以決策樹建構疫情下因子選股策略-以台灣電子工業股為例

系級：金融所碩一

學號：109352007

姓名：黃詩媛

一、 研究動機

2020 年初疫情爆發，造成同年 3 月時市場恐慌股市大跌，而後台灣股市卻表現亮眼，大盤不斷創新高，其中尤其是半導體、5G 類股表現最佳，帶領台股走向新高，且在全球低利率的環境下，資金不斷流入報酬率較高的股票市場，股票已然成為多數投資人的主要投資標的。

本研究主要探討電子工業類股在疫情下的因子選股策略，希望能透過過去 2015-2019 年的資料，得出有效的選股策略，建構出 2020 年疫情期間下有效的選股策略，驗證因子選股的有效性，故本研究以 2015-2019 年台灣經濟新報的電子工業類股(含半導體、光電業、通訊網路業等等)的股東權益報酬率、本益比與研究發展費用率之基本面因子為決策樹模型訓練及測試樣本，並以類股股票報酬大於台灣加權股價報酬為目標，以期能夠得出在 2020 疫情期間勝過大盤的選股策略。

二、 研究架構

本研究為決策樹因子分析，大致上會如下圖之研究架構對於以決策樹方法對電子類股進行分析。

1. 取得 TEJ 電子類股資料 2015-2020 年月收盤價、月報酬率、本益比、股東權益報酬率與研究發展費用率的資料。
2. 篩選流動性較好的股票，本研究從 406 家公司篩選出 150 家流動性最佳的公司。
3. 進行資料清理，去除空值與無效樣本(例如本益比或股東權益報酬率為零或負值)。
4. 對報酬率進行標籤，當類股報酬率勝過大盤表現為 1，遜於大盤表現為 0。

5. 將樣本資料分割成訓練與測試資料，在使用決策樹方法找出有效的因子選股策略。
6. 驗證 2020 年疫情期間股災的狀況下，決策樹選股策略是否有效，進行回測分析。

三、 研究方法

- **基本面因子**

本研究選取**本益比、股東權益報酬率與研究發展費用率**作為分析因子。以下介紹各因子選取原因。

1. **本益比**(Price-to-Earning Ratio)

本益比為**現在股價除以預估未來每年每股盈餘**，假設本益比為 20 倍，代表需 20 年才能回本，也就是說本益比能幫助我們判斷目前股價是昂貴或是便宜，也代表投資人對這家公司的展望，本益比愈高代表投資人愈看好這間公司未來能賺愈多錢，當成長性越高，本益比通常也越高，符合電子類股成長性高的特性。若 EPS 為負值，則應用本益比計算還本時間無意義，故本研究會將負或零的本益比去除。

2. **股東權益報酬率**(Return On Equity)

股東權益報酬率為**稅後淨利除以股東權益**，稅後淨利就是公司本期的獲利，股東權益為總資產扣除負債，代表還完負債後公司的淨值，最主要是股本加上保留盈餘與資本公積，因此股東權益報酬率也就是公司用自有資本賺錢的能力，高 ROE 代表公司可以用同樣的股東權益賺到更多的錢，但使用這項衡量公司獲利的因子也需考慮公司負債，避免是因公司槓桿程度高，股東權益報酬率才高的情況發生。

股東權益報酬率若為負值也無意義，本研究會去除股東權益報酬率為負值或零的資料。

3. **研究發展費用率(Research & Development)**

本研究所選擇之電子高科技產業為台灣重點發展產業，也是此次疫情下表現最佳的類股，尤其是半導體與通訊業，而這些產業日新月異，非常倚靠研發能量，例如台積電製成超前，相對投入的研發成本也高，故本研究選擇研究發展費用率作為選股因子之一。

- **決策樹模型**

本研究應用決策樹模型對選取的基本面因子進行分析，決策樹以分類模型為主，亦可用於回歸，決策樹會依據給定的特徵設定條件判斷是否滿足設定目標來進行分類，因為決策樹每一個葉節點都是一個條件判斷式，不需要太多的計算，不論輸入的資料為類別尺度、連續尺度都可以進行分類。

當決策樹一般是以二元樹為主，在二元樹中，每個節點代表提問者所提的一個問題，最高層的節點稱為根節點(Root node)，根節點再依問題的滿足與否分裂成左右兩個子節點，如左子節點可代表 True，而右子節點代表 False，依此進行逐層分裂。若分裂後節點的實例僅含單一類別，則此節點就不再行分裂，這種節點稱為葉節點 (Leaf node)，決策樹會計算後選取具最大資訊增益的特徵(因子)進行分裂，並於分裂後的各子節點依此程序進行，直至全部節點都是葉節點。

