

Python 資料分析與機器學習應用
期末報告

利用機器學習建立最佳投資組合

第 N 組

許智評、林姝延、賴宇辰、王奕翔、曾國睿

目錄

壹、研究動機及目標	P.3
貳、資料來源及蒐集方式	P.4
參、研究工具	P.4 – P.5
肆、研究流程	P.5 – P.6
伍、模型介紹及研究方法	P.6 – P.12
陸、預期產出	P.12 – P.33
柒、結果討論	P.33 – P.35
仈、研究限制	P.35
玖、研究時程安排	P.35 – P.36
拾、小組分工	P.36 – P.37

研究摘要:

我們的研究主軸為利用機器學習技術建立最佳股票投資組合，以幫助投資者更有效的實現更高投資回報。首先，我們收集和整理股票市場相關的資料，包括股票價格、公司財報資訊、技術指標和新聞等。接著，我們建立了股票選股模型，預測股價走勢並根據預測結果建立投資組合。最後，我們會透過實證研究和模擬交易，評估各方法及模型的投資回報和風險管理能力。我們的研究產出有望為投資者提供有效的投資決策依據，並幫助投資者在股票市場中實現更好的回報並有效管理風險。

壹、研究動機及目標

股票市場是許多投資人的投資途徑，然而，對大多數投資人而言，如何選擇最佳的投資組合以實現投資目標卻是很大的挑戰。傳統的投資方法通常是基於人為的判斷和決策，需要花費大量時間和精力。而機器學習的興起為股票投資提供了一個新的解決方案，可以透過優化選股的過程，幫助投資者更有效的達到更好的報酬。因此我們希望透過這門課所學落實機器學習在股票投資中的應用，以各模型預測特定公司股價上漲或下跌幅度，找出能得出最佳投資組合的模型，以提供投資人決策建議。

我們的研究主要在探索如何利用機器學習課堂所學，建立最佳股票投資組合。具體目標包含：

- 收集和整理股票市場相關的資料，包括股票價格、財務報表、技術指標、新聞等消息面指標等。
- 建立股票選股模型，利用監督式學習方法分析和評估不同的股票選股策略。
- 找出最佳的股票投資組合。
- 進行實證研究和模擬交易，評估機器學習模型的投資回報和風險管理能力。

貳、資料來源及蒐集方式

1. 基本面相關公司資訊:

於TEJ 財經資料庫取得過去各公司歷史股價、公司資產負債表、損益表及現金流量表等財報資訊，以評估特定公司的績效和市場前景。

2. 技術指標資訊:

於三竹股市、Goodinfo! 台灣股市資訊網、Yahoo 奇摩股市、TWSE 臺灣證券交易所等平台或資料庫透過爬蟲或手動方式取得技術面模型所需資料。

3. 消息面資訊:

使用人工方式從網路上蒐集社交媒體上的討論和評論、博客文章、網路論壇等資訊，同時透過不同投資經歷及方式的投資人獲得更多的市場情報和觀點，進一步分析特定公司股價趨勢。並從台灣info此網站確認股價走勢等資訊。

參、研究工具

我們的研究工具涵蓋了模型方法和各模組，這些工具可以幫助我們更有效地進行模型建構和股價趨勢預測。

- 模型:

- Random Forest (隨機森林):

是一種集成學習(Ensemble Learning)方法，用於解決監督式機器學習問題，特別是在分類和回歸任務中表現出色。Random Forest由多個決策樹組成，每個決策樹都是獨立且隨機生成的。它通過隨機選擇特徵子集和訓練數據的子集來建構每個決策樹，這種隨機性使得每個決策樹都具有差異性。當進行預測時，Random Forest結合了所有決策樹的預測結果，通過投票或取平均值來得出最終預測結果。

- 多元回歸模型:

多元邏輯回歸是指有多個自變量(即特徵)的邏輯回歸模型，使用的是邏輯函數(sigmoid 函數)來將線性組合的結果映射到 0 到 1 之間的概率值。在多元邏輯回歸中，線性組合不再是單個自變量的加權和，而是多個自變量的加權和。

- XGBoost 模型:

XGBoost 模型是一種集成學習方法，基於決策樹進行建模，並利用 Gradient Boosting 迭代方法降低當前模型的誤差。通過擬合殘差的方

式，將學習器的預測結果和實際值之間的差距進行最小化，以提高準確性。

- SVR模型:

SVR (Support Vector Regression) 是 SVM (Support Vector Machine) 在回歸問題上的一個變體。SVM 是一種監督式機器學習算法，主要用於分類問題。然而，SVM 也可以用於回歸問題，而這種回歸版本的 SVM 就是 SVR。SVR 通過將回歸問題轉化為一個優化問題，以找到一個最佳的超平面來進行回歸預測。它使用了一個稱為「支持向量」的訓練數據點集合，這些數據點位於超平面的邊界上。SVR 的目標是使這些支持向量與實際目標值之間的誤差最小化。同時，還引入了一個稱為「容忍度」的參數，用於控制支持向量與超平面之間的距離。

- 套件 / 模組:

- Pandas套件: 進行數據處理及數據清理。
- Scikit-learn套件:
 - RandomForestRegressor: 用來找出特徵給模型使用。
 - SVM: 用在 SVM 分析中。
 - LinearRegression: 用在邏輯回歸模型之中。
- XGBoost 套件: 使用在 XGBoost 的模型中。

肆、研究流程

以下為我們的研究流程：

1.選擇產業並蒐集資料_基本面 & 技術面

1-1 選定五個產業

1-2 從各產業中選擇兩間公司

1-3 蒐集公司相資料，包括財務報表、市場趨勢、技術指標、投資消息等

2. 建立模型_基本面 & 技術面

2-1 確定適當的機器學習模型

2-2 設計特徵選擇方法和模型評估指標

3.訓練、驗證及測試模型_基本面 & 技術面

3-1 將資料分為訓練集、驗證集和測試集

- 3-2 調整模型參數以達到最佳效果
- 3-3 以訓練及訓練模型
- 3-4 在測試集上評估模型的表現
- 4.根據模型預測建立投資組合_基本面 & 技術面
 - 4-1 基於模型預測建立最佳投資組合
- 5.進行投資_基本面 & 技術面 & 消息面
 - 5-1 根據投資組合進行投資
 - 5-2 監控投資組合表現
- 6.模型成效比較_基本面 & 技術面 & 消息面
 - 6-1 比較各模型表現
 - 6-2 呈現模型成效
 - 6-3 提供投資建議

伍、研究方法及模型介紹

下列為我們進行投資的設定:

- 研究初始設定投資期間為 2023 年 3 月到 5 月
- 起始資金為 10 萬新台幣
- 最多僅能進行一次買進與一次賣出
- 選定以下 5 種產業, 各產業分別選擇 2 間公司作為投資標的選項
 - 半導體產業
 - 台積電 (2330)
 - 聯電 (2303)
 - 電子零組件產業
 - 宏碁 (2353)
 - 南電(8046)
 - 金融業
 - 華南金(2880)
 - 富邦金(2881)
 - 通訊產業

- 台灣大(3045)
- 華碩 (2357)
- 生技醫療產業
 - 高端疫苗 (6547)
 - 神隆 (1789)

