機器學習期末報告-

Stock Prediction in Python.

Presenter: 彭歆雯
Adviser: 蔡宗憲老師

Department of Computer Science and Information Engineering,
Chung Cheng Institute of Technology,
National Defense University, Taoyuan, Taiwan, R.O.C.

Report

2021/12/14

Outline

- Software Requirement
- Environment Setup
- DataSet
- Pre-processing
- Model Description
- Parameters Tuning
- Result Analysis
- Conclusion
- Reference

Software Requirement

- Windows 10
- Anaconda (python 3.8)
- Jupyter notebook 6.3.0
- Spyder 4.2.5

Environment Setup

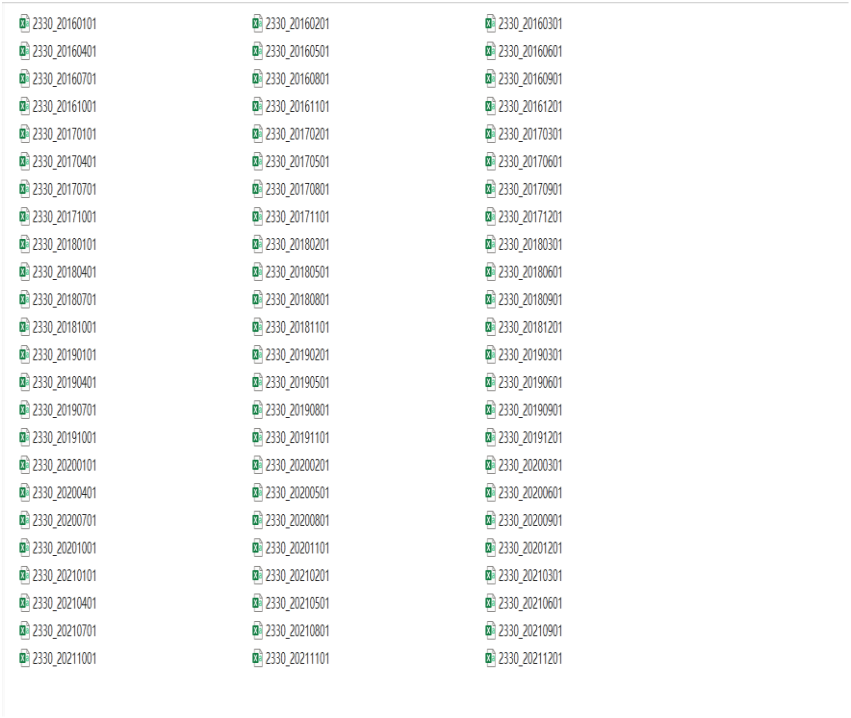
- c:\> pip install quandl
- c:\> pip install numpy
- c:\> pip install matplotlib
- c:\> pip install pandas
- c:\> pip install plotly
- c:\> pip install requests
- c:\> pip install pytrends
- c:\> pip install pystan
- c:\>conda install -c conda-forge fbprophet

(必須先pip install pystan，才能conda install -c conda-forge fbprophet，且需要版本python ≥3.5)

DataSet

- 從台灣證券交易所「個股日成交資訊」取得台積電(股票代碼：2330)自2016年1月1日至2021年12月13日之歷史股價，並依每月存成.cvs檔。

```python
import requests
import pandas as pd

dates = [20200601, 20201001, 20201101]
stockNo = 2330
url_template = "https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=html&date={}&st

for date in dates :
    url = url_template.format(date, stockNo)
    file_name = "{}_{}.csv".format(stockNo, date)

    data = pd.read_html(requests.get(url).text)[0]
    data.columns = data.columns.droplevel(0)
    data.to_csv(file_name, index=False)

```



Pre-processing

- 因自台灣證券交易所所存取的資料中日期為民國年且中文字顯示為亂碼，透過此階段修正日期為西元年及標題欄位，以利後續讀取資料時可正常執行。

	A	B	C	D	E	F	G	H	I
1	?交?	? 漱?⊥	? 漱?	? ???:?霓		?嗥 ??賭?	漱蝑		
2	105/01/04	43800291	6.14E+09	142.5	143.5	139	139.5	-3.5	14188
3	105/01/05	46502108	6.44E+09	139	140	137	138	-1.5	15836
4	105/01/06	53873344	7.32E+09	138	138	135	135.5	-2.5	15926
5	105/01/07	63475065	8.43E+09	134.5	135	130.5	133	-2.5	19061
6	105/01/08	52641383	7.02E+09	132	135	132	134	1	11784
7	105/01/11	48536267	6.43E+09	133	134.5	130.5	133	-1	10980
8	105/01/12	35133164	4.66E+09	133	134	131.5	133	0	9492
9	105/01/13	50237061	6.72E+09	133.5	135.5	133	133.5	0.5	9488
10	105/01/14	39498490	5.18E+09	130.5	132	130.5	131.5	-2	14501
11	105/01/15	79648971	1.09E+10	137.5	138	135.5	137	5.5	20280
12	105/01/18	37008681	5.05E+09	135	138	134.5	137	0	11947
13	105/01/19	24995537	3.42E+09	137.5	138	136	138	1	8146
14	105/01/20	44423192	6E+09	137.5	137.5	133.5	134.5	-3.5	13799
15	105/01/21	31957378	4.32E+09	134	136.5	134	135	0.5	9956
16	105/01/22	33926219	4.66E+09	136.5	138.5	136	138.5	3.5	12684
17	105/01/25	27213347	3.8E+09	140	140	139	139.5	1	9105

	Date	Volume	Turnover	Open	High	Low	Close	Change	Transaction
0	2016-01-04	43800291	6137797502	142.5	143.5	139.0	139.5	-3.5	14188
1	2016-01-05	46502108	6436850512	139.0	140.0	137.0	138.0	-1.5	15836
2	2016-01-06	53873344	7321139064	138.0	138.0	135.0	135.5	-2.5	15926
3	2016-01-07	63475065	8432791819	134.5	135.0	130.5	133.0	-2.5	19061
4	2016-01-08	52641383	7022206705	132.0	135.0	132.0	134.0	1.0	11784