而在決策樹中有兩種特徵選擇的方式，一為熵(Entropy)，二為基尼不純度(Gini Impurity)，由於我們希望獲得的資訊量要最大，因此經由分割後的資訊量要越小越好，以下為兩方法資訊量公式：

熵(Entropy)資訊量公式：

$$H(x) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

基尼不純度(Gini Impurity)資訊量公式：

$$G(x) = -\sum_{i=1}^{n} p(x_i)(1 - p(x_i))$$

本研究以基尼不純度為主要的特徵選擇方式，也就是 CART 樹方法，兩方法結果在本研究準確率幾乎相似，而相對熵(Entropy) 基尼不純度較不複雜，故運算較快速。

四、 實證結果

1. 樣本選取

本研究從台灣經濟新報(TEJ)抓取資料，抓取 2015 年 1 月 1 日至 2020 年 12 月 31 日電子類股(含半導體、電腦與週邊設備、光電、通訊網路、電子零件、電子通路、資訊服務與其他電子業)之月資料，含證券代碼、年月日、市值(百萬元)、收盤價、報酬率、本益比、研究發展費用率與股東權益報酬率，共 406 家公司。

	證券代碼	年月日	證期會代碼	市值(百萬元)	收盤價(元)_月	報酬率%_月	本益比-TSE	研究發展費用率	ROE - 綜合損益
0	1471 首利	2015-01-30	1471	1805.0	12.40	-1.5386	0.0	NaN	NaN
1	1471 首利	2015-02-26	1471	1742.0	11.97	-3.4601	0.0	NaN	NaN
2	1471 首利	2015-03-31	1471	1720.0	11.82	-1.2717	0.0	0.10	-3.31
3	1471 首利	2015-04-30	1471	1696.0	11.65	-1.4051	0.0	NaN	NaN
4	1471 首利	2015-05-29	1471	1662.0	11.42	-2.0188	0.0	NaN	NaN
...	...	...	...	...	...	...	...	...	...
28091	9912 偉聯	2020-08-31	9912	376.0	6.99	1.8951	0.0	NaN	NaN
28092	9912 偉聯	2020-09-30	9912	405.0	7.54	7.8685	0.0	2.77	-5.19
28093	9912 偉聯	2020-10-30	9912	385.0	7.17	-4.9073	0.0	NaN	NaN
28094	9912 偉聯	2020-11-30	9912	405.0	7.53	5.0208	0.0	NaN	NaN
28095	9912 偉聯	2020-12-31	9912	451.0	8.39	11.4210	0.0	NaN	NaN

2. 流動性篩選

本研究將 406 家公司在 2015-2020 年市值進行加總平均，並取出前 150 家市值最高的公司，通常來說，市值越大的股票往往擁有更高的流動性。

3. 資料清理

因研究發展費用率與股東權益報酬率是**季**資料，報酬率與本益比為**月**資料，故在此將報酬率與本益比以平均(例如 3 月的季資料是以 1、2、3 月的資料做平均)的方式轉換為季資料，接著再合併報酬率、本益比、研究發展費用率與股東權益報酬率，再統一去除空值，本益比與股東權益報酬率小於或等於零的資料無意義，故也去除。

4. 目標標籤

由於本研究目標希望得到勝過大盤表現的策略，故新增一行 target 將類股報酬率大於大盤表現的資料標籤為 1，反之則為 0。與上方資料清理得的資料進行以日期為準的合併，再進行決策樹分析。

證券代碼	年月日	研究發展費用率	ROE - 綜合損益	報酬率%	本益比	報酬率%_bench	target
1582 信錦	2015-03-31	1.32	1.35	3.989167	13.086667	0.998600	1
2059 川湖	2015-03-31	3.10	5.47	5.561200	27.923333	0.998600	1
2301 光寶科	2015-03-31	2.98	0.74	3.693200	12.936667	0.998600	1
2303 聯電	2015-03-31	7.75	1.68	1.726067	20.950000	0.998600	1
2308 台達電	2015-03-31	7.09	2.59	1.617600	23.543333	0.998600	1
...	...	...	...	...	...	...	...
8114 振樺電	2020-09-30	5.25	-2.57	0.074100	20.400000	2.599433	0
8131 福懋科	2020-09-30	1.26	3.94	-1.151833	11.176667	2.599433	0
8150 南茂	2020-09-30	4.52	8.72	-3.637700	8.386667	2.599433	0
8163 達方	2020-09-30	3.65	4.15	2.207633	13.083333	2.599433	0
8213 志超	2020-09-30	0.00	12.38	4.380367	5.960000	2.599433	1