我們將採用基本面、技術面和消息面三種預測方法，以預測未來股票價格和建立最佳投資組合。基本面分析將透過分析公司財報來預測公司股價。技術面分析將透過分析過去股票價格的走勢和各種技術指標來預測未來價格走勢。消息面分析將參考五種不同的消息來源，並進行實測，以測試獲利情況並與其他兩種方法進行比較。詳細介紹及方法如下所述：

1. 基本面預測：

透過分析公司的財報，了解其財務和經營狀況來評估或預測公司價值，並決定投資組合。

首先自 TEJ 財經資料庫取得各公司的股價及涵蓋資產負債表、損益表及現金流量表的綜合財報資料，再進行資料合併和清理。接著建立模型，預測股價並計算模型的準確率。最後，根據預測結果買進前三高的個股和放空預期下跌的個股，建立投資組合。

研究方法詳述如下：

- 資料清理

為了提升和確保數據品質，在進行模型訓練前先進行以下資料處理：

- 多餘欄位處理：刪除股價相關的多餘欄位，僅保留 15 日均價
- 數據時間處理：根據日期欄位進行分組，並求算 3、6、9、12 月的平均股價
- 財報資料缺失值處理：對於缺失值超過 40% 的欄位，刪除整欄以避免對後續分析及模型訓練造成影響
- 資料清理：財報數據僅保留資料型態為整數或浮點數的欄位
- 資料合併，將股價資料以及財報資訊按照公司及年月進行合併
- 新增列：將每公司最後新增一行，用以進行 2023 年 6 月的股價預測
- 移動列：為了使用前一季的財報資料來預測下一季度的股價，因此將各公司財報資訊後移一行，並將欄位名稱加上 “_shifted”

- 相關性處理

- 找出並刪除與目標變數 y 相關性過低的變數: 設定閾值為 0.2, 刪除和 y 相關性絕對值 < 0.2 的變數, 幫助簡化模型並避免不必要的特徵。此作法可以降低模型噪音, 避免 overfitting, 並提高模型的預測性能和解釋能力。
- 避免多重共線性: 設定閾值為 0.7, 找出兩特徵相關性在 0.7 以上的變數 pairs, 並刪除 pairs 中較不重要, 也就是與 y 相關性較低的變數。因為財報數據較容易出現多重共線性的情形, 此做法可以避免多重共線性過度影響模型表現, 提高模型穩定度及解釋能力, 同時幫助簡化模型, 提高模型訓練、測試及預測效率。
- 繪製變數相關性熱力圖: 視覺化呈現各變數間的相關性
- 模型應用:
 - 利用多元回歸模型、XGBoost 模型、SVR 模型對時間公司的股價及財報在 2013 年至 2023 年 3 月的數據進行機器學習, 並預估未來一季股價。模型訓練完成後, 利用 2023 年第 1 季的財報數據, 對我們的目標投資期間(2023 年 3 月到 5 月)推估未來可能的走勢, 並取漲跌幅作為判斷是否進行買賣的依據。
- 模型訓練:
 - 模型選擇
 - 多元邏輯回歸模型
 - SVM 模型
 - XGBoost 模型
 - 設定 X 、 y : X 為公司財報數據、 y 為股價
 - 切 training / testing dataset: 因為股價及財報皆為時間序列資料, 因此以時間順序將 dataset 切分為訓練及測試子集, 並利用迴圈找出最佳切分比例。
 - 找模型最佳參數: 進行參數調整, 找出最佳模型。

2. 技術面預測:

透過分析過去價格的走勢來預測未來價格走勢, 並以歷史有重現性等來決定是否進行投資。

以爬蟲技術以及有收錄台股資料的第三方資料庫(yfinance、TWStock)取得過去一段時間的開盤價、最高價、最低價、收盤價、MACD、KD、RSI 值、成交量和月均線數據。

透過對上述資料進行資料清理(ta、scikit-learn)和數據切割, 建立、訓練、測試模型, 並判斷各公司的股價未來走勢。最後, 根據預測結果建立最佳投資組合。

研究方法詳述如下:

- 資料清理
 - 數據時間處理: 將訓練集從資料中提取出來, 並與預估區間切開。
 - 透過ta資料庫添加yfinance資料庫沒有之數值: RSI值與KD值。
 - 清除其餘指標: 刪除與本次無關之目標特徵值。
 - 新增dataframe: 將預估值填入新的frame中避免混淆。
 - 分隔各家公司之分析結果: 透過一家公司一個檔案, 避免混淆。
- 模型應用

同基本面, 利用隨機森林(Random forest)模型、SVR 模型、XGboost 模型分別對這九種指標在2022年至2023年2月的數據進行機器學習, 並預估未來60個交易日後的數字。模型訓練完成後, 利用2023年3月到5月前60個交易日的數據, 對我們的目標投資期間(2023年3月到5月)推估未來可能的走勢, 並取漲跌幅作為判斷是否進行買賣的依據。
- 模型訓練:
 - 模型選擇:

隨機森林(Random forest)模型

SVR 模型

XGBoost模型

 - 設定 X、y: X 股票股價、成交量、技術面數據、y 為股價
 - 切 training / testing dataset
 - 找模型最佳參數
- 預估股價視覺化
 - 將預估股價透過折現圖繪製, 並找出3月至5月之最高股價與最低股價。
- 模型比較
 - 將三種模型產出之折線繪於同一張圖, 觀察預估之趨勢。
- 相關性處理
 - 找出與目標變數相關性最高之變數: 將所有特徵值對目標變數之相關性進行作圖, 透過相關性的長條圖, 幫助我們了解模型著重使用那些數據進行預測。

3. 消息面預測:

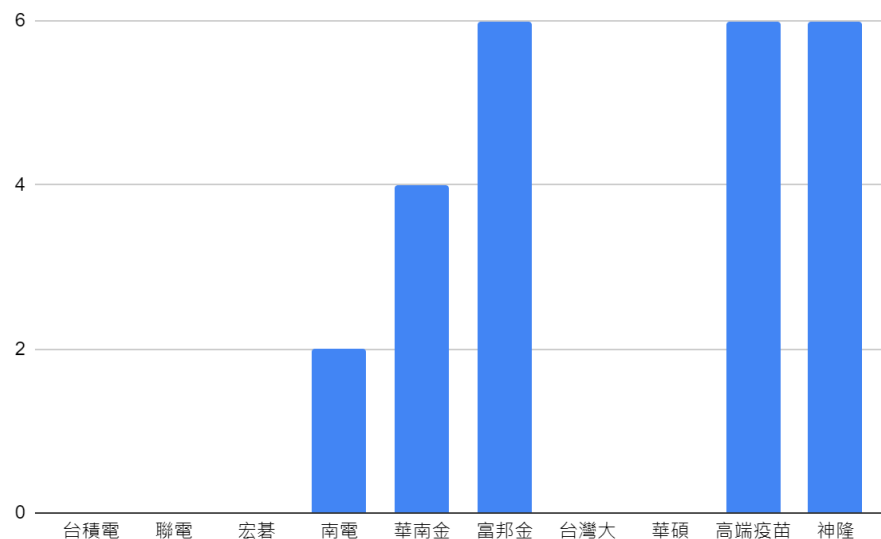
身為一個從未接觸過股市的大學新鮮人，希望學習身邊周遭投資人的經驗與觀點，盼望在預定的投資期間，利用習得的投資技能，預估若實際進行投資的獲利情況。

