

Presentation
2024/01/11

論壇投資人情緒對台股之影響： 文字探勘與機器學習之應用

組員：嚴臨、李知祐、陳沛蓉、蘇振賢

目 錄

01

研究動機

02

研究流程

03

資料的來源與處理

04

回歸分析結果

05

機器學習：預測模型

06

總結

07

參考資料



01

專題研究動機



02

專題研究流程

研究動機



研究目的



文獻探討

- 探討情緒指標對大盤是否具有解釋力
- 了解情緒指標的預測能力

情緒指數編製

三因子

時間序列

機器學習

假說設立

假說設立



資料收集

PTT Stock版

TEJ 台灣經濟新報資料庫

玩股網

實證結果與分析

結論

03

資料：來源與處理

03

資料：來源與處理

本章節分為以下三部分：

(1)

爬蟲方法與情緒資料來源

(2)

情緒指數編制

(3)

財務資料來源與計算

(1)
爬蟲方法與
情緒資料來源

PTT的資料要怎麼獲取？

資料獲得-網路爬蟲

我們利用爬蟲將PTT上面的文章、日期、版友的留言等資訊獲取，而其中，我們實際需要的資料包含：

資料形式：曰資料

資料區間：2018/01/02-2023/10/30

```
[{'Tag': '推',
  'UserId': 'clamperni',
  'Content': '早',
  'Updatetime': '11/07 08:30'},
 {'Tag': '→',
  'UserId': 'zzzzzzzzzy',
  'Content': '小恩早雯晴早',
  'Updatetime': '11/07 08:30'},
 {'Tag': '推',
  'UserId': 'tearness',
  'Content': 'https://i.imgur.com/0',
  'Updatetime': '11/07 08:30'}, ...]
```



```
[{'早',
  '小恩早雯晴早',
  'https://i.imgur.com/0',
  '早',
  '空蛙最後救贖~' }]
```

