資料分析與學習基石 Project A 報告

F74076213 吳定洋

一、 前言

在當今社會中，許多人都說不能只靠主動式收入，要想辦法去增加自己的被動式收入，大家想方設法去嘗試各種投資，股票、權證、期貨、虛擬貨幣、NTF，各種形形色色的投資產品，這當我將選用股票為目標。

在投資中如果像要從中穩定獲利，又常常需要投入極大的時間去研究，所以我想要蒐集資料，並建構模型預測股市走向，從歷史股市資料中判斷出未來幾天漲跌趨勢，雖然可能沒辦法 100%準確，但希望對於投資者來說有個參考標的，提高投資報酬率。

二、 資料收集

原本我的目標是要從 TWSE 臺灣證券交易所爬取長榮從民果 95 年開始的資料，然而在開始實行時發現該網站最早只民國 99 年開始。之後開始爬取資料又發生一個奇怪的問題，爬取時有時候會連不上該網站，發現可能是用同一個網路 ip 連續爬取多筆資料時會發生問題，如果我一次將 2010 年到 2022 年的資料爬取，會一直 error，所以我變成是 1 個月 1 個月抓資料並將需要用到的 column 保留後存成該年該月的 csv，最後在將全部的 csv 合成成一份。

三、 資料前處理

第一步，將資料 Date 從民國改成西元年，因為 pandas 的 to_datetime 不能用民國年。完成後資料欄位如下。

	Date	Open	High	Low	Close	Volume	UpAndDown
0	2010-01-04	17.95	18.25	17.80	18.15	3,629	0.3
1	2010-01-05	18.35	18.85	18.35	18.80	5,557	0.65
2	2010-01-06	18.95	19.00	18.40	18.50	3,031	-0.3
3	2010-01-07	18.50	19.10	18.40	18.60	5,313	0.1
4	2010-01-08	18.70	19.80	18.65	19.60	6,358	1
...	...	...	...	...	...	...	...