Model Description

- Facebook於2017年2月份公開一種基於Python及R語言的時間序列的預測模型-「Prophet」。
- Prophet是透過加法模型來描述資料：

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

  - 在所有時間點t的數值$y(t)$是由：
    - 趨勢的影響$g(t)$
    - 季節性的影響$s(t)$
    - 假日的影響$h(t)$
    - 誤差ε_t

Model Description

- Trend model：
  - Logistic：用於容量C會飽和預測。

$$g(t) = \frac{C(t)}{1 + \exp(-k \cdot (t - m))}$$

    - with *C* the carrying capacity, *k* the growth rate, and *m* an oset parameter.
  - Linear ：用於容量C不飽和預測，最大容量會隨時間等因素有所改變。

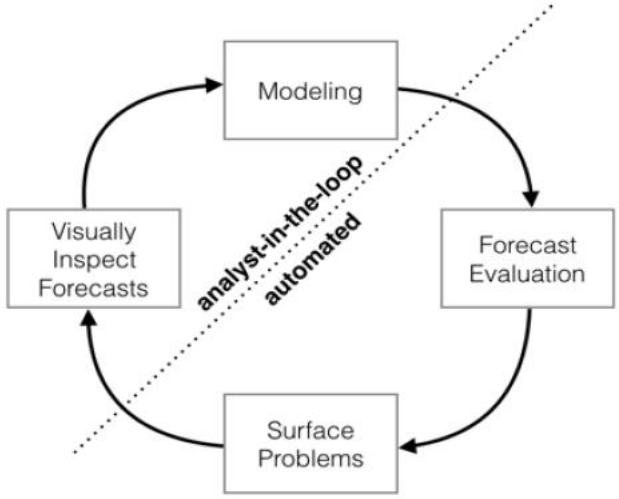
$$g(t) = k \cdot t + m$$

Model Description

- 而Prophet還加入*Change Points*的想法，讓Trend model在不同時間區間內，能有不同的成長率K。

$$k(t) = k_0 + \sum_j \mathbf{1}(t \leq t_j) \cdot \delta_j$$

- 預測模型Prophet透過各參數的微調，可以讓時間序列型的資料能輕鬆預測。



Parameters Tuning

- 透過stocker執行股價預測。
- 先讀取歷史資料，並使用plot視覺化台積電(股票代碼：2330)自2016-01-04至2021-12-13止之歷史股價及顯示此期間內之最高值、最低值及最近值股價。

```
from stocker import Stocker
stock = Stocker(stockNo, df)
```

```
stockNo : 2330 Stocker Initialized. Data covers 2016-01-04 00:00:00 to 2021-12-13 00:00:00.
```

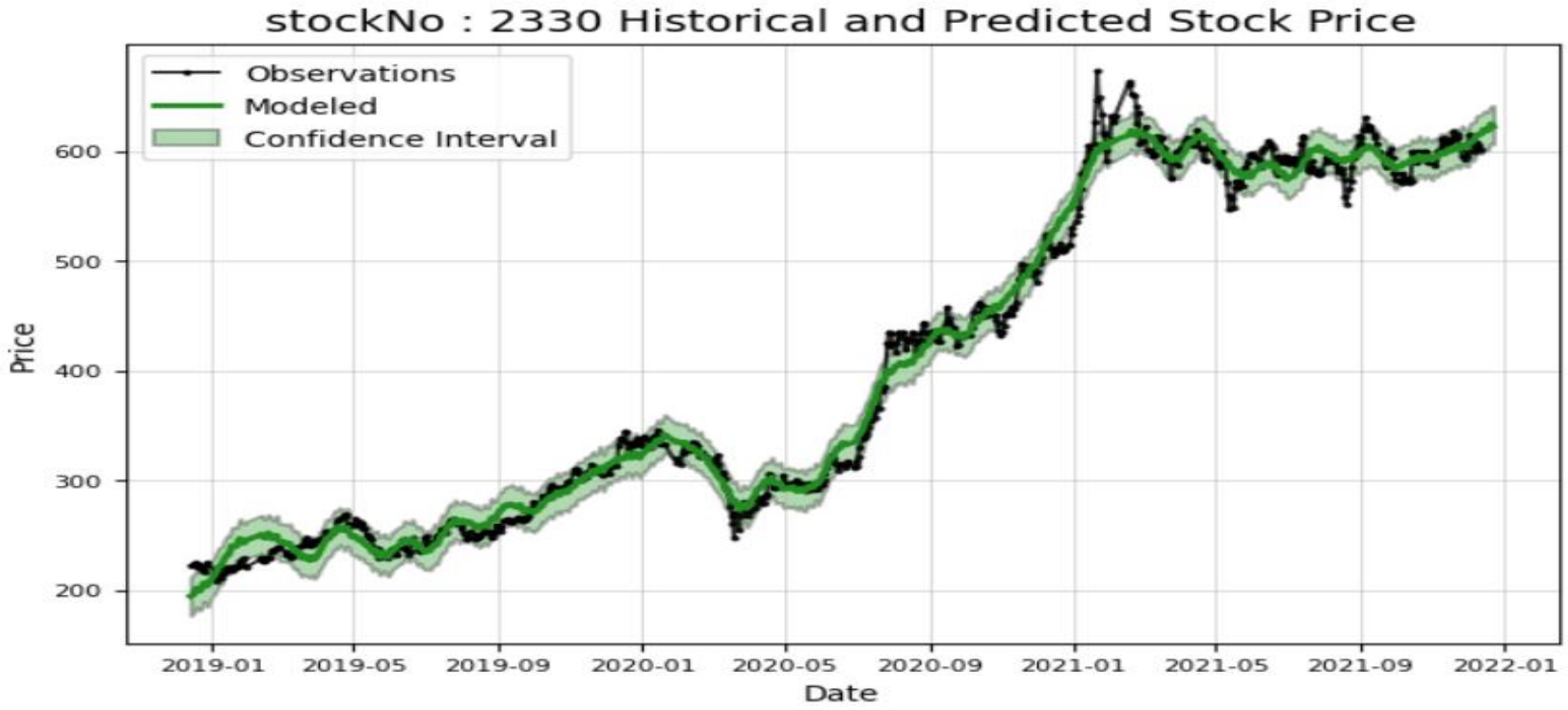
```
stock.plot_stock()
```

```
Maximum Adj. Close = 673.00 on 2021-01-21 00:00:00.
Minimum Adj. Close = 131.50 on 2016-01-14 00:00:00.
Current Adj. Close = 601.00 on 2021-12-13 00:00:00.
```