資料前處理-缺失值處理

由於PTT版是一個由多人自願性發文維護的論壇，因此偶爾會有當天盤中/盤後閒聊沒人發文或是標題打錯導致爬蟲抓不到的狀況：

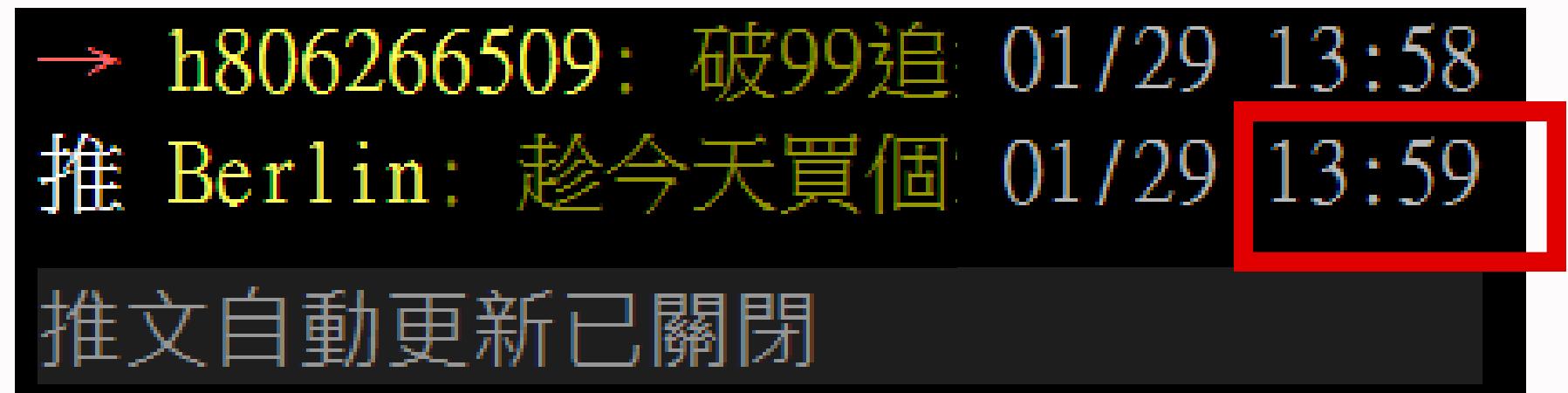


資料前處理-缺失值填補

正常情況：

盤中：當天8：30 ~ 14：00

盤後： 14：00 ~ 隔早8：30



留言會停在下一篇
盤中/後閒聊發布時

資料前處理-缺失值填補

異常情況：

推 f204137: <https://i.imgur.com> 02/11 13:57
推 amadeousMT: 橘子也只有跌一點 02/11 13:58
噓 iuakob: SOD 你是不是說某P 02/11 14:00
→ KadourZiani: 一堆散戶在放空 02/11 14:01

•
•
•



留言會繼續在上一篇
篇文章底下出現

推 abcgo: 聯電ADR怎麼了..... 02/12 08:21
→ appledick: 軋空列車新年正式 02/12 08:24
推 tomice: 早 02/12 09:04

推文自動更新已關閉

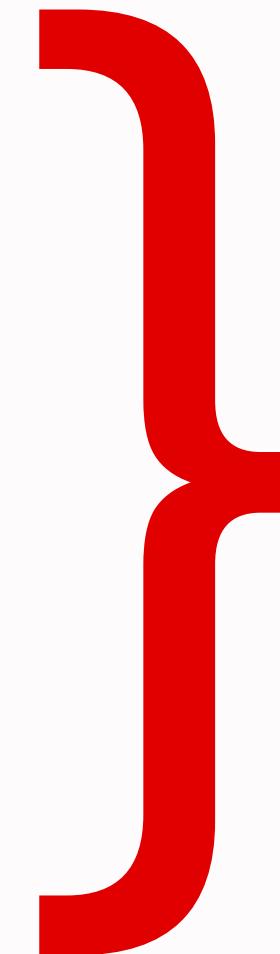
資料前處理-缺失值填補

異常情況：

噓 iuakob: SOD 你是不是說某P 02/11 14:00
→ KadourZiani: 一堆散戶在放空 02/11 14:01

•
•
•

推 abcgo: 聯電ADR怎麼了..... 02/12 08:21
→ appledick: 軋空列車新年正式 02/12 08:24



02/11 盤後閒聊

(2)