圖 一 csv 檔更改 Date 後資料欄位

第二步，畫出 K 線圖、seasonal、trend 圖觀察趨勢。

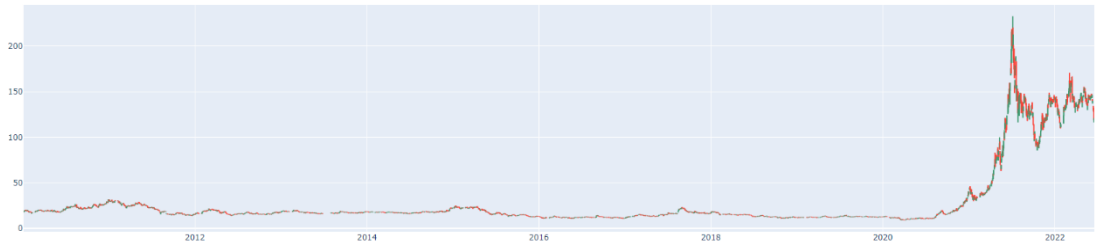

圖 二 K 線圖

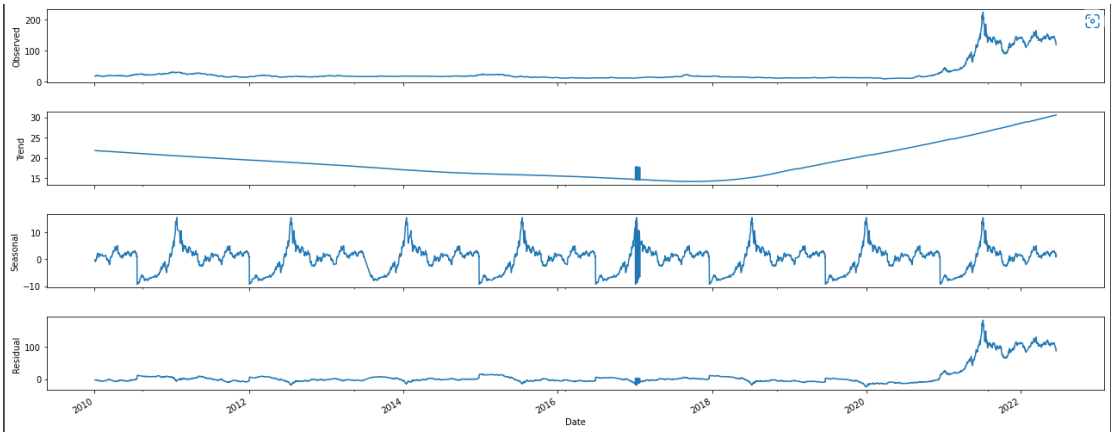

圖 三 趨勢圖

從 K 線可以直接看出很明顯有在 2021 年左右股價急遽增長，大概是跟疫情有極大的相關。在後續處理選擇用來訓練的 data 上要注意，留到後面說明。

第三步，加入更多的 feature 以利之後使用，我加入了 9 日指數平滑移動平均線(EMA)、5 日、10 日、15 日、30 日的簡單移動平均線（SMA）、相對強弱指標(RSI)、指數平滑異同移動平均線(MACD)等常見的指標。