Parameters Tuning

- 使用 `stock.create_prophet_model()` 函式預測台積電(股票代碼：2330) 之未來10日的股價。

```
Predicted Price on 2021-12-23 00:00:00 = $622.39
```



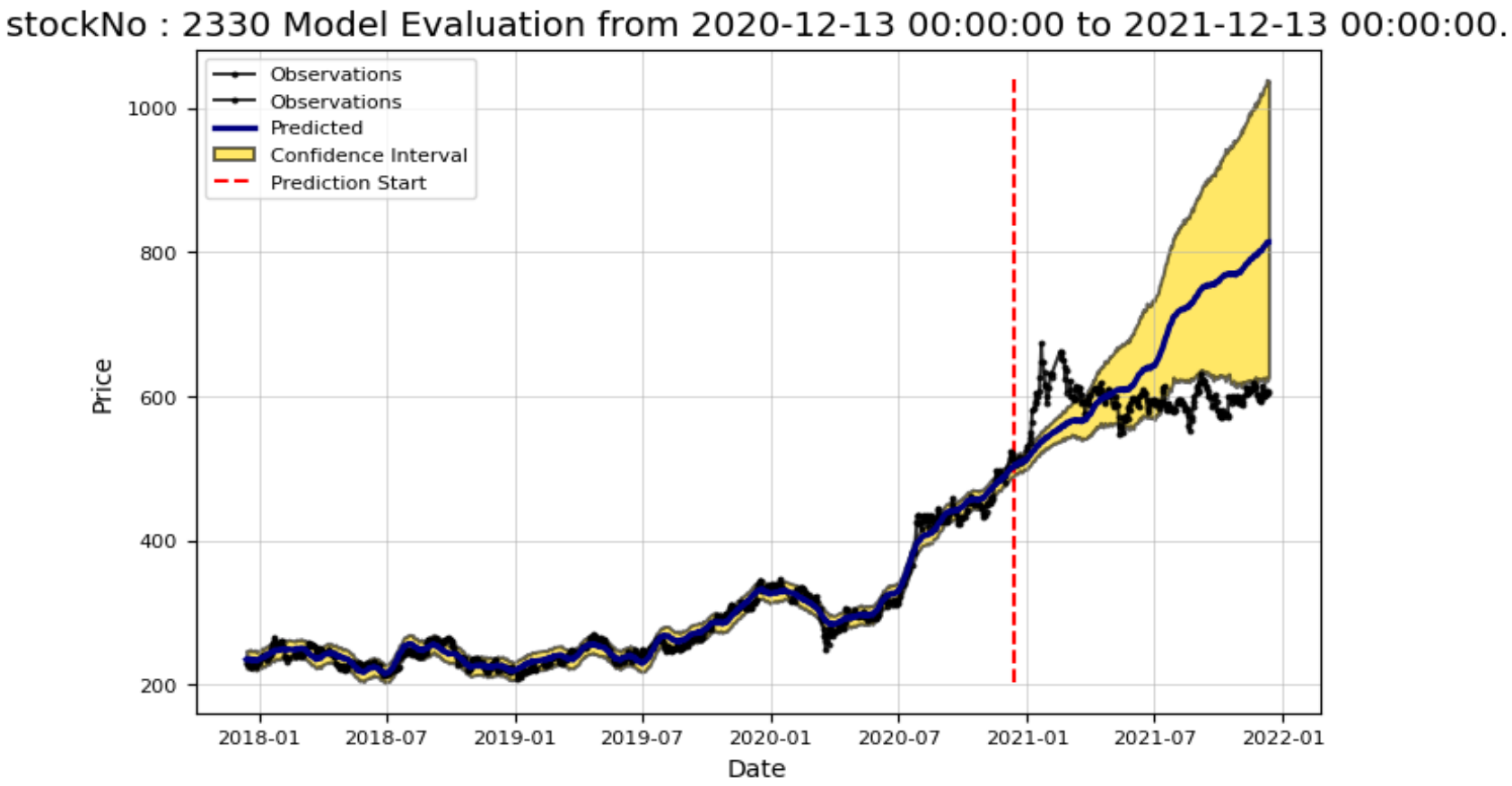
Parameters Tuning

- 而為創建最佳模型，對蒐集之數據為區分三組執行：
  - Training set：2016-01-04至2019-12-12。
  - Validation set：2019-12-13至2020-12-12。
  - Testing set：2020-12-13至2021-12-13。

Parameters Tuning

- 為了確認此模型之預測結果是否準確，透過使用`stock.evaluate_prediction()`將2017-12至2020-11月三年份資料作為training set及將2020-12至2021-12月一年份資料作為testing set。
- Stocker計算之指標包含：
  - 預測及實際之價格
  - 訓練及測試集之平均誤差
  - 正確預測價格變化方向之時間百分比
  - 實際價格落在預測之信心區間內之時間百分比。
- 由下圖顯示，用預設參數之模型訓練結果不甚理想。

```
Prediction Range: 2020-12-13 00:00:00 to 2021-12-13 00:00:00.

Predicted price on 2021-12-11 00:00:00 = $814.15.
Actual price on    2021-12-10 00:00:00 = $605.00.

Average Absolute Error on Training Data = $7.75.
Average Absolute Error on Testing  Data = $90.87.

When the model predicted an increase, the price increased 48.17% of the time.
When the model predicted a  decrease, the price decreased  53.85% of the time.

The actual value was within the 80% confidence interval 36.48% of the time.
```