情緒指數編製方法



資料拿到了，那下一步呢？

情緒指數編製

01

尋找合適的情緒辭典

編制專屬的情緒詞庫

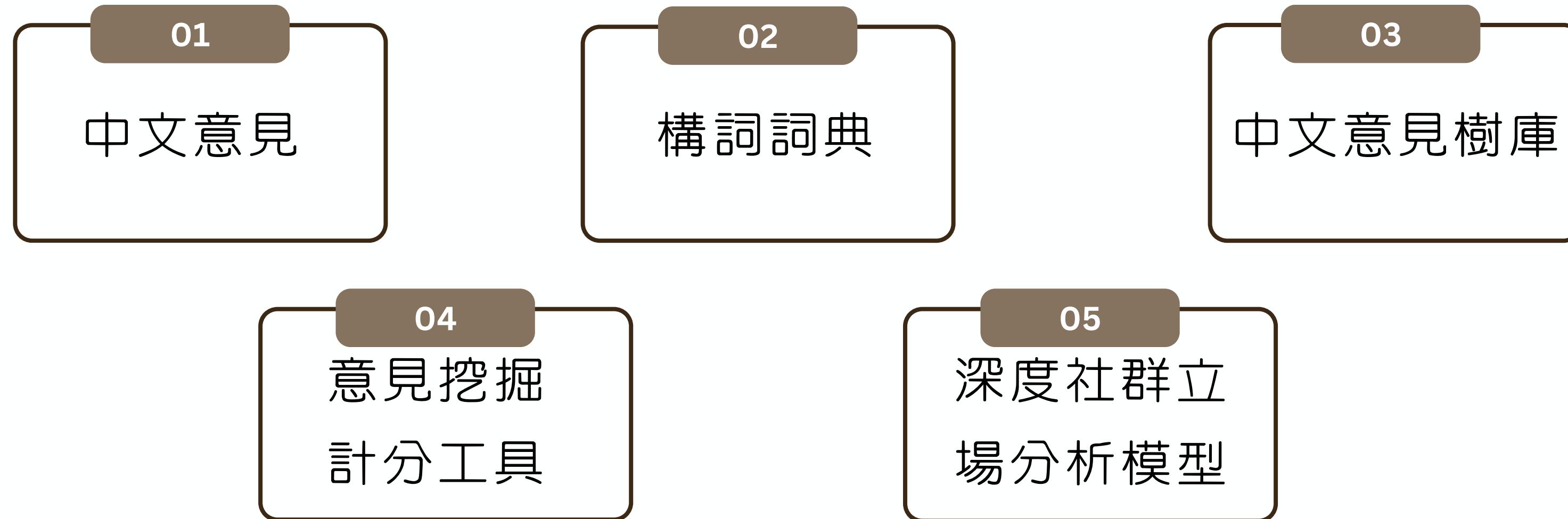
02

03

計算情緒分數

Step 1：尋找合適的情緒辭典

在製作模型之前，我們先查詢現今網路上是否有適合用於「中文」情緒分析之字辭典，最終找到了中研院製作的CSentiPackage，套件中包含多個可以用於中文情感語意分析研究的工具，例如：



ANTUSD

各欄位解釋如下：

Score	the CopeOpi numerical sentiment score
Pos	the number of positive annotations
Neu	the number of neutral annotations
Neg	the number of negative annotations
Non	non-opinionated annotations
Not	not-a-word annotations (which is collected from real online segmented data)

A	B	C	D	E	F	G
1	字詞	Score	Pos	Neu	Neg	Non
2	一下子爆	-0.38875	0	0	1	0
3	一下子爆	-0.20839	0	0	1	0
4	一夕成名	0	1	0	0	0
5	一大打擊	-0.40599	0	0	1	0
6	一大步	0.048718	1	0	0	0
7	一大指標	0.027758	1	0	0	0
8	一大突破	0.205476	1	0	0	0
9	一大問題	-0.06945	0	0	1	0
10	一大勝利	0.321512	1	0	0	0
11	一大障礙	-0.04173	0	0	1	0
12	一己之力	0.159852	1	0	0	0
13	一分打點	0.28774	1	0	0	0
14	一反	-0.454	0	0	1	0
15	一反前態	-0.252	0	0	1	0
16	一夫當關	0.224588	1	0	0	0
17	一巴掌	-0.221	0	0	1	0
18	一心想	0.293364	1	0	0	0
19	一手造成	-0.17479	0	0	1	0
20	一文不值	-0.11268	0	0	1	0
21	一日千里	0.04947	1	0	0	0

Step2：編制專屬的情緒詞庫

由於PTT屬於網路論壇，因此具有許多不存在於ANTUSD中的「鄉民用語」，為了讓情緒分析更準確，我們需要擴充辭典，並將PTT的"鄉民用語"也賦予情緒分數，完成這項工作需經過以下幾個步驟：

推 s155260	: 好厲害哦然後雞排發不起
推 lmc66	: 下週下船後準備佈局鋼鐵
→ guilty13	: 厲害個雕~嚇喊矇到就繼續喊，你也在信？
→ s155260	: 隨便賣一張都能讓YU酸閉嘴 可惜是盜來的對帳單
推 kmshy	: 影片有聯絡訊息 之前有先例被桶 保重
→ Nick0907	: 救救AIPC~~
→ mirror487	: 賺那麼多雞排該發了吧
推 eric1719	: 共享盜帳單 分身術變出朋友
推 Xenia1050	: k大 QQQQQ
噓 a331330	: 多軍大聲說出來！長榮250！下週金融帶頭衝
→ Xenia1050	: 我請阿文救我了 希望他能看到
推 winfisa	: 不要再提雞排了，他會回 要吃自己買不起喔
推 yiersan	: 老酥停損了 笑死

