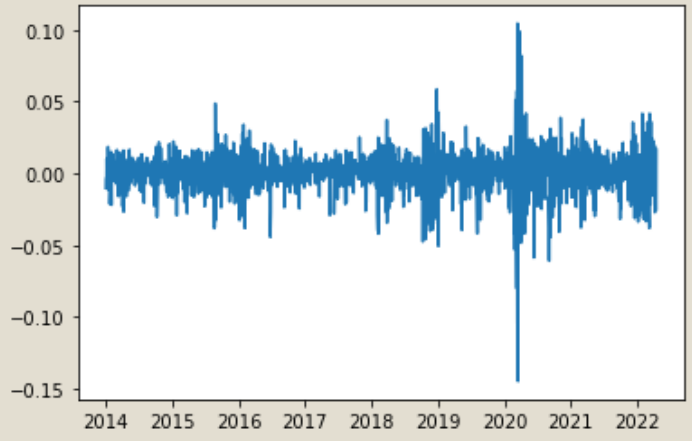


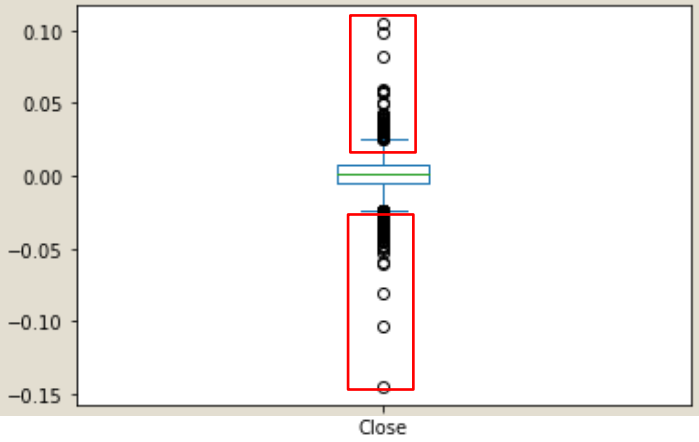

```
Mean: 0.0007943870522762759
Mode: 0    0.0
Name: Close, dtype: float64
Median: 0.0013314104634460922
```

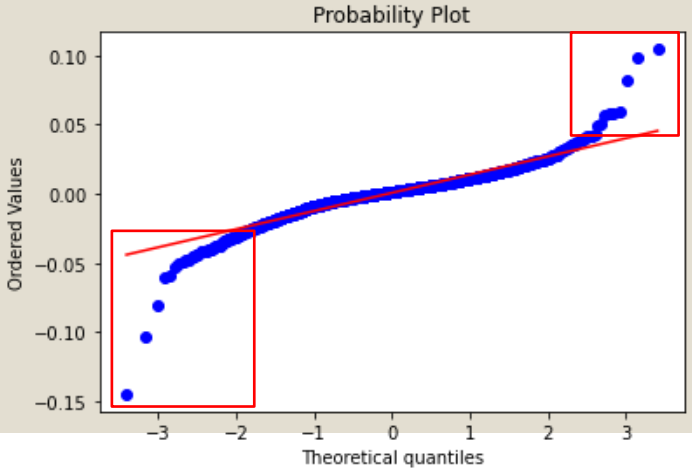
Test for Normal Distribution

Log Transformation:










```
Mean: 0.0006978561143539093
Mode: 0    0.0
Name: Close, dtype: float64
Median: 0.0013305249224602146
```

Test for Normal Distribution

Kolmogorov Smirnov Test (ks-test)

- it is commonly used as a test for normality to see if your data is normally distributed. It's also used to check the assumption of normality in Analysis of Variance.
- the test compares a known hypothetical probability distribution (normal distribution) to the distribution generated by data
- For Normal distribution, Statistic Value = 0, P-Value > set significant value (0.05)

Note: The significant value can be any values (0 – 1) depends on the confident level

Result:

```
ks_stat: 0.47871176849303665, p_value: 0.0
Probably NOT Guassian
```

Test for Normal Distribution

Shapiro Wilk test

- Stats Model developed specifically only for normal distribution.
- For Normal distribution, Statistic Value = 0, P-Value > set significant value

Note: The significant value can be any values (0 – 1) depends on the confident level

Result:

```
sw_stat: 0.8971880078315735, p_value: 2.877398088214O2e-35
Probably NOT Guassian
```

Difficulties and Take-away

Difficulties:

- Too much misinformation regards to the usage of LSTM

Take-away:

- Stock data is not predictable
- Reason are due to the nature of the data
  - Not Stationary
  - Not Normal Distribution

THE END

What is sequential data?

Whenever the points in the dataset are dependent on the other points in the dataset the data is said to be Sequential data. A common example of this is a Timeseries such as a stock price or a sensor data where each point represents an observation at a certain point in time

Symbol	Date	Open	Close
AAPL	2019 - 12 -26	71.21	72.48
GOOG	2019 - 12 -26	1346.17	1360.40
NFLX	2019 - 12 -26	334.60	332.63
AAPL	2019 - 12 -27	72.78	72.45
GOOG	2019 - 12 -27	1362.99	1351.89
NFLX	2019 - 12 -27	332.96	329.09
GOOG	2019 - 12 -30	1350.00	1336.14