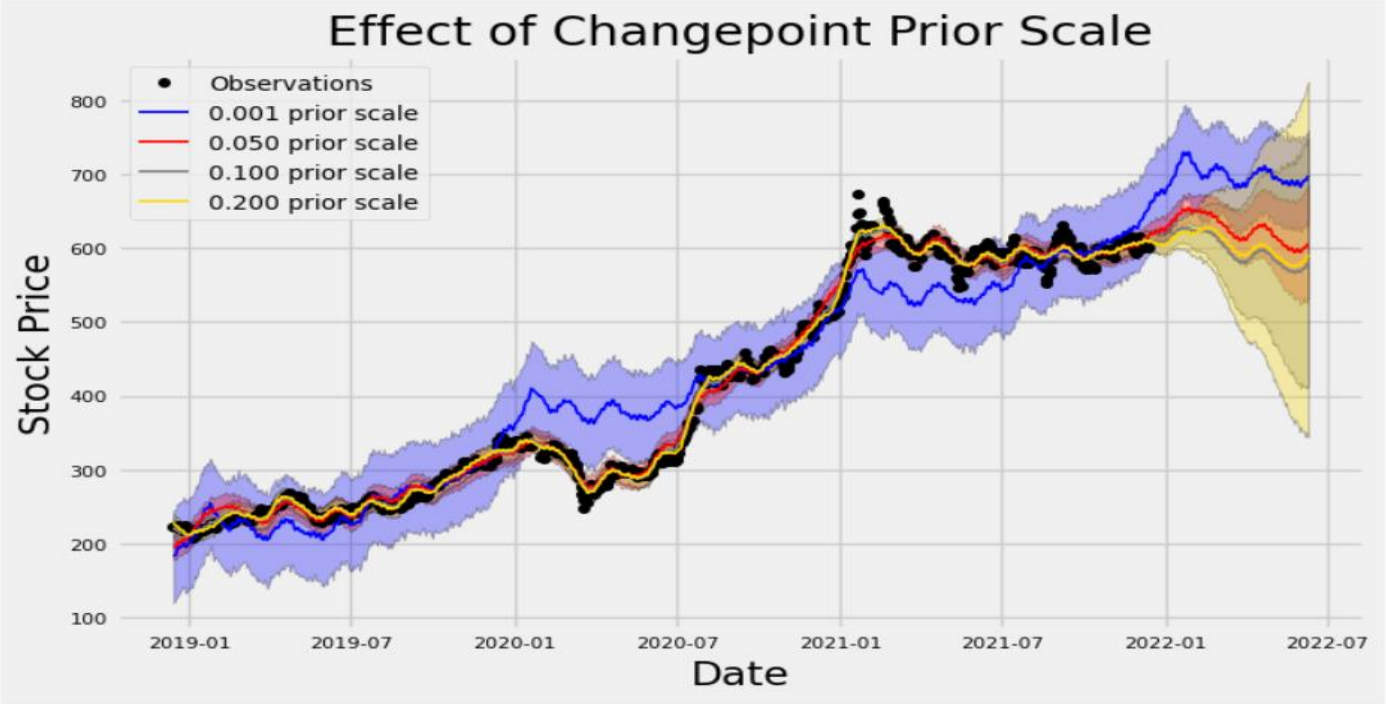
Parameters Tuning

- 由於預測與實際狀況相差過大，可以透過參數調整讓模型預測更加準確，調整不同的`changepoint_prior`讓模型以不同的強度去訓練資料(deault:changepoint_prior_scale=0.05)，找到較適合之值，並避免overfitting及underfitting。
- 透過`stock.changepoint_prior_analysis(changepoint_priors=[])`析模型不同的changepoint_prior_scale參數值對預測結果之變化。
- 使用3年份資料做訓練，並顯示6個月之預測結果。

Parameters Tuning

- 由下圖顯示changepoint_prior_scale0.01與實際值相差最大，0.200 與實際值最相近，而預設值0.05介於兩端之間。

```
stock.changepoint_prior_analysis(changepoint_priors=[0.001, 0.05, 0.1, 0.2])
```