Step2-1：斷詞

本專題中的斷詞是指將PTT的留言斷成一個個詞，以讓字典判斷其分數，此步驟利用「**jieba斷詞**」工具執行，並將所有留言分別斷詞後儲存。

```
[‘ ’, ‘羨慕’, ‘yo’, ‘叔’, ‘新建’, ‘成本’, ‘8’, ‘塊’, ‘500’, ‘張’],
```

```
[‘ ’, ‘有人’, ‘今年’, ‘總資產’, ‘報酬’, ‘率’, ‘贏正’, ‘2’, ‘的’, ‘嗎’, ‘?’],
```

```
[‘ ’, ‘老蘇’, ‘真的’, ‘好’, ‘可憐’, ‘哪’]
```

[]內是一則留言，" "內則是一個個斷好的詞。

Step2-2: Word 2 Vec

```
array([[-3.0827671e-01, -8.0822521e-01, -5.6796223e-01, ...,
       -1.7690266e+00,  1.9922721e+00,  6.5078333e-02],
      [-1.8237098e+00, -1.8612362e-01, -6.3750273e-01, ...,
       -1.4657029e+00, -1.8036233e-01, -2.2183588e+00],
      [ 6.5391026e-02, -1.6161119e+00, -1.3539549e+00, ...,
       -1.4877031e+00,  2.7499900e+00, -1.0258013e+00],
      ...,
      [-8.7284511e-03,  1.0449322e-03,  6.4855663e-04, ...,
       -6.4985296e-03, -1.7772641e-03, -1.1236409e-02],
      [-3.7686035e-02,  2.0650290e-02, -3.8139217e-03, ...,
       -1.0175283e-02,  8.1982976e-03, -5.2785235e-03],
      [-2.6321745e-02,  2.1798255e-02, -1.6470809e-02, ...,
       -2.2229983e-02,  7.6693702e-03,  1.8507373e-02]], dtype=float32)
```

將單詞轉變成詞向量，藉由詞與詞之間的距離來判斷哪些詞屬性或意思相近，將「詞」本身的資訊進行數值化。

Step2-3： 降維

由於做完W2V 之後會有極大量的單詞（20萬以上），因此在維度過高的時候會發生**維度詛咒**（curse of dimensionality）——預測/分類能力通常是隨著維度數（變數）增加而上升，但在模型樣本數沒有繼續增加的情況下，預測/分類能力上升到一定程度之後，預測/分類能力會隨著維度的繼續增加而減小。當遇到這種維度過高的情況，就**需要降維**。