參考多種不同的消息來源並使用模擬股市的APP進行實測，以測試獲利情況並與其他兩種方法進行比較。我們也將參考周遭投資人的經驗和觀點，以期在預定的投資期間，利用習得的投資技能，預估若實際進行投資的獲利情況。

為符合消息面藉由各方消息以進行股票買賣的原則，最終，決定將股票消息員分成多種面向，吸收各個面向的投資策略，去決定最終的投資組合。

選擇多個面向的向方法如下:

1. 對十家公司進行篩選，淘汰四家較不受各個消息源待見的產業，以下為票數最終出來的結果(一共6票，包括5種消息源及自己)。



以下為各公司被淘汰的原因:

- 神隆(6票):
 - 波動大難以預估
 - 在三月時，一度跌破季線
 - 疫情趨緩
- 高端(6票):

- 半年線一路下探
- 2月份跌破季線, 且季線有下彎趨勢
- 疫情趨緩
- 富邦金(6票):
 - 金控股好在其殖利率高, 並不適合短期股票買賣
 - 富邦金年線自年初以來一路下探
 - 不在0050的金控, 不想關注
- 華南金(4票):
 - 金控股好在其殖利率高, 並不適合短期股票買賣
 - 漲幅不大, 不易從中獲利
 - 財報評分不高

2. 綜合各消息源提供的絕對獲利投資點(3個月提供3次機會), 及**最佳投資點**(3選1)。自行消化後, 統整出最好的投資組合, 進行投資。

消息源	時間點	投資策略
A(大學生投資經歷2年)	5/9(宏碁29、華碩290)	自己畫趨勢線 考量市場情緒(包含觀察FED)
B(業餘投資者)	3/20(台積電510)、5/16(華碩302)	低點的時候買, 別人恐懼我貪婪
C(Line、Decard、Facebook等股票相關資訊)	4/24(台積電508、台灣大103)、5/9(聯電49.8)	毫無規律, LINE的可信度極低, 一旦發現無法詐騙及消失。 Dcard、Facebook內容不可全信以為真
D(職業投資者)	4/24(聯電50、華碩279)、5/02(台積電496.5)、5/16(台積電506)、5/9(宏碁28.8)	合適的選股技巧 找買點 適當停利 寫程式輔助
E(業餘投資者)	5/16(台積電504)、5/9(宏碁28.7)	聽內線股居多, 沒有什麼特別投資策略