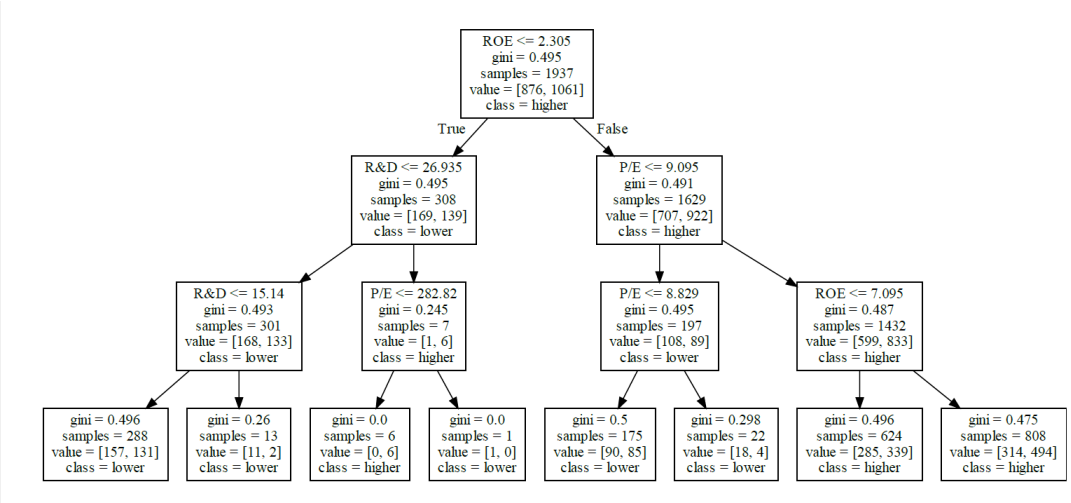
5. 決策樹分析

運用上方整理好的資料取 2015-2019 年的資料進行訓練與測試，**研究發展費用率、股東權益報酬率與本益比**為自變數因子，target 欄位為應變數，總共 2724 筆資料，並以 8：2 分割為訓練及測試資料，2179 筆資料為訓練資料，545 筆資料為測試資料，接下來運用不同因子進行決策樹分析。