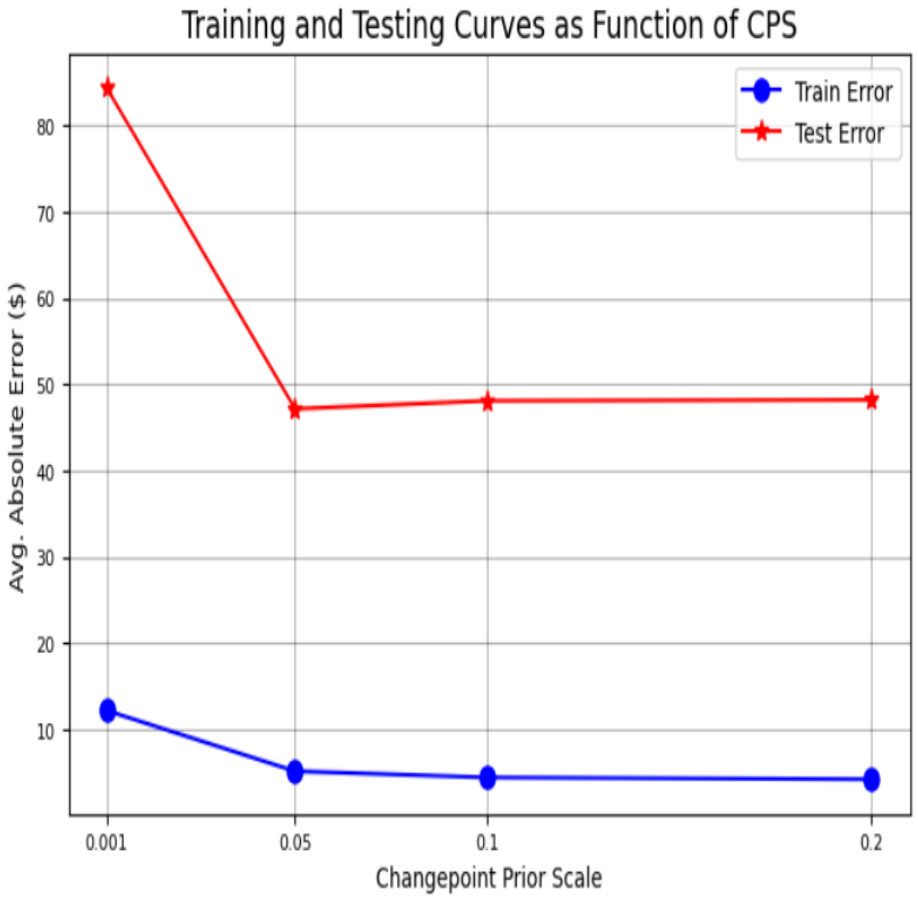


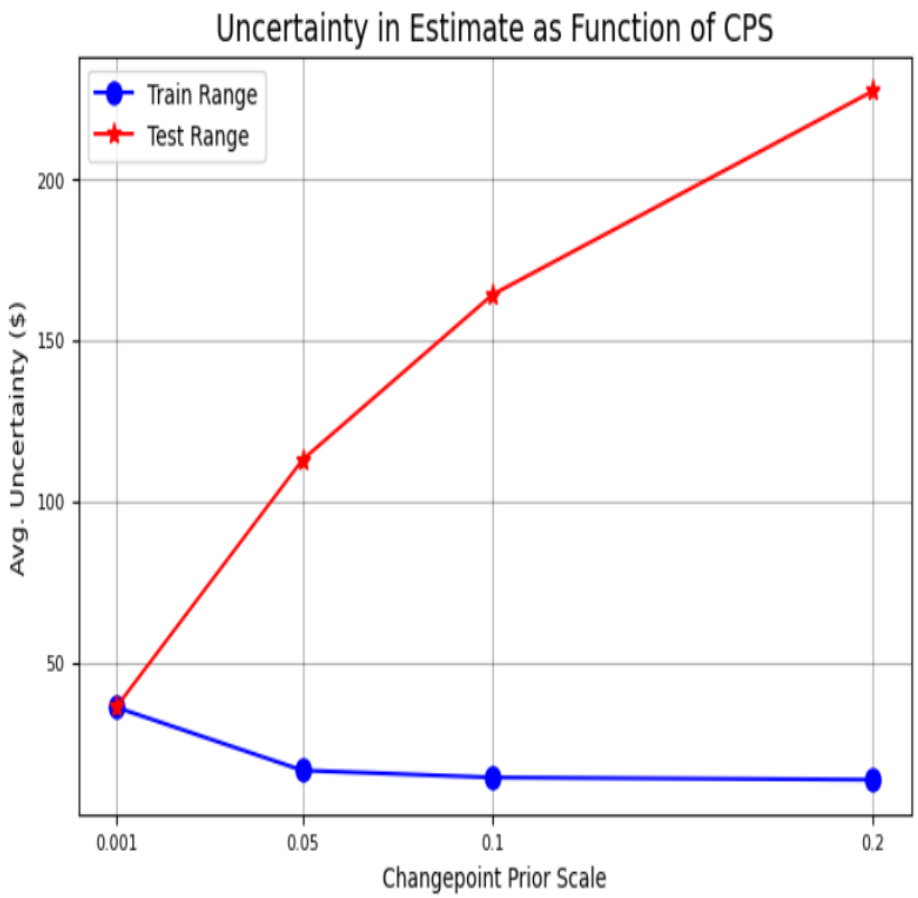
Parameters Tuning

- 使用`changepoint_prior_validation()`對此模型進行數值評估：
  - 訓練誤差training error
  - 訓練範圍training range (信心區間)
  - 測試誤差testing error
  - 測試範圍testing range (信心區間)
- 透過下圖顯示，隨著changepoint_prior_scale之值越大train error越低且training data的uncertainty也越低；而changepoint_prior_scale之值越大，降低test error，但testing data的uncertainty越趨上升。

```
stock.changepoint_prior_validation(start_date='2019-12-13', end_date='2020-12-12',
                                   changepoint_priors=[0.001, 0.05, 0.1, 0.2])
```

```
Validation Range 2019-12-13 00:00:00 to 2020-12-11 00:00:00.

      cps  train_err  train_range   test_err  test_range
0   0.001  12.182030    36.125320  84.406083   36.118652
1   0.050   5.151109    16.550593  47.185801  112.925216
2   0.100   4.426912    14.354172  48.119678  164.188232
3   0.200   4.214696    13.663605  48.222283  227.325023
```