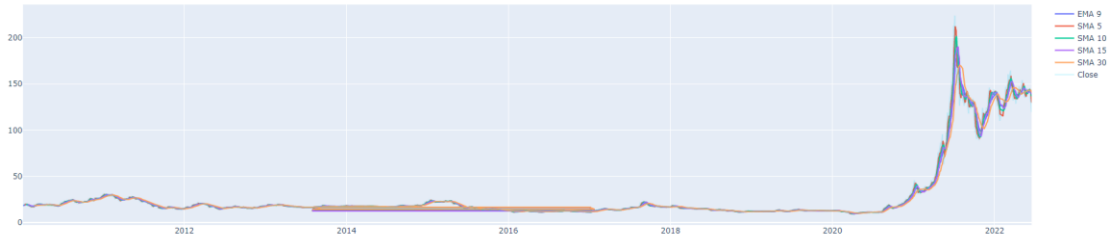

圖 四 各種均線

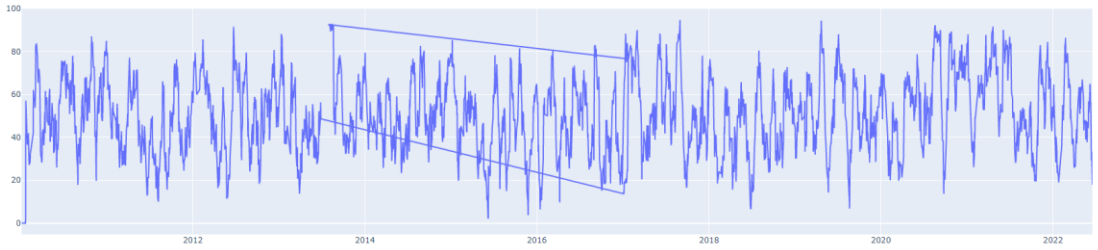

圖 五 RSI

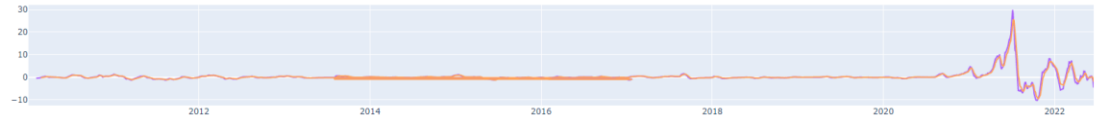

圖 六 MACD

以下是截至目前的 dataframe。

	Date	Open	High	Low	Close	Volume	UpAndDown	EMA_9	SMA_5	SMA_10	SMA_15	SMA_30	RSI	MACD	MACD_signal
0	2010-01-04	17.95	18.25	17.80	18.15	3,629	0.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	2010-01-05	18.35	18.85	18.35	18.80	5,557	0.65	18.150000	NaN	NaN	NaN	NaN	0.000000	NaN	NaN
2	2010-01-06	18.95	19.00	18.40	18.50	3,031	-0.3	18.492105	NaN	NaN	NaN	NaN	0.000000	NaN	NaN
3	2010-01-07	18.50	19.10	18.40	18.60	5,313	0.1	18.495018	NaN	NaN	NaN	NaN	0.000000	NaN	NaN
4	2010-01-08	18.70	19.80	18.65	19.60	6,358	1	18.525545	NaN	NaN	NaN	NaN	0.000000	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3050	2022-06-13	134.00	134.50	130.00	130.50	63,339	-8.5	142.325519	142.5	143.05	142.433333	143.116667	29.230769	-0.954968	-0.035552
3051	2022-06-14	129.00	132.00	128.50	131.50	42,825	1	141.142967	139.9	141.65	142.133333	142.866667	33.333333	-1.599668	-0.348375
3052	2022-06-15	131.50	132.50	128.00	128.00	51,324	-3.5	140.178671	137.3	140.40	141.300000	142.416667	27.941176	-2.365748	-0.751849
3053	2022-06-16	130.00	130.50	119.00	119.50	91,675	-8.5	138.960804	133.8	139.15	140.366667	141.800000	22.891566	-3.617055	-1.324891
3054	2022-06-17	118.50	121.00	116.50	119.50	45,646	0	137.014723	129.7	136.65	138.800000	140.733333	17.948718	-4.556204	-1.971153

3055 rows × 15 columns

圖 七 加入各指標後的 dataframe

第四步，將我 Close(收盤價)往前一天 shift，如上圖 2010-01-04 的 Close 會變成 18.80，因為我股價預測是要預測未來的股價，我要在 2010-01-04 預測隔天的顧價，所以必須要將我要當作 y_pred 的資料往前移一天，以便我模型訓練。此外有許多的 column 是 NaN，因為指標需要用到前幾天的資料，例如 2010-01-04 的 5 日均線要用到更之前五天的資料，那是我沒有的，所以這天的 5 日均線為空，其他的指標也是如此，因此，這邊用最簡的方法，直接把有空的資料的

row 刪除掉。

第五步，將資料分成 train、test、valid，這邊不能夠分成 70%、15%、15%，因為 2021 後股價暴漲太誇張，如果後面 15%是 test，百分之百會預測出不堪入的結果。我認為 train 必須要包含到股價暴漲後的 data，才行，而 test 我將使用 2022 年 5 月中後的資料。

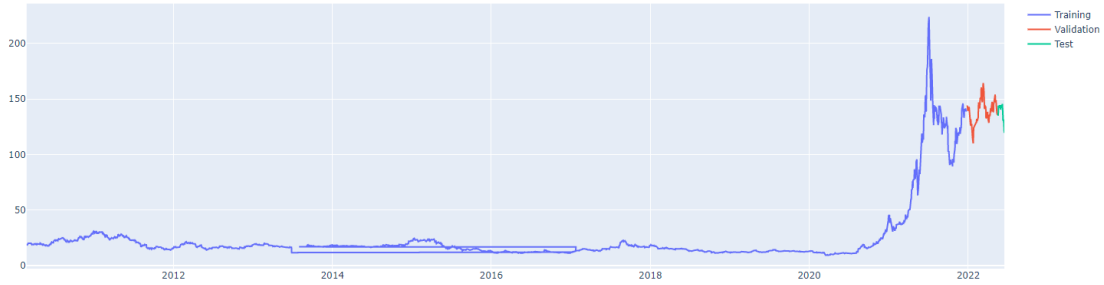

圖 八 第一次嘗試 data split

四、 第一次訓練嘗試

我將使用 xgboost 的 XGBRegressor，進行模型預測，此外我也運用了 HW3 學到的 Grid Search 找取最好的超參數。

所有的 feature 如下：

	Open	High	Low	Close	Volume	UpAndDown	EMA_9	SMA_5	SMA_10	SMA_15	SMA_30	RSI	MACD	MACD_signal
3000	137.5	138.5	137.0	136.0	26149	2.0	142.806535	140.0	145.40	144.166667	141.466667	47.572816	-0.539784	0.765914
3001	134.0	136.5	130.0	135.0	60598	-1.5	142.275882	138.6	144.00	144.066667	141.433333	48.039216	-0.958781	0.420975

圖 九 所有 feature

```
parameters = {
        'n_estimators': [100, 200, 300, 400],
        'learning_rate': [0.001, 0.005, 0.01, 0.05],
        'max_depth': [8, 10, 12, 15],
        'gamma': [0.001, 0.005, 0.01, 0.02],
        'random_state': [42]
}

eval_set = [(X_train1, y_train1), (X_valid1, y_valid1)]
model = xgb.XGBRegressor(eval_set=eval_set, objective='reg:squarederror', verbose=False)
clf = GridSearchCV(model, parameters)

clf.fit(X_train1, y_train1)

print(f'Best params: {clf.best_params_}')
print(f'Best validation score = {clf.best_score_}')
```