- 因子：研究發展費用率、股東權益報酬率與本益比

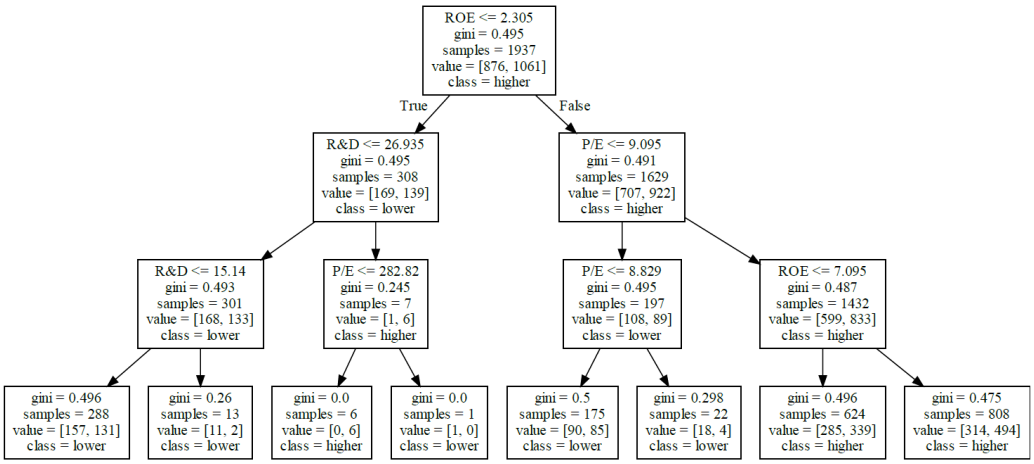
決策樹預測準確率為 0.5484536082474227，層度為 3 層。

在此得到**股東權益報酬率大於 7.095 且本益比大於 9.095** 的選股策略。



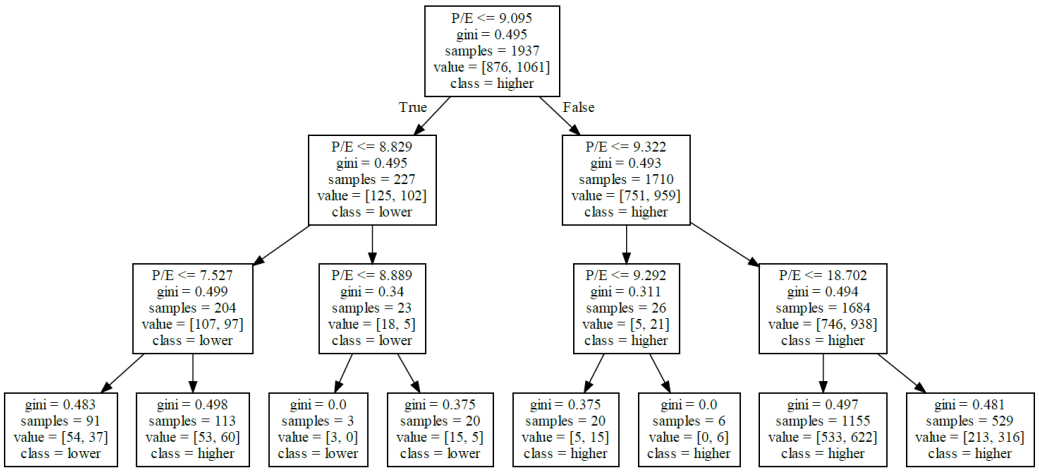
➢ 因子：股東權益報酬率

決策樹預測準確率為 0.5587628865979382，層度為 3 層。



➢ 因子：本益比

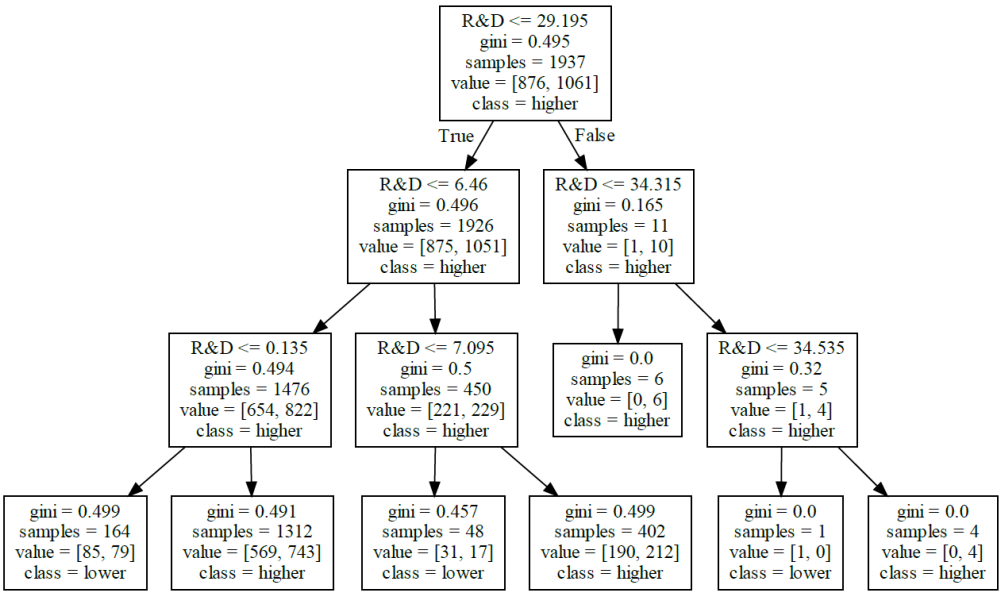
決策樹預測準確率為 0.5917525773195876，層度為 3 層。



在此得到**本益比大於 18.702** 的選股策略。

➢ 因子：研究發展費用率

決策樹預測準確率為 0.5670103092783505，層度為 3 層。



6. 回測實證

從上述的決策樹分析中，可歸納出兩個策略。

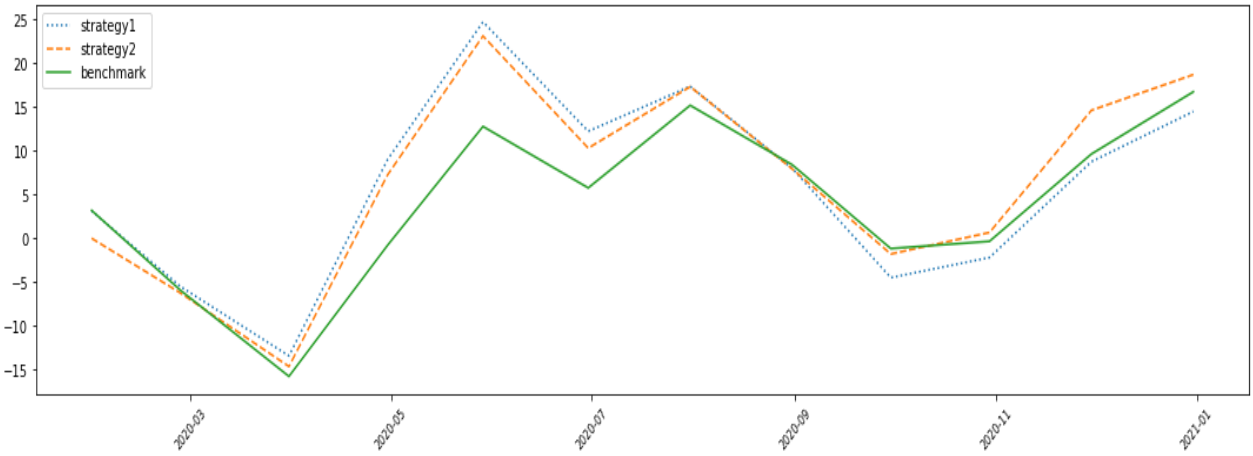
➢ 第一個是應用研究發展費用率、股東權益報酬率與本益比這三個因子的策略，是**股東權益報酬率大於 7.095 且本益比大於 9.095** 的選股策略。

➢ 第二為則是由各單一因子決策樹分析中取**準確率最佳的選股策略**，以**本益比大於 18.702** 為篩選條件的選股策略。

本研究以這兩個策略篩選出在 2020 年股東權益報酬率與本益比符合上述條件的個股，**策略一篩選出 42 家公司，策略二篩選出 59 家公司。**

以這兩個策略篩選出的類股分別做同一月份的月報酬加權平均，再和大盤(benchmark)的月報酬率在 2020 年做回測實證(累積報酬率)。