由於只有一次買賣點機會，最終選擇5/9(二)進行投資，購買宏碁(28.7)及聯電(49.95)。以9:1分配。於5/24賣出(宏碁31、聯電49.3)。

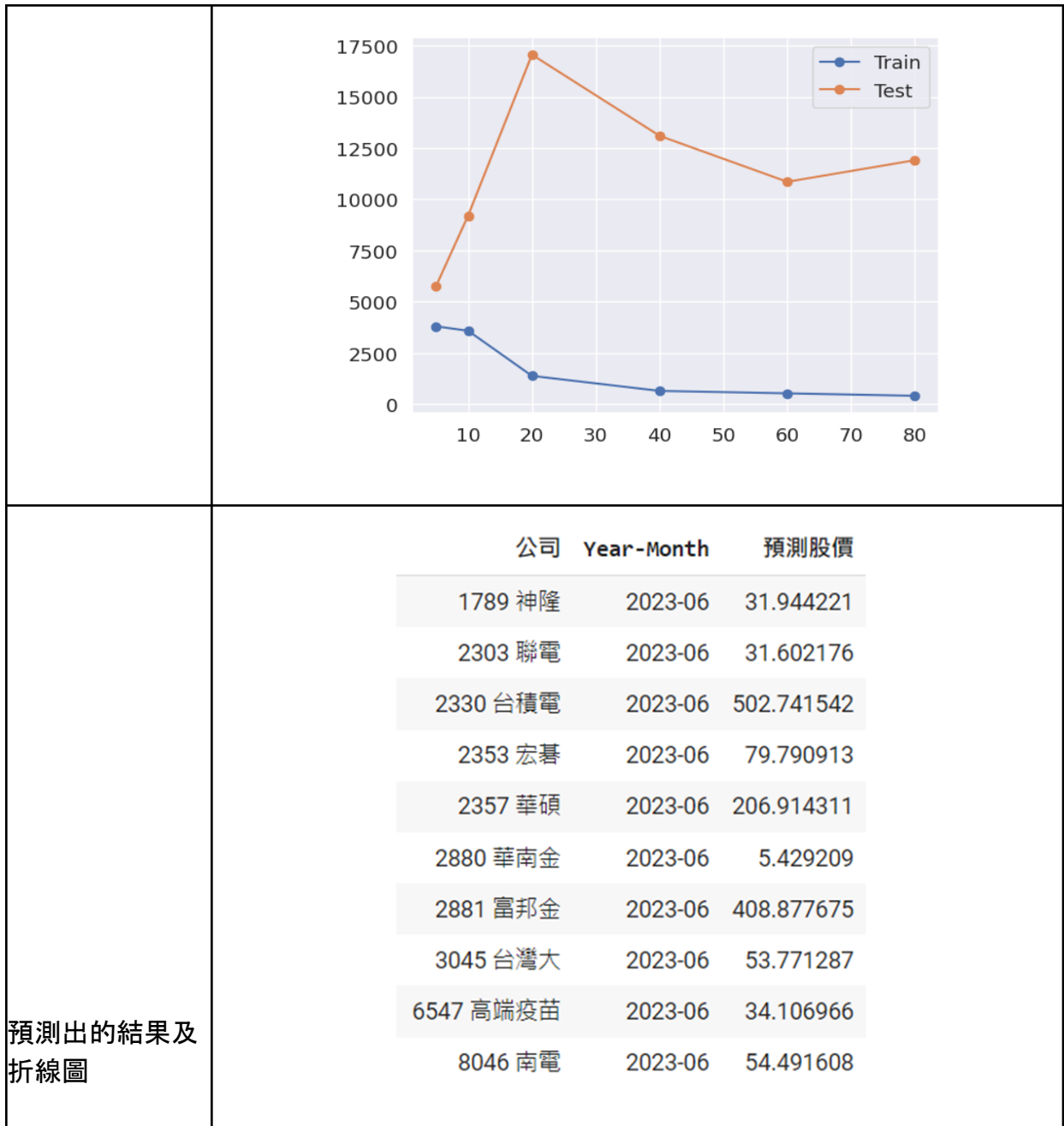
陸、研究產出

以下是我們研究預期產出的列點式說明：

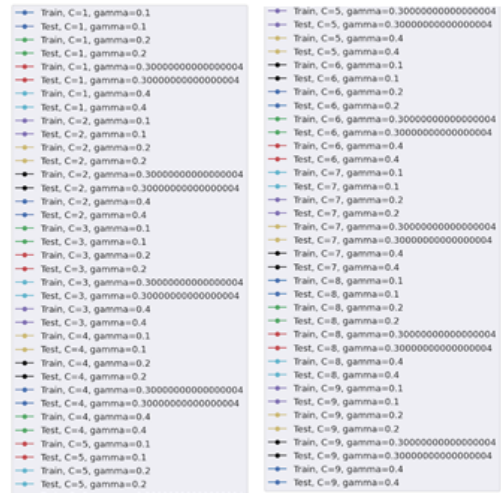
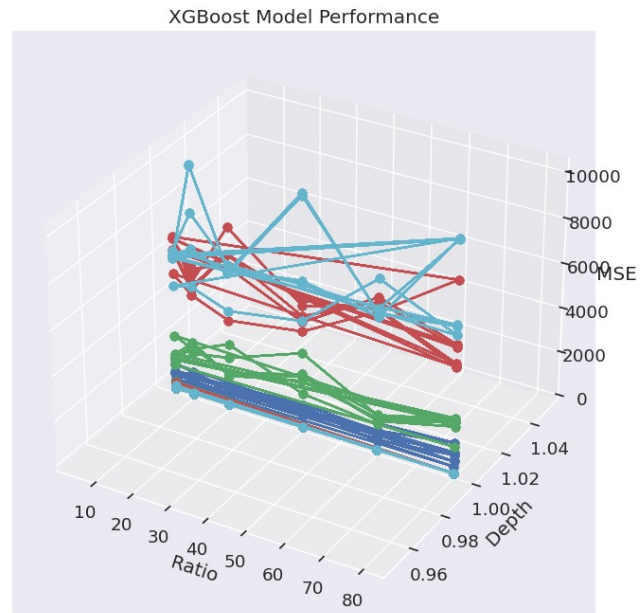
- 各模型好壞衡量：根據均方誤差 (MSE)、最佳相對誤差 (Minimum Relative Error) 衡量各模型好壞。
- 各模型的投資組合：根據預測結果，產出各模型在各期間所建議的投資組合，包括投資公司、比例等，以提供投資者作為參考。
- 期間內平均報酬：計算期間內每個模型的平均報酬，並進行各期報酬的比較，以瞭解每個模型的表現和趨勢。
- 各模型投資成效比較：根據計算的報酬率或其他指標，進行各模型的投資成效比較，以找出最適合的投資策略。

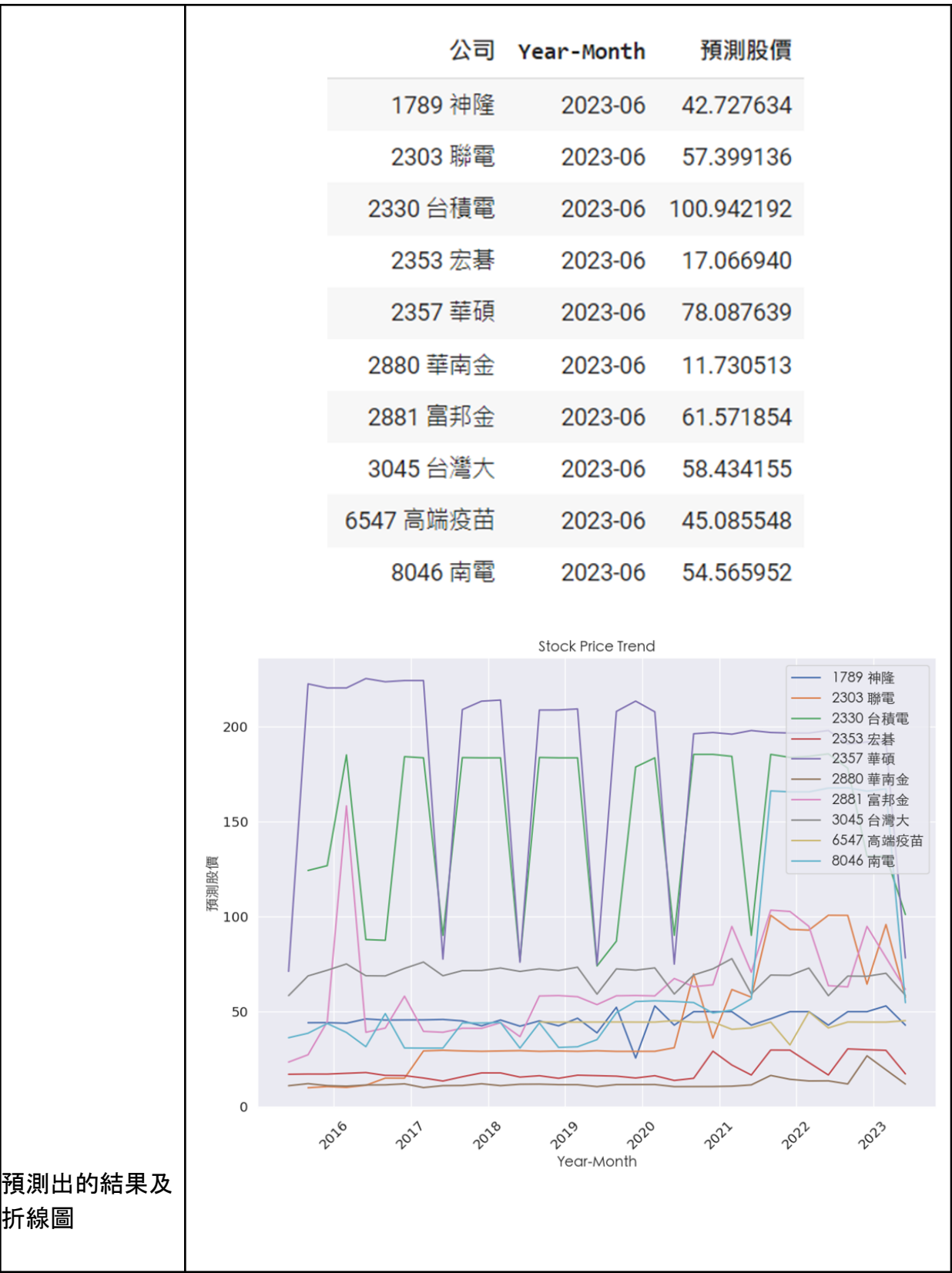
基本面模型衡量：

多元邏輯回歸	
MSE	<ul style="list-style-type: none">● 訓練集: 530.35● 測試集: 10853.61
training data占比	由圖可得當 testing ratio 為 60% 時為最佳模型，因此後續將以此 ratio 進行模型訓練。