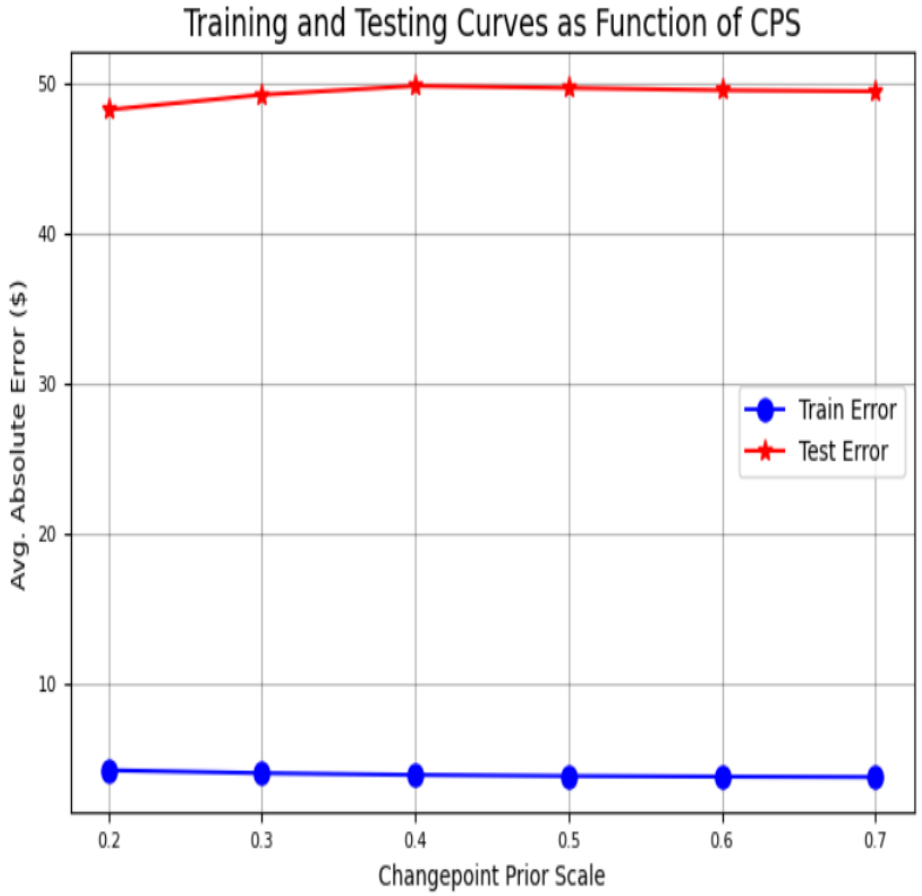


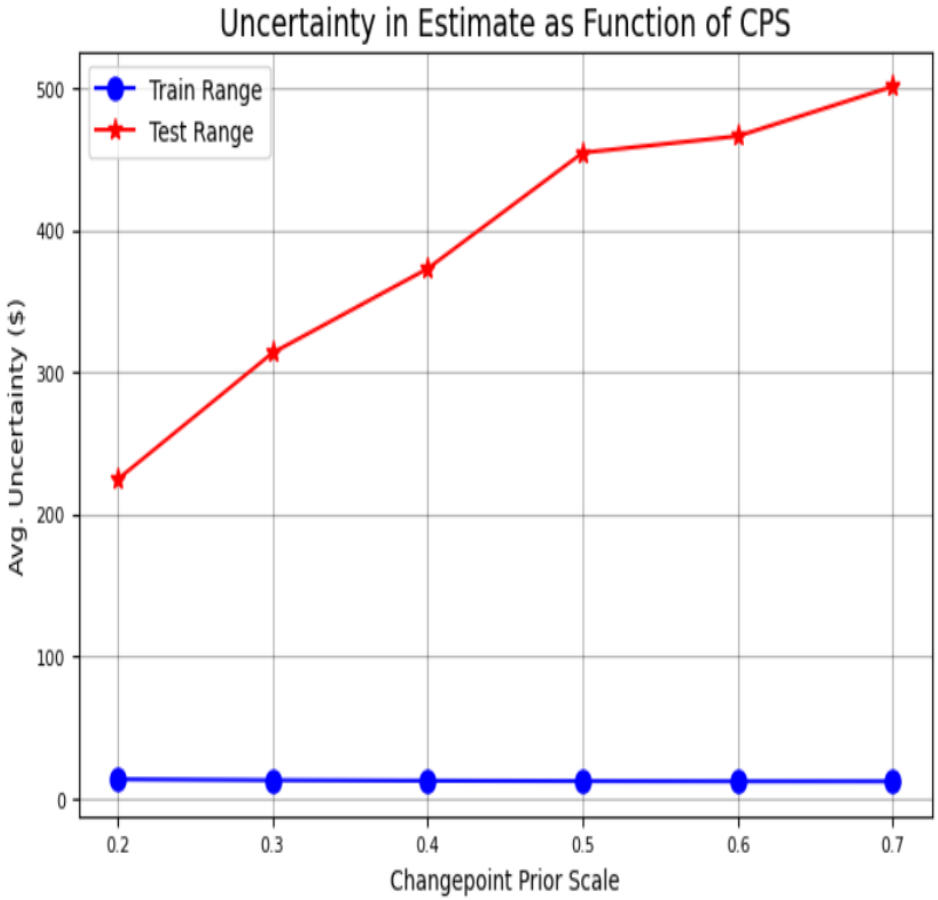


```
stock.changepoint_prior_validation(start_date='2019-12-13', end_date='2020-12-12',
                                   changepoint_priors=[0.2, 0.3, 0.4, 0.5, 0.6, 0.7])
```

```
Validation Range 2019-12-13 00:00:00 to 2020-12-11 00:00:00.
```

	cps	train_err	train_range	test_err	test_range
0	0.2	4.214696	13.644523	48.222283	224.532783
1	0.3	4.024882	12.979843	49.222478	313.683315
2	0.4	3.907530	12.591617	49.812885	372.987413
3	0.5	3.832927	12.357641	49.675163	454.591740
4	0.6	3.789867	12.222704	49.522126	466.232566
5	0.7	3.762327	12.182127	49.448852	501.244947