圖 十 grid search

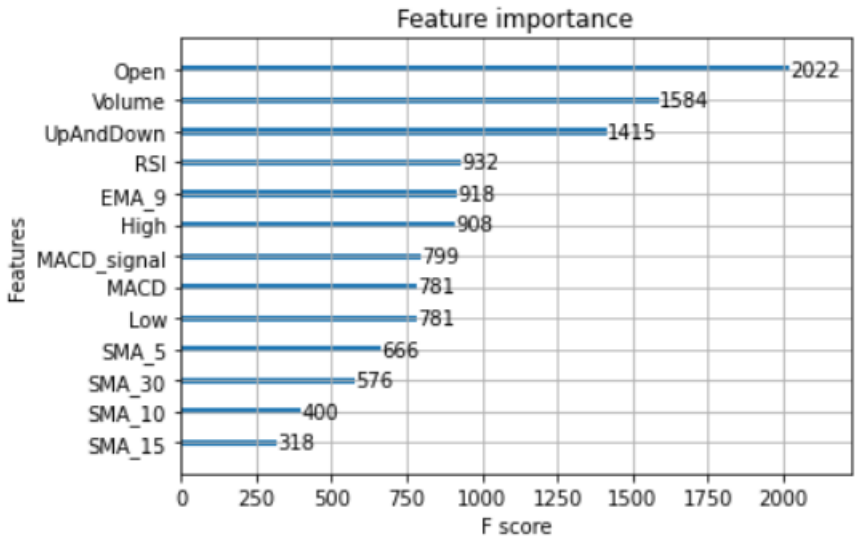

圖 十一 各 feature 在此預測中的重要性

預測結果如下。

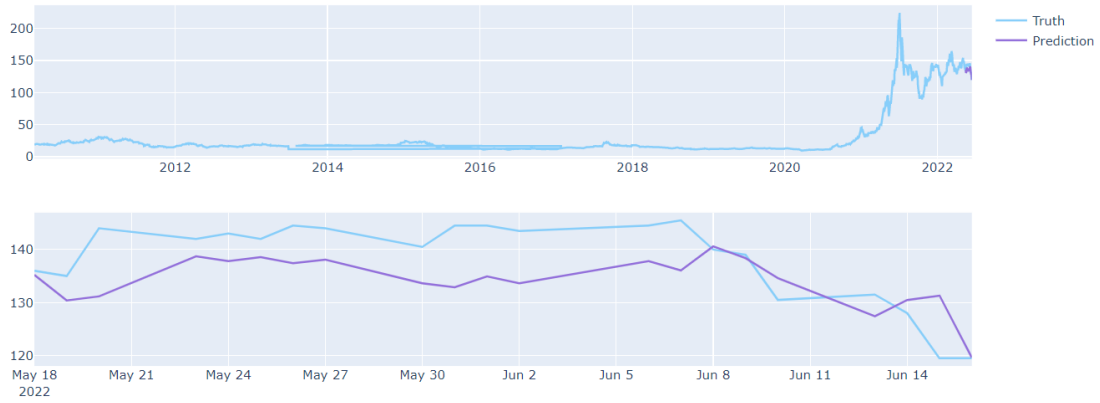

圖 十二 第一次嘗試預測結果

在精確度上的 mean_squared_error = 47.32192415643213、Best validation score = 0.70132211234。雖然沒辦法準確月測股價，但在大趨勢下，六月多會下跌的趨勢有預測出來。

五、 第二次訓練嘗試

做想到小組的報告 Store Sales - Time Series Forecasting，那份 data 也是 Time Series 的預測，在訓練模型時有嘗試過全部的 train 資料帶入訓練，和使用時間較接近 test 時間的部分資料進行訓練，發現後者結果比較好，因此我將嘗試只使用 2021 年後的資料進行訓練。