下圖 strategy1 為**股東權益報酬率大於 7.095 且本益比大於 9.095** 的選股策略，strategy2 為**本益比大於 18.702** 的選股策略，benchmark 為大盤報酬率。可看出在 **2020 年 9 月以前兩策略明顯勝過大盤**，而後貼近大盤表現，**策略一年底表現略輸大盤，策略二則是勝過大盤表現的。**



五、 結論

本研究共選取研究發展費用率、股東權益報酬率、本益比與報酬三個因子，目標為是否勝過大盤來進行決策樹分析，而無論是三個因子或是個別因子的決策樹分析，預測準確率表現都不甚理想，大概都在 55%左右，在回測實證中也只有策略二表現較佳，原因猜測如下：

◆ **基本面外的影響因素**

基本面因素對觀察公司的營運狀況相當重要，但股價不只會被財報資料影響，仍會被其他總體或個別事件影響，或許可以研究企業財報公布前後之

影響，觀察是否顯著影響公司股價，或是是否只在財報公布時間有顯著影響。

- **因子選擇**

本研究只選取三個重要因子，往後可多選取不同的基本面因子，進行迴歸分析取出對於電子類股報酬率更具顯著影響的因子做分析。

- **時間延遲因素**

通常股價會領先基本面數值，因股市瞬息萬變，可能在利多消息發布幾天內股價就會反應，而財報為每季發布一次，所以可將財報基本面因子往前移一季，較能切合當時公司表現。

- **分析方法**

本研究使用決策樹分析雖能較好的視覺化投資策略，但由於抓取因子少，所以在參數調整部分也有所限制(像是本研究之決策樹深度基本上只能小於等於 3)，故分析出的選股策略較為粗糙，往後可多進行交叉驗證提升準確率，或可應用其他機器學習方法，例如由決策樹延伸出的隨機森林方法進行分析，能處理較為複雜和龐大的資料。

六、 參考資料

許政文(2019)，以決策樹建構選股策略--以台灣上市公司為例，國立高雄師範大學事業經營學系碩士論文。

蔡正修(2007)，台灣上市電子類股價指數走勢預測之研究，國立成功大學統計學系碩士論文。