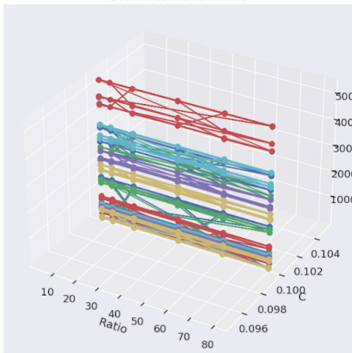


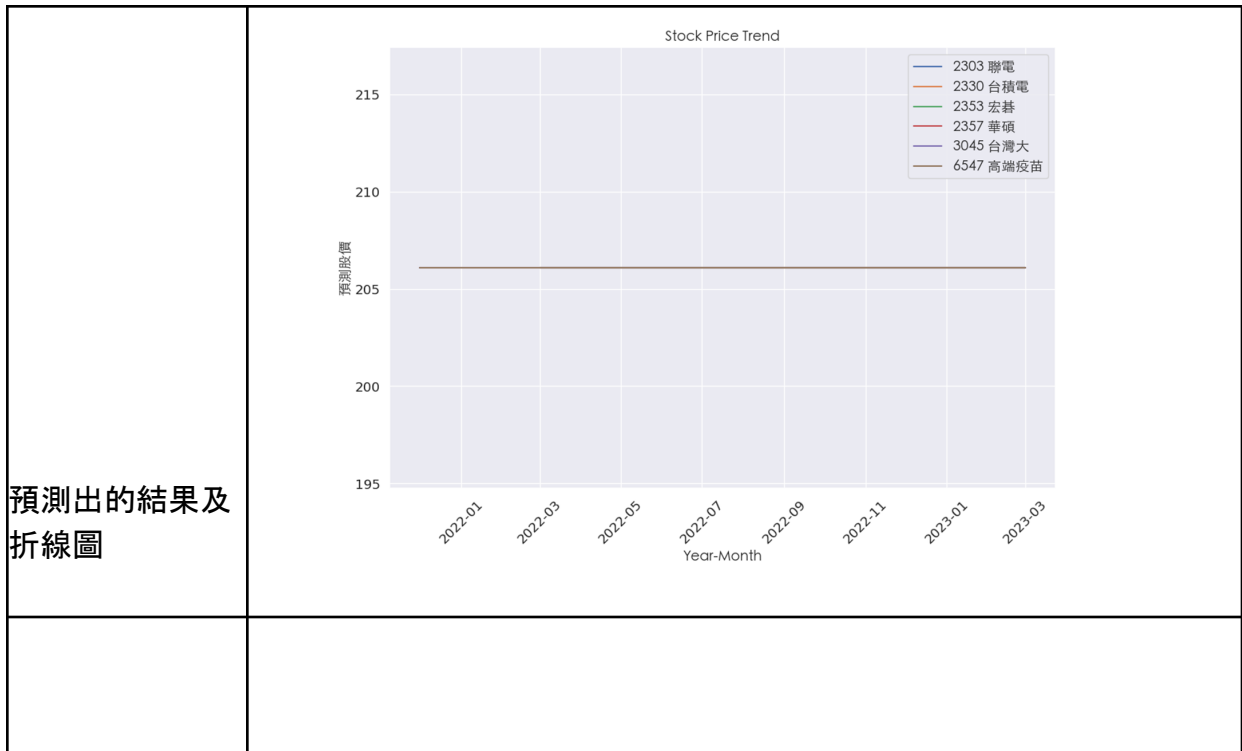
	<p>Stock Price Trend</p>
XGBoost	
MSE	<ul style="list-style-type: none"> ● 訓練集: 1.89 ● 測試集: 278.69
training data占比	<p>根據結果可得最佳 testing ratio 為 80%, 且得最佳訓練深度 depth 為 9、最佳學習率 eta 為 0.4。</p>





預測出的結果及折線圖

SVR	
MSE	<ul style="list-style-type: none">● 訓練集: -● 測試集: -
training data占比	根據圖片結果可得模型最佳訓練比例為 80%、最佳 C value 為 1000、最佳 gamma 為 1e-08
training data占比	<div><p>SVR Model Performance</p><p>Legend:</p><ul style="list-style-type: none">Train, C=0.1, gamma=1e-08Test, C=0.1, gamma=1e-08Train, C=0.1, gamma=1e-07Test, C=0.1, gamma=1e-07Train, C=0.1, gamma=1e-06Test, C=0.1, gamma=1e-06Train, C=0.1, gamma=1e-05Test, C=0.1, gamma=1e-05Train, C=0.1, gamma=0.0001Test, C=0.1, gamma=0.0001Train, C=0.1, gamma=0.001Test, C=0.1, gamma=0.001Train, C=0.1, gamma=0.01Test, C=0.1, gamma=0.01Train, C=0.1, gamma=0.1Test, C=0.1, gamma=0.1Train, C=1, gamma=1e-08Test, C=1, gamma=1e-08Train, C=1, gamma=1e-07Test, C=1, gamma=1e-07Train, C=1, gamma=1e-06Test, C=1, gamma=1e-06Train, C=1, gamma=1e-05Test, C=1, gamma=1e-05Train, C=1, gamma=0.0001Test, C=1, gamma=0.0001Train, C=1, gamma=0.001Test, C=1, gamma=0.001Train, C=1, gamma=0.01Test, C=1, gamma=0.01Train, C=1, gamma=0.1Test, C=1, gamma=0.1Train, C=10, gamma=1e-08Test, C=10, gamma=1e-08Train, C=10, gamma=1e-07Test, C=10, gamma=1e-07Train, C=10, gamma=1e-06Test, C=10, gamma=1e-06Train, C=10, gamma=1e-05Test, C=10, gamma=1e-05Train, C=10, gamma=0.0001Test, C=10, gamma=0.0001Train, C=10, gamma=0.001Test, C=10, gamma=0.001Train, C=10, gamma=0.01Test, C=10, gamma=0.01Train, C=10, gamma=0.1Test, C=10, gamma=0.1Train, C=100, gamma=1e-08Test, C=100, gamma=1e-08Train, C=100, gamma=1e-07Test, C=100, gamma=1e-07Train, C=100, gamma=1e-06Test, C=100, gamma=1e-06Train, C=100, gamma=1e-05Test, C=100, gamma=1e-05Train, C=100, gamma=0.0001Test, C=100, gamma=0.0001Train, C=100, gamma=0.001Test, C=100, gamma=0.001Train, C=100, gamma=0.01Test, C=100, gamma=0.01Train, C=100, gamma=0.1Test, C=100, gamma=0.1Train, C=1000, gamma=1e-08Test, C=1000, gamma=1e-08Train, C=1000, gamma=1e-07Test, C=1000, gamma=1e-07Train, C=1000, gamma=1e-06Test, C=1000, gamma=1e-06Train, C=1000, gamma=1e-05Test, C=1000, gamma=1e-05Train, C=1000, gamma=0.0001Test, C=1000, gamma=0.0001Train, C=1000, gamma=0.001Test, C=1000, gamma=0.001Train, C=1000, gamma=0.01Test, C=1000, gamma=0.01Train, C=1000, gamma=0.1Test, C=1000, gamma=0.1</div>



基本面買賣點紀錄:

統一以 2023 年 3 月 1 日作為進場時機, 計算從進場到 2023/ 6/2 的股票獲利。

基本面預測股價紀錄及投資標的:

以下紀錄 Regression、XGBoost 模型股價統整及投資標的決策:

1. Regression 模型

股價統整_Regression 模型					
	台積電 (2330)	聯電 (2303)	宏碁 (2353)	南電(8046)	華南金(2880)

3 月平均股價	519.04	51.47	26.03	256.15	22.54
6 月預測股價	502.74	31.60	79.79	54.49	5.43
3 月~6 月 股價漲幅	-3.14	-38.60	206.48	-78.73	-75.91
	富邦金(2881)	台灣大(3045)	華碩 (2357)	高端疫苗 (6547)	神隆 (1789)
3 月平均股價	58.45	97.74	276.54	59.77	26.08
6 月預測股價	408.88	53.77	206.91	34.11	31.94
股價漲幅	599.58	-44.99	-25.18	-42.94	22.47

由上表可得各公司股價漲幅，取其中漲幅為正的公司，分別為「宏碁 (2353)」、「富邦金(2881)」及「神隆 (1789)」，並按以下步驟進行投資組合建立：

1. 根據漲幅大小計算投資組合權重：將三間公司股價漲幅加總，計算各公司漲幅佔總漲幅比例，決定投資權重
2. 計算預計投入金額：預計投入金額各公司投資權重 * 投資起始資金 10 萬 得 預計投入金額
3. 計算買進股數：預計投入金額 / 2023 年 3 月 1 日 開盤價，並取整數位，得各公司買進股數
4. 計算投入資金：各公司買進股數 * 2023 年 3 月 1 日 開盤價 得投入資金 (若總計 > 10 萬 則投資佔比最低之個股股數 - 1)