這次使用 2021 年後的資料，85%train、15%valid、5%test。

圖 十三 第二次嘗試 data split

訓練的模型預測結果如下。

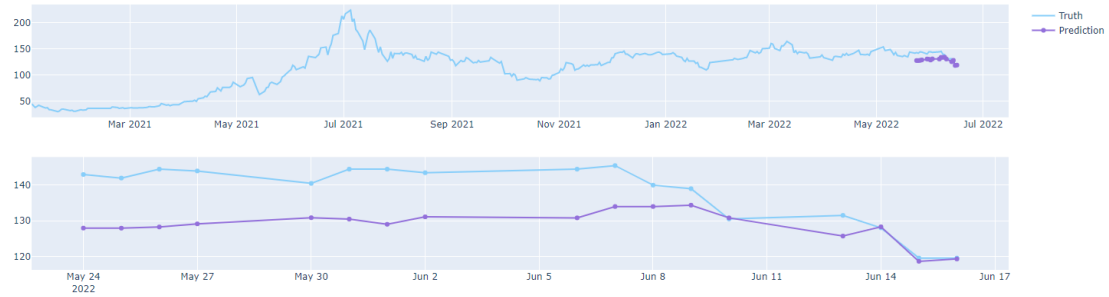

圖 十四 第二次嘗試圖

在後半段的預測比前次更為精確，但大趨勢卻變得比較難觀察出來。而且準確度也下降很多。mean_squared_error = 118.18366555907585。Best validation score = -1.591637674252826。

本來以為依據小組報告的經驗，如果 time Series 的資料在某年之後有突然的劇烈變化，應該將該年之前資料捨棄，而只使用劇烈變化之後較平穩的資料，結果發現可能不是那麼簡單，應該還有其他隱藏在背後的因素，而非直觀的看起來劇烈變化。

六、 第三次訓練嘗試

這次我將減少 feature 數量只留下以下 feature 做訓練。

	Close	EMA_9	SMA_5	SMA_10	SMA_15	SMA_30	RSI	MACD	MACD_signal
3000	136.0	142.806535	140.0	145.40	144.166667	141.466667	47.572816	-0.539784	0.765914
3001	135.0	142.275882	138.6	144.00	144.066667	141.433333	48.039216	-0.958781	0.420975

圖 十五 第三次嘗試 feature

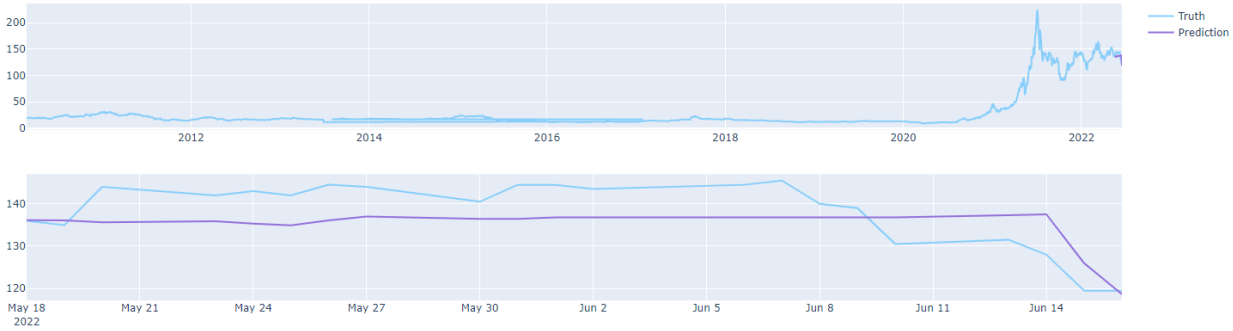

圖 十六 第三次嘗試預測結果

預測結果如下：

在精確度上的 mean_squared_error = 41.79040244124398、Best validation score = 0.6468560004694961，結果還是不大理想。

七、 繼續篩選 feature 做嘗試

目前得到最好的 feature 選擇是全部使用，只將 Volumn 移除。

	Open	High	Low	Close	UpAndDown	EMA_9	SMA_5	SMA_10	SMA_15	SMA_30	RSI	MACD	MACD_signal
3000	137.5	138.5	137.0	136.0	2.0	142.806535	140.0	145.40	144.166667	141.466667	47.572816	-0.539784	0.765914
3001	134.0	136.5	130.0	135.0	-1.5	142.275882	138.6	144.00	144.066667	141.433333	48.039216	-0.958781	0.420975