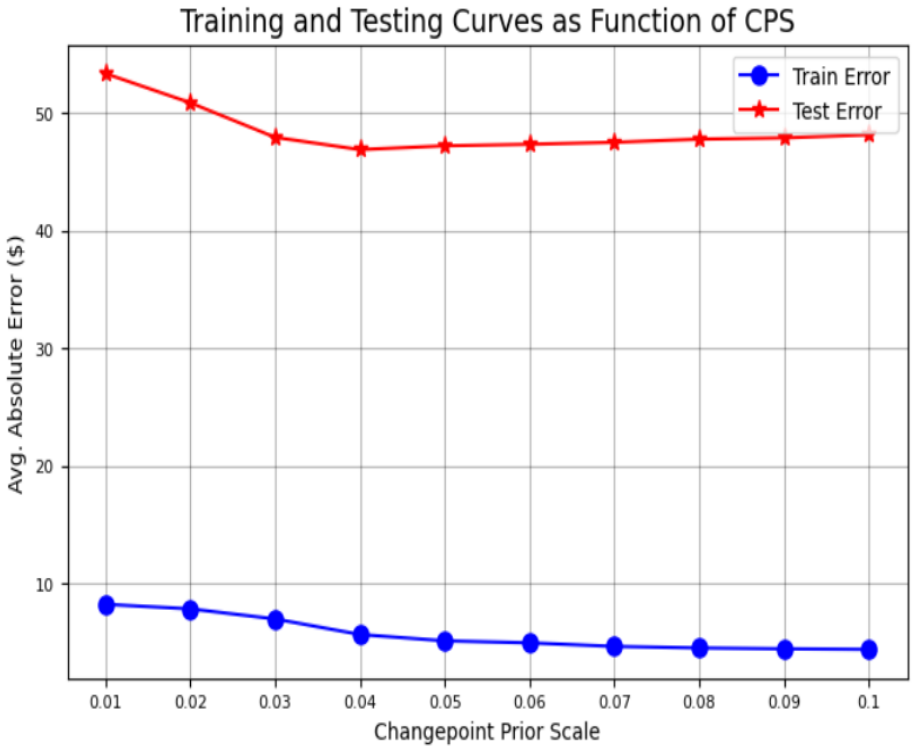


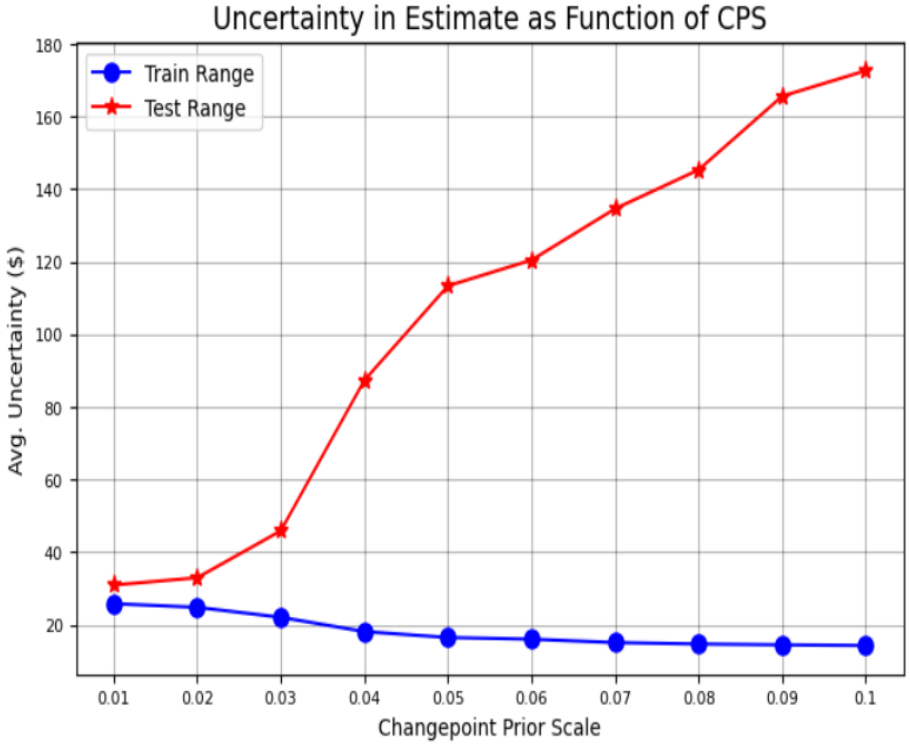


```
stock.changepoint_prior_validation(start_date='2019-12-13', end_date='2020-12-12',
                                   changepoint_priors=[0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1])
```

```
Validation Range 2019-12-13 00:00:00 to 2020-12-11 00:00:00.
```

	cps	train_err	train_range	test_err	test_range
0	0.01	8.253699	25.846606	53.325913	30.969211
1	0.02	7.865408	24.833043	50.842867	32.993463
2	0.03	7.011641	22.095966	47.905493	45.909768
3	0.04	5.689161	18.158374	46.863948	87.350618
4	0.05	5.151109	16.526496	47.185801	113.353032
5	0.06	4.981911	16.065955	47.316847	120.457847
6	0.07	4.678973	15.130887	47.485591	134.588866
7	0.08	4.547827	14.743411	47.753249	145.321584
8	0.09	4.473469	14.515838	47.860781	165.589736
9	0.10	4.426912	14.329942	48.119678	172.615011





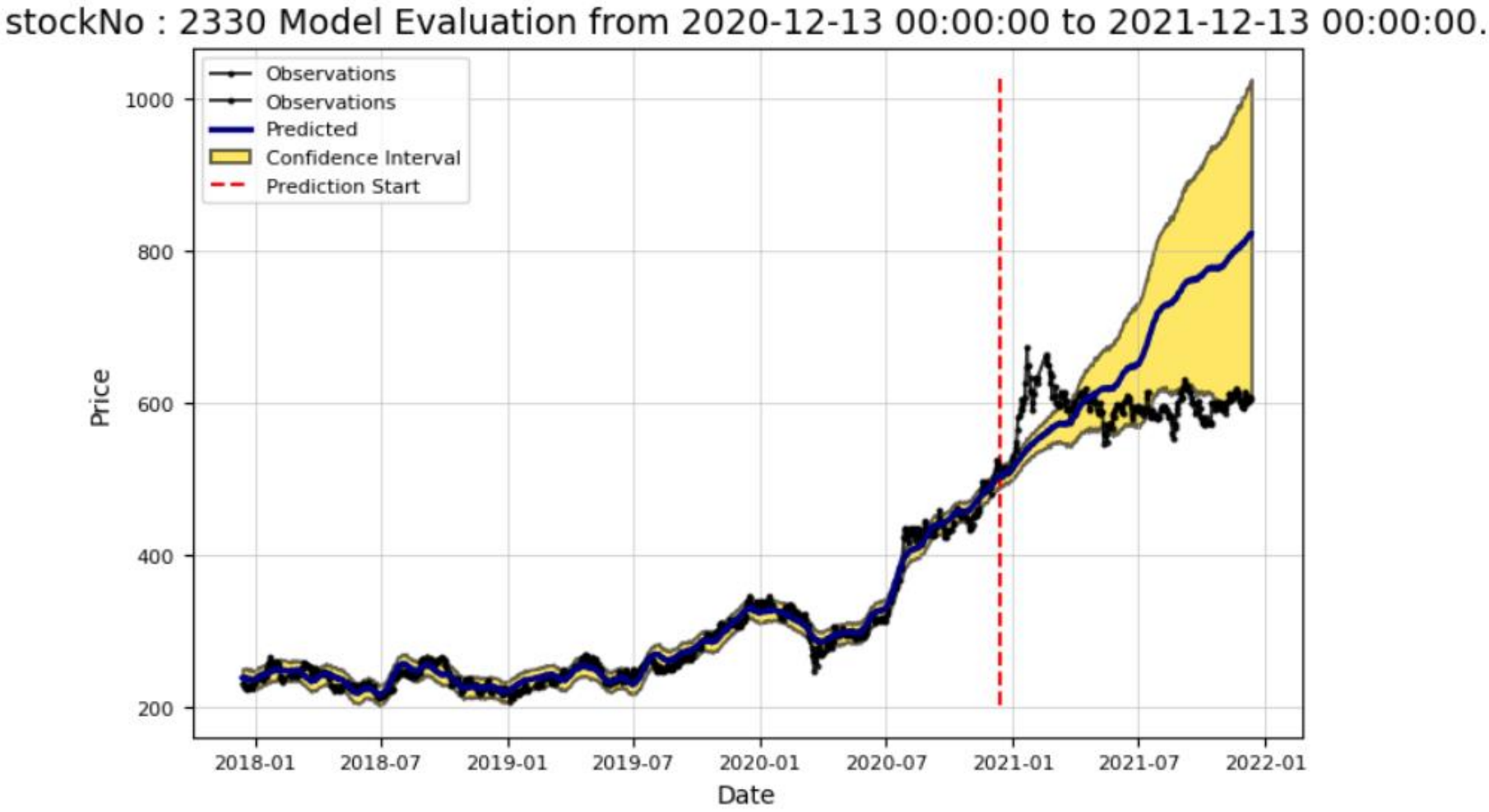
Result Analysis

- 經過測試後，設定changepoint_prior_scale = 0.04，並以此優化後之模型來預測未來10日之股價。。

```
stock.changepoint_prior_scale = 0.04
stock.evaluate_prediction()
```