計算結果如下表所示：

投資佔比_Regression 模型 * 起始資金為 10 萬新台幣							
公司名	3~6 月股價漲幅	投資權重	2023/03/1開盤價	2023/06/02收盤價	買進股數	投入資金	最終金額
富邦金(2881)	599.58	0.72	59.2	60.6	1216	71987.2	73689.6
宏碁 (2353)	206.48	0.25	25.4	33.95	984	24993.6	33406.8
神隆 (1789)	22.47	0.03	26.3	26.25	114	2998.2	2992.5
總和						99979	110088.9

股價資料來源: <https://tw.stock.yahoo.com/quote/2881.TW/technical-analysis>

Regression 投資成效:

公司買進股數 * 2023 年 6 月 2 日 收盤價可得最終金額, 並用來衡量截至 2023 年 6 月 2 日的投資成效, 以之起始投入資金為 \$ 99,979, 至 2023 年 6 月 2 日投資組合價值為 110,088.9, 因此可得投報率及年化報酬率如下:

- 投報率 $(110088.9 - 99979) / 99979 = 0.10\%$
- 換算為年化報酬率 = 47.01%

最初投入資金(本金)	投資期間	單位
<input type="text" value="99979"/>	<input type="text" value="3"/>	<input type="text" value="月"/>
最終金額	年化報酬率(%)	
<input type="text" value="110088.9"/>	<input type="text" value="47.01"/>	

2. XGBoost 模型

使用模型預測出各公司 2023 年 6 月股價後，計算 3 月至 6 月股價漲跌幅，將漲幅 > 0 之個股納入投資組合，如下表黃色標註。

股價統整_XGBoost 模型					
	台積電 (2330)	聯電 (2303)	宏碁 (2353)	南電(8046)	華南金(2880)
3 月平均股價	519.04	51.47	26.03	256.15	22.54
6 月預測股價	100.94	57.40	17.07	24.57	11.73
3 月 ~6 月 股價漲幅	-80.55	11.53	-34.45	-78.70	-47.95
	富邦金(2881)	台灣大(3045)	華碩 (2357)	高端疫苗 (6547)	神隆 (1789)

3 月平均股價	58.45	97.74	276.54	59.77	26.08
6 月預測股價	61.57	58.43	78.09	45.09	42.73
股價漲幅	5.35	-40.22	-71.76	-24.57	63.82

由上表可得各公司股價漲幅，取其中漲幅為正的公司，分別為「神隆 (1789)」、「聯電 (2303)」及「富邦金(2881)」，並按以下步驟進行投資組合建立：

1. 根據漲幅大小計算投資組合權重：將三間公司股價漲幅加總，計算各公司漲幅佔總漲幅比例，決定投資權重
2. 計算預計投入金額：預計投入金額各公司投資權重 * 投資起始資金 10 萬 得 預計投入金額
3. 計算買進股數：預計投入金額 / 2023 年 3 月 1 日 開盤價，並取整數位，得各公司買進股數
4. 計算投入資金：各公司買進股數 * 2023 年 3 月 1 日開盤價 得投入資金 (若總計 > 10萬 則投資佔比最低之個股股數 - 1)

計算結果如下表所示：

投資佔比_XGBoost 模型 * 起始資金為 10 萬新台幣							
公司名	3~6 月股價漲幅	投資權重	2023/03/1 開盤價	2023/06/02 收盤價	買進股數	投入資金	最終金額
神隆 (1789)	63.82	0.79	26.3	26.25	3004	79005	78855

聯電 (2303)	11.53	0.14	49	51.4	286	14014	14700.4
富邦金 (2881)	5.35	0.07	59.2	60.6	117	6926	7090.2
總和						99945	100645.6

股價資料來源: <https://tw.stock.yahoo.com/quote/2881.TW/technical-analysis>

XGboost 投資成效:

公司買進股數 * 2023 年 6 月 2 日 收盤價可得最終金額, 並用來衡量截至 2023 年 6 月 2 日的投資成效, 以之起始投入資金為 \$ 99,945, 至 2023 年 6 月 2 日投資組合價值為 100,645.6, 因此可得投報率及年化報酬率如下:

- 投報率 = $(100645.6 - 99945) / 99945 = 0.7\%$
- 年化報酬率 = 2.83%

最初投入資金(本金)	投資期間	單位
<input type="text" value="99945"/>	<input type="text" value="3"/>	<input type="text" value="月"/>
最終金額	年化報酬率(%)	
<input type="text" value="100645.6"/>	<input type="text" value="2.83"/>	

前 10 重要變數:

- 回歸模型:
 - 每股稅前淨利: 20.549377524001002
 - 稅前純益 / 實收資本: 1.678745880087783
 - 有息負債利率: 0.7809608171300852

- 每股現金流量: 0.6359487109664078
- ROA(C)稅前息前折舊前: 0.4214210096591013
- 利息支出率: 0.3913548610804077
- 淨值/資產: 0.10883783501570973
- 當期所得稅資產—流動: 2.0074889326835824e-05
- 遞延所得稅資產: 6.021222207682248e-06
- 普通股股本: 2.0516128150204164e-06
- XGBoost 模型
 - 稅前純益 /實收資本: 0.8838732242584229
 - 當期所得稅資產—流動: 0.07893867045640945
 - 普通股股本: 0.015522556379437447
 - 淨值/資產: 0.006122684106230736
 - 採權益法之長期股權投資: 0.004142888356000185
 - 遞延所得稅資產: 0.0027675617020577192
 - 每股稅前淨利: 0.0023191585205495358
 - ROA(C)稅前息前折舊前: 0.001818023039959371
 - 利息支出率: 0.0014838390052318573
 - 應收帳款及票據: 0.0014059750828891993
- SVR 模型:

因為在訓練模型時並沒有跑出正常的結果，因此在結果上並不會討論到此模型所較重視的重要特徵。

根據以上結果，下列重要變數可以做為基本面投資者進行投資決策時的重要依據，說明敘述如下：

- 每股稅前淨利 (EPS_Earnings Per Share): EPS 反映了公司在每股股份上的盈利能力，高盈利能力可能意味著更高的股價。當一家公司的 EPS 增長時，通常會對股價有正面影響。
- 稅前純益 /實收資本 (PBT_Profit Before Tax / Paid-in Capital): 這個比率反映了公司每單位實收資本所產生的稅前利潤。這可以衡量公司有效運用資本的能力。高比率可能表明公司的利潤能力較強，可能吸引更多投資者，進而對股價有正向影響。
- ROA_Return on Asset: 是一個衡量公司資產運營效率的指標。它表示公司每單位資產所產生的利潤。高ROA表示公司在運營資產方面效率較高，可能對股價產生正向影響。

- 利息支出率(Interest Expense Ratio): 這個比率反映了公司利息支出占總收入的比例。高利息支出率可能表明公司面臨較高的財務負擔，可能對股價產生負面影響。
- 淨值/資產 (Equity/Assets Ratio): 這個比率衡量公司淨值與總資產之間的關係，反映了公司的財務結構。高淨值/資產比可能表示公司有較低的財務風險，可能對股價產生正面影響。