圖 十七 目前最好 feature 選擇

預測結果如下：

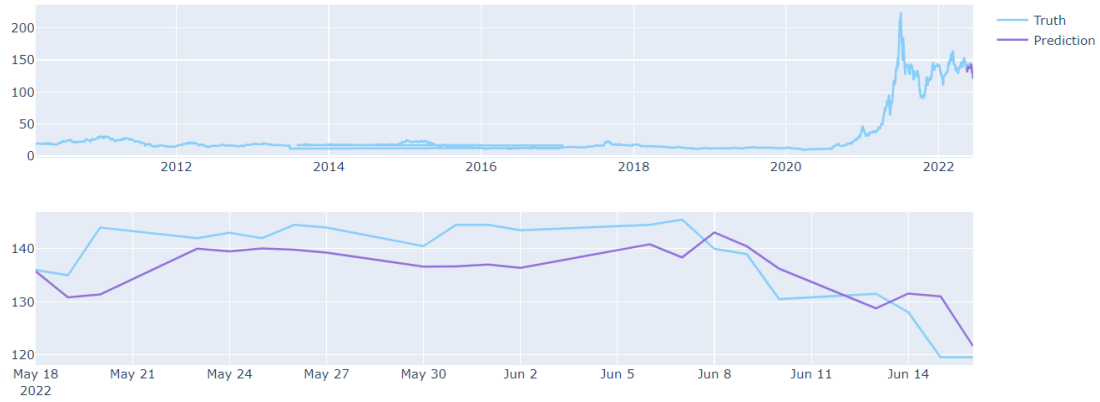

圖 十八 目前最好預測結果

mean_squared_error = 32.866894513789916、Best validation score = 0.7247024366629243，是截至目前我在 feature 上選擇的訓練結果最好的。

八、 問題討論

在做這股價預測模型，我有更加了解 Time Series 的操作，首先是，如果選用的資料集在某個時間段有劇烈變化，可以考慮把它拿掉，或是把劇烈變化後漸趨穩定的資料集也概括進到 train 資料中，至於要拿掉或是使用我認為可能必須要看 trend，還有資料的數量，如果 trend 為穩定持續的向上，且劇烈變化後趨於穩定的資料夠多，那應該是可以把劇烈變化之前的資料不納入訓練，但是以上條件如果有一個沒有達成，那可能需要把加入更多資料才行，畢竟還有 seasonal，這有週期較為穩定的隱藏 feature 在其中。

這次報告我只使用 xgboost 的 XGBRegressor，因為在作業 3 的時候，我勝過助教給分數也是用 XGBRegressor，而且使用上非常方便，是一個使用上很簡單且結果相較於其他簡單 Regressor 中，最令人滿意的。在後續團體報告中，也是 XGBRegressor 讓我們預測結果有所進步。

九、 後續目標

在股市預測中，我認為除了我這次用到的各種 feature 之外，大盤的考慮、國際局勢動盪、各種財經相關新聞，都很重要，如果能夠將這些都變成 ferture 來訓練模型，我相信能夠大大提升預測結果，如果後續我要繼續製作股市預測模型，我會著重將國際局勢動盪、各種財經相關新聞變成訓練 feature 下手，我認為這兩項是能夠有效解決股市劇烈變化後導致資料變得沒有規律，導致預測差

強人意的破冰船。

還有自動化的訓練模型也是我如果後繼續製作想要做的，做成像是 APP 那樣，我只要在 GUI 介面輸入我要的公司股票代號，APP 就會開始幫我跑預測，在預測中著重使用不同 feature，並且回傳多種預測結果。但首先，我可能必須要找到穩定，且能提供我股市歷年資料的網站，這次我選用的網站有一些奇怪的問題，導致我要自動化擷取一堆股市資料會有問題。

一〇、 結論

這次報告讓我更加瞭解要製作股市預測模型應該著重從哪邊下手，讓我更加熟悉 Time Series 的資料該如何操作，也提供我一個機會做我一直以來想嘗試的用機器學習或深度學習預測股價，還有應用各種資料前處理，怎麼從現有的資料中生出更多可用資料(RSI、EMA…)，希望後續能夠繼續完善這樣的作品。