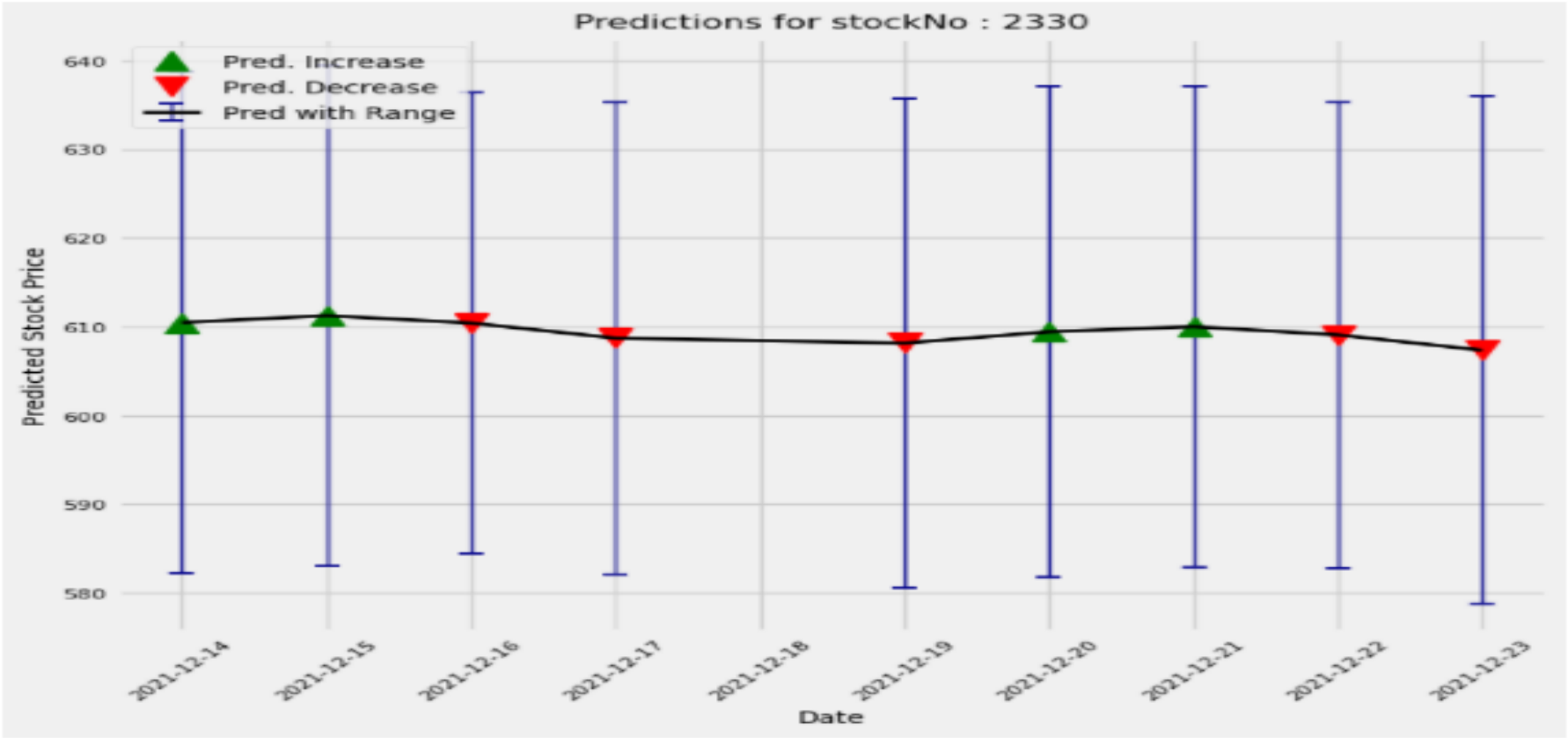
```
Prediction Range: 2020-12-13 00:00:00 to 2021-12-13 00:00:00.

Predicted price on 2021-12-11 00:00:00 = $822.05.
Actual price on    2021-12-10 00:00:00 = $605.00.

Average Absolute Error on Training Data = $8.19.
Average Absolute Error on Testing  Data = $94.92.

When the model predicted an increase, the price increased 48.69% of the time.
When the model predicted a  decrease, the price decreased  55.77% of the time.

The actual value was within the 80% confidence interval 43.03% of the time.
```

Result Analysis



```
stock.predict_future(days=10)
```

```
Predicted Increase:

         Date    estimate    change       upper       lower
1 2021-12-14  610.465624  1.610760  625.108351  596.842149
2 2021-12-15  611.250552  0.784929  625.326203  597.099401
7 2021-12-20  609.425433  1.274780  623.477792  595.819866
8 2021-12-21  610.006566  0.581133  622.849897  595.810524

Predicted Decrease:

          Date    estimate    change       upper       lower
3   2021-12-16  610.414040 -0.836512  623.696171  597.692376
4   2021-12-17  608.710323 -1.703717  622.937907  596.312737
6   2021-12-19  608.150652 -0.559671  622.068345  594.556412
9   2021-12-22  609.080617 -0.925949  622.620382  596.347270
10  2021-12-23  607.348598 -1.732019  621.736206  593.109490
```



Conclusion

- Prophet是由Facebook開發之資料庫，是專為單變數時間序列資料的資料自動化預測而設計。
- 可透過如何擬合Prophet模型，並使用模型進行樣本內及樣本外之預測。
- Prophet確實是進行快速準確的時間序列預測的好選擇！

Reference

- https://weikaiwei.com/finance/stocker/
- http://hn28082251.blogspot.com/2019/05/python-stocker-new-session1.html
- https://towardsdatascience.com/stock-prediction-in-python-b66555171a2
- https://op8867555.github.io/posts/2018-05-29-facebook-prophet.html