技術面模型衡量:

SVR					
	台積電 (2330)	聯電 (2303)	宏碁 (2353)	南電(8046)	華南金(2880)
最佳相對誤差	0.0%	0.0%	0.0%	0.0%	0.0%
training data 占比	20%	20%	70%	20%	90%
	富邦金(2881)	台灣大(3045)	華碩 (2357)	高端疫苗 (6547)	神隆 (1789)
最佳相對誤差	0.0%	0.0%	0.0%	0.0%	0.0%
training data 占比	60%	30%	20%	20%	90%
RandomForest					

	台積電 (2330)	聯電 (2303)	宏碁 (2353)	南電(8046)	華南金(2880)
最佳相對誤差	0.39%	0.16%	0.41%	0.60%	0.00%
training data 占比	95%	95%	95%	90%	85%
	富邦金(2881)	台灣大(3045)	華碩 (2357)	高端疫苗 (6547)	神隆 (1789)
最佳相對誤差	0.01%	0.01%	0.06%	2.36%	0.00%
training data 占比	95%	95%	95%	80%	90%
XGboost					
	台積電 (2330)	聯電 (2303)	宏碁 (2353)	南電(8046)	華南金(2880)
最佳相對誤差	0.0%	0.0%	0.0%	0.0%	0.0%
training data 占比	20%	20%	20%	20%	20%
	富邦金(2881)	台灣大(3045)	華碩 (2357)	高端疫苗 (6547)	神隆 (1789)

最佳相對誤差	0.0%	0.0%	0.0%	0.0%	0.0%
training data 占比	20%	20%	20%	20%	20%

SVR、XGboost之最佳相對誤差，是透過計算預測值(y_{pred})和實際值(y_{test})之間的差異之絕對值，並統計大於閾值 (epsilon)之比例，作為最佳相對誤差。

RandomForest之最佳相對誤差，是透過計算coefficient of variation, CV, 來計算出誤差並且轉換成百分比。

技術面買賣點紀錄:

首先，透過列出所有公司之股價最高與最低點日期，當最高點日期發生在最低點日期之後，即給該公司代表上漲的紅色標記，反之則給予代表下跌之綠色。如下圖：

SVR					
	台積電 (2330)	聯電 (2303)	宏碁 (2353)	南電(8046)	華南金(2880)
Low	5月22日	4月11日	3月16日	3月17日	4月26日
High	5月26日	5月4日	4月11日	5月26日	4月12日
	富邦金(2881)	台灣大(3045)	華碩 (2357)	高端疫苗 (6547)	神隆 (1789)
Low	3月29日	4月17日	4月7日	5月24日	5月19日

High	4月11日	5月4日	5月26日	3月30日	3月27日
RandomForest					
	台積電 (2330)	聯電 (2303)	宏碁 (2353)	南電(8046)	華南金(2880)
Low	4月14日	4月14日	3月31日	4月20日	3月1日
High	5月19日	5月17日	5月30日	3月16日	4月25日
	富邦金(2881)	台灣大(3045)	華碩 (2357)	高端疫苗 (6547)	神隆 (1789)
Low	3月1日	3月2日	3月1日	5月25日	3月10日
High	5月4日	4月25日	3月16日	3月16日	4月11日
XGboost					
	台積電 (2330)	聯電 (2303)	宏碁 (2353)	南電(8046)	華南金(2880)
Low	4月14日	4月17日	3月31日	4月20日	3月1日
High	5月4日	5月4日	3月1日	3月16日	3月7日

	富邦金(2881)	台灣大(3045)	華碩 (2357)	高端疫苗 (6547)	神隆 (1789)
Low	3月1日	3月1日	3月1日	5月24日	3月10日
High	5月4日	4月24日	3月16日	3月1日	3月23日

將三個模型皆是預測具有上漲趨勢的留下，則得到下表。

SVR					
	台積電 (2330)	聯電 (2303)	富邦金(2881)	台灣大(3045)	華碩 (2357)
Low	5月22日	4月11日	3月29日	4月17日	4月7日
High	5月26日	5月4日	4月11日	5月4日	5月26日
RandomForest					
	台積電 (2330)	聯電 (2303)	富邦金(2881)	台灣大(3045)	華碩 (2357)
Low	4月14日	4月14日	3月1日	3月2日	3月1日
High	5月19日	5月17日	5月4日	4月25日	3月16日

XGboost					
	台積電 (2330)	聯電 (2303)	富邦金(2881)	台灣大(3045)	華碩 (2357)
Low	4月14日	4月17日	3月1日	3月1日	3月1日
High	5月4日	5月4日	5月4日	4月24日	3月16日

接下來我們計算了這五家公司在最高價與最低價的漲幅百分比。如下圖：

SVR					
	台積電 (2330)	聯電 (2303)	富邦金(2881)	台灣大(3045)	華碩 (2357)
Low	456.00	42.43	55.17	97.30	269.51
High	499.84	43.12	56.28	99.03	270.92
漲幅 (%)	9.61	1.63	2.01	1.78	0.52
RandomForest					
	台積電 (2330)	聯電 (2303)	富邦金(2881)	台灣大(3045)	華碩 (2357)

Low	443.17	40.82	55.39	92.51	255.60
High	539.41	50.43	60.57	97.83	282.91
漲幅 (%)	21.72	23.54	9.35	5.75	10.68
XGboost					
	台積電 (2330)	聯電 (2303)	富邦金(2881)	台灣大(3045)	華碩 (2357)
Low	282.05	25.63	35.15	61.24	161.94
High	322.82	29.71	38.04	62.42	178.30
漲幅 (%)	14.45	15.92	8.22	1.93	10.10

漲跌幅百分比之前三名，我們以黃色標註，可以看到在三個模型中，台積電皆有入選漲幅的前三名。因此，選擇台積電進行投資。

透過預測結果可知，台積電的進場點與獲利了結點有以下幾個可能：

進場點：4月14日(RandomForest、XGboost模型預測結果相同)、5月22日

獲利了結點：5月4日、5月19日、5月26日

進場點部分，我們會選擇4月14日，除了他有兩個模型支持之外，若於5月22日進場，沒有機會碰到其他兩個模型預測的高峰。

獲利了結點，我們選擇三個日期的中間時段5月19日。

技術面投資成效：

接下來計算實際收益：

4月14日：買入台積電零股190股，投入新台幣98040元。

5月19日：賣出台積電零股190股，收入新台幣101,080元

經計算， $((101,080-98040)/98040)*100\%=3.1\%$

最初投入資金(本金)	投資期間	單位
<input type="text" value="98040"/>	<input type="text" value="3"/>	<input type="text" value="月"/>
最終金額	年化報酬率(%)	
<input type="text" value="101080"/>	<input type="text" value="12.99"/>	

消息面投資成效：

一共有4票選擇5/9(二)進行投資，為配合只有一次買賣點機會的限制，選擇5/9(二)購買宏碁(28.7)及聯電(49.95)。以9:1分配。於5/24賣出(宏碁31、聯電49.3)。

賣出的時機點設立於宏碁漲至8%的時刻。

之所以以9:1的方式形成組合，原因是提議購買宏碁的消息源有3票，而聯電僅1票，且為facebook新聞所推薦，消息來源並非十分可信，因而僅以1成的資金進行投資。

公司	買股	賣股	淨利(損)
宏	3張(86100)	3張(93000元)	6900元

基	元)		
聯 電	250股 (12489元)	250股(12325元)	-164元

結果:

一共營利**6736元**

年化報酬率:**29.79%**

柒、結果討論

一. 基本面:

1. SVR模型問題:

在本次的研究結果中, 我們的 SVR 模型不盡理想, 所預測出來的數值都完全相同, 在經過幾次的調參後(如上述所示), 還是得到一樣的結果, 根據他人的經驗, 這可能跟特徵的選擇以及訓練資料的多寡有關。首先是特徵的選擇, 因為 SVR 對於特徵較為敏感, 因此在特徵的縮放上會比較顯著, 如果要改善, 可能必須得做特徵歸一化, 但因為不確定這樣歸一化是否會造成數據上的錯誤, 反而做出一個不正確的模型; 第二個有可能的原因是訓練資料的多寡, 因為本次研究是依據財報裡的參數來當作特徵, 因此在訓練資料上可能會遠遠不足, 在這樣的情況下, 有可能也會讓 SVR 做出無法討論的結果。加上還有許多我們為思考到的因素, 因此在本次報告中我們選擇不呈現此模型結果。

2. 模型預測結果:

在基本面的研究結果中, 可以看到雖然我們的年化報酬率十分高, 甚至超越其他兩組, 但是以所預測出的數值來說, 可以說和現實有非常大的差異, 這牽涉到許多因素, 有可能是跟我們所訓練的模型有關, 有可能是跟我們dataset的選擇方式有關, 也還有可能是跟其他市場波動等現實因素我們沒有考慮進去, 需要更多的因素考量在內。然而, 本次研究成果中各模型分別使用的重要特徵大部分都是相同的, 從這點上來看可以認為, 這些被選出來的特徵是可能當作進行投資決策時的重要參考依據。

3. 基本面的投資本質:

依照傳統的基本面投資, 都是以財報當作參考對象, 加上許多新聞以及個人觀點後所作出的投資行為, 而這樣的投資基本上都是以長期為主, 但因為時間的關係, 本次研究與傳統基本面投資的理念有著不小的差異, 但之所以堅

持做是想要跟技術面做出比較，加上在網路上鮮少有人是以機器學習來做基本面分析，因此才有這次的報告。

二. 技術面：

1. 模型相對誤差過低：

三種模型的相對誤差都過低，推測可能是因為最佳相對誤差極低的可能原因為，我們預測之目標—收盤價，也被作為特徵值輸入模型中，造成最佳相對誤差極低。未來考慮將收盤價獨立出來，進行分析以避免模型受到干擾過大，也能單獨對收盤價進行更多的分析。

2. 投資策略之差異：

因本次投資的策略在最開始，是以最高最低價之出現先後順序進行第一波篩選，這可能會忽略某些震盪幅度較大、具有價差獲利潛力的股票，但這同時也伴隨著持有者必須承受股價震盪之壓力。因此不同類型的投資人，可能會對相同的預測結果有著不同的投資策略。以在第一關就被我們淘汰的南電來說，4/20-255買入零股392股，投入新台幣99960元，5/9-287.5零股賣出，收穫112700元，報酬率：12.75%，年化報酬率61.58%，這就是被我們忽略掉的價差獲利潛力。

3. 模型準度的提升：

我們的模型是以60天前的股價預測60天後的股價，透過資料的橫移，達到一次預測三個月後股價的結果。這樣不考慮靠近預測目標之日期的股價(比如昨天的股價對今天的股價之影響)，可能造成預測結果有沒被考慮到的特徵值，這是我們可以優化的部分。此外亦可添加各國股市、市場情緒等因子，作為預測的特徵值輸入，達到更準確的預測。

三. 消息面：

1. 消息源可信度：

每個消息源的可信度是消息面最大的問題，就如同最後投資的結果，哪怕是網路上的新聞也不見得一定是真的。此外，由於5種消息源中，只有一位為職業操盤手，其餘4種消息源的可信度及長期投資成功率，都相對的比程式、職業操盤手更低。

2. 投資策略的不同：

事實上，雖淘汰了4間相較不受待見的公司，但事實上，這幾間公司在過去三個月期間，也有能夠大量獲利的區段，但因為投資策略的差異，導致這四間先被淘汰，屬實也是有些可惜。

3.最終投資組合的調整：

以結果論來看，這樣的投資策略是成功的，雖然有一檔股票並無獲利，但我認為這是受限於一次買入及放空的限制，若無這限制，聯電確實也是一個不錯的投資選項。已分配比例較重的股票，決定做空的時機，我認為這是這次消息面能夠成功的主要原因之一。

捌、研究限制

由於專案時程短，我們必須限縮研究範圍，選定的產業和公司數量會比較有限。這可能會對模型的準確度和預測能力產生一定的限制，因為我們無法觀察到更廣泛的市場變化和趨勢。同時，限制了公司數量也可能導致樣本不夠大，使得模型的可信度和泛化能力受到影響。因此，我們需要謹慎選擇產業和公司，並在研究方法和模型設計上做更多優化，以確保研究結果的有效性和可靠性。

玖、研究心得

基本面：

藉由這次的研究我們發現了用機器學習進行股價預測的潛力，可以提供交易策略及投資建議，且帶入不同季度資料後便可使模型應用於任何年月，可省去散戶許多上網查找資訊，並進行決策的投資時間成本。

然而我們也在研究的過程中了解到股價預測具有一定程度的挑戰性，因為股價受到多種因素影響，除了公司財報數據等基本面資訊外，也受到市場風險或其他總體經濟狀況、政策變化、國際局勢等未考慮因素影響，因此，即便用了較佳的特徵和模型，仍會出現超出模型預測能力的情形。

技術面：

本次專題讓我遇到很多原本沒想過的問題，比如資料庫可能有誤、怎麼平移資料預測未來股價、需要排除許多可能影響股價的外來因素……，諸多問題不斷地在過程中冒出等著我們解決，這是本次我認為最大的收穫。

關於七研究本身，我認為技術面雖然已經有一點點抓出未來趨勢、甚至接近實際狀況，但距離用來實際投資還是差了太多，可能要多利用美股、多種資料預測方式(用去年同一時間的股價預測、用一季前的資料預測)，來增加預測的精確度。

消息面：

從五種不同的消息源，也意味著五種不同面向去投資股市，可以發現縱然每種投資策略不同，但憑據提供的線圖，可以做出正確的篩選及投資。更能夠藉由此種方式，降低風險。

此外，在這次投資中，聯電的股票是以虧損為最終結果。這也能顯現出，縱然獲得各種消息，可能能夠協助股票的買賣，但消息及新聞的可信度都需再查證再進行投資。

拾、小組分工

以下是本研究的分工安排，我們將分為三個小組分別進行基本面、技術面、消息面相關研究及模型建立，每個小組成員如下：

1. 基本面研究小組：

林姝延
賴宇辰

2. 技術面研究小組：

曾國睿
王奕翔

3. 消息面研究小組：

許智評

參考資料：

- [財訊](#)

- [TIMCO](#)
- [台灣info](#)
- [Yahoo! 股市](#)
- 三竹股市 APP
- TWStock

投資一定有風險，基金投資有賺有賠，申購前應詳閱公開說明書