我們使用的降維方式為：

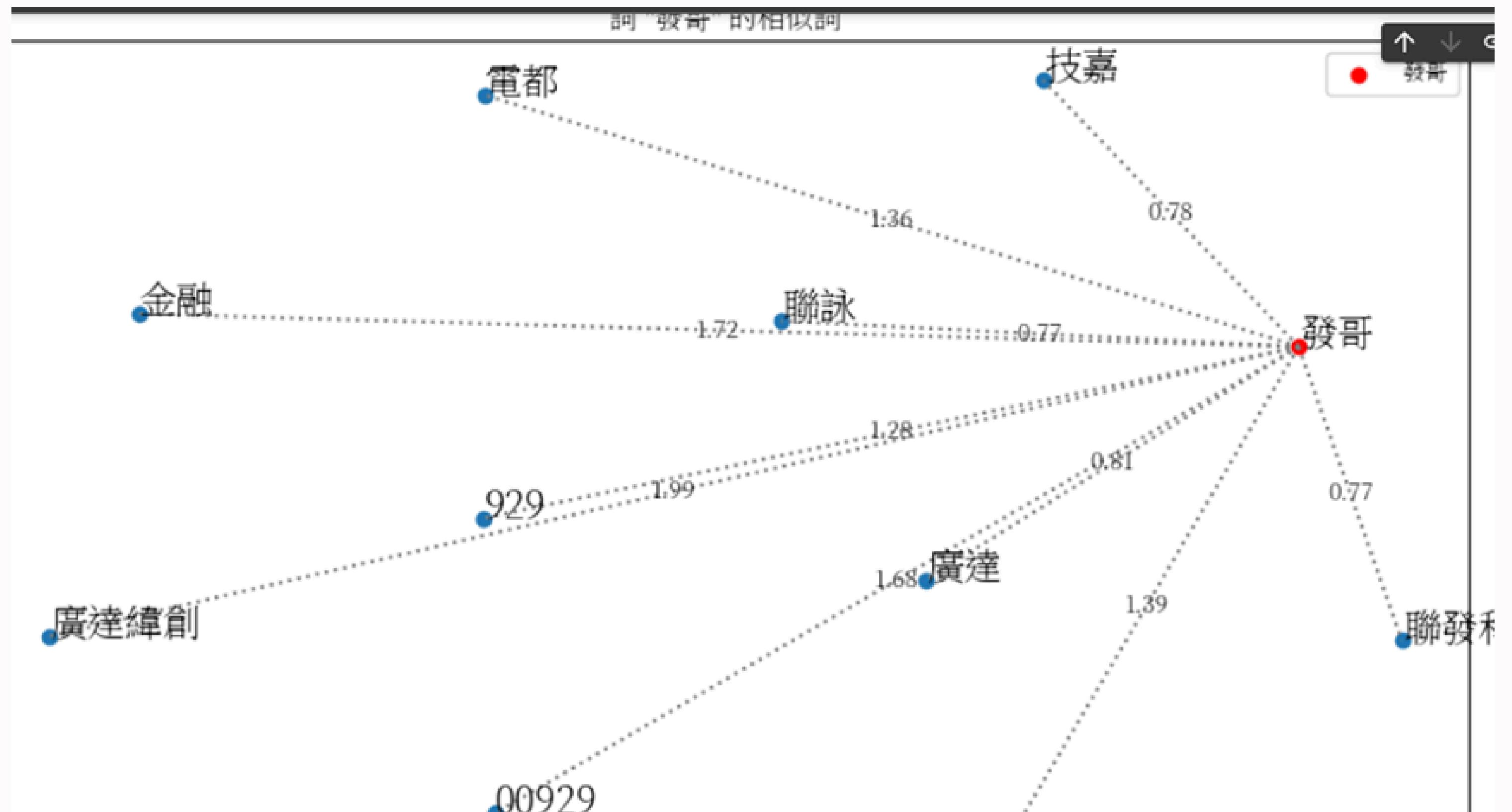
PCA

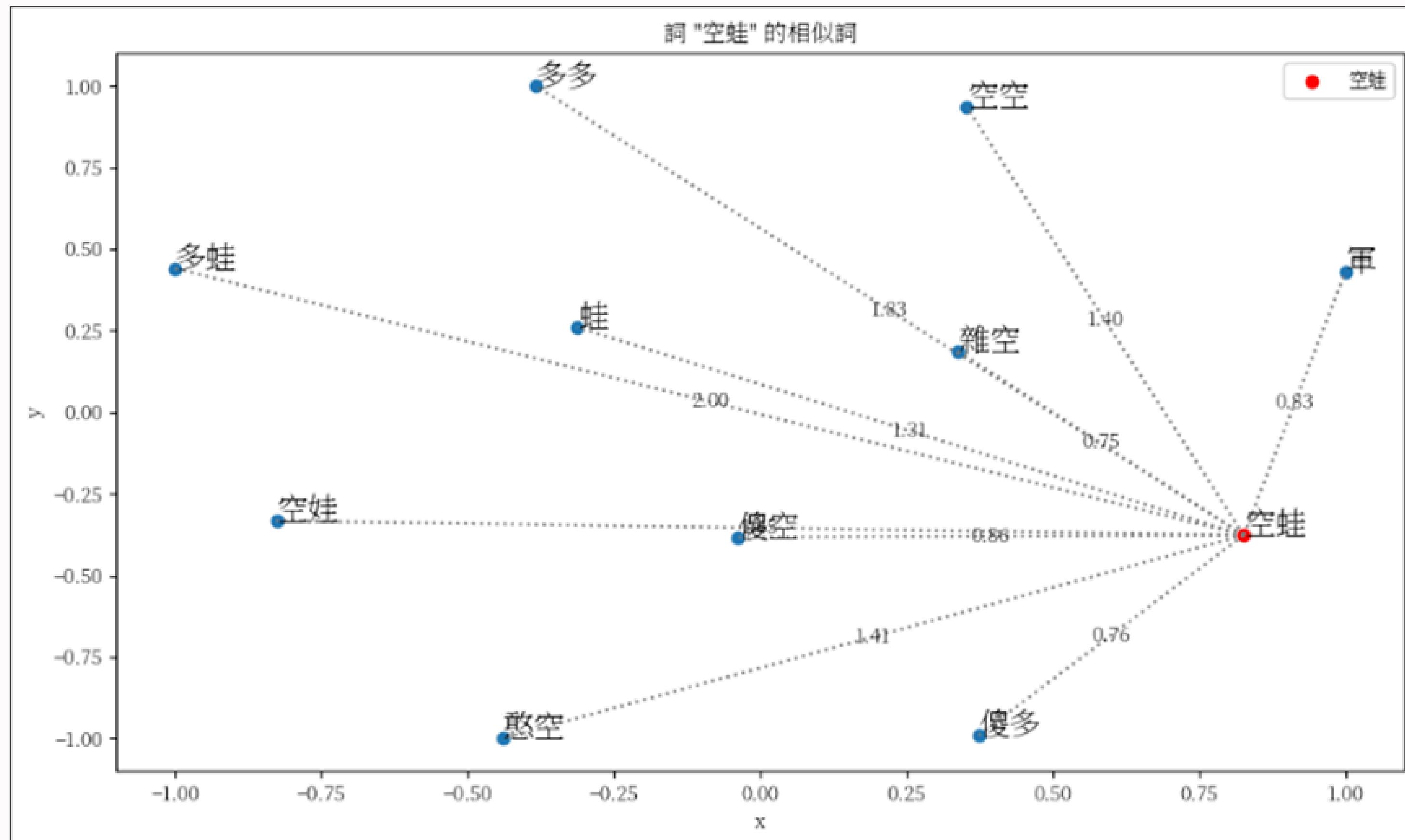
降維方法：PCA

線性回歸模型中的解釋變量之間由於存在精確相關關係或高度相關關係。看似相互獨立的指標本質上是相同的，是可以相互代替的。

PCA主要步驟：

- (1) 將數據標準化
- (2) 建立共變異數矩陣 (covariance matrix)
- (3) 利用奇異值分解 (SVD) 求得特徵向量 (eigenvector) 跟特徵值 (eigenvalue)
- (4) 特徵值會由大到小排列，選取 k 個特徵值與特徵向量
- (5) 原本的數據投影 (映射) 到特徵向量上，得到新的特徵數





Step3-1：計算分數

得到了每個詞與周遭所有詞的距離後，接著便要幫每個新詞定義分數。這一步我們使用的是 Inverse distance weighted (IDW)—— 使用已知稀疏點集進行多元插值，分配給未知點的值是使用已知點可用值的加權平均值計算。

$$d_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$$

$$w_i = \frac{1/d_i}{\sum_1^n 1/d_i}$$

$$Z_0 = \sum_{i=1}^{i=n} w_i * Z(X_i, Y_i)$$

- 計算未知點與所有點之 **距離**
- 計算**權重**: 將 **距離** 轉為 **權重**，兩點距離越小，已知點對新點的結果影響就越大，因此權重與距離為反比
- 計算插值點的值: 函數值進行 **加權平均**，得到新點的函數值，即為 **新詞的分數**

Step3-2：歸一化(Normalization)

得到每個單詞的情緒分數後，我們將每日留言中所有單詞情緒分數加總，並且將盤中留言及盤後留言分開計算，在得到每日盤中/盤後分數後，由於每日的留言數目、風向、情緒都大不相同，因此為了統一以及方便之後的運算，需進行歸一化(Normalization)，我們使用以下公式，將盤中/盤後每日情緒分數控制於[-1,0]之間，降低單位轉換或是區間對數據的影響

$$X_{nom} = \frac{X_t}{|X_{min}|} \subset [-1, 0]$$

(3)

財務資料來源與計算

五因子建構

資料來源

TEJ 台灣經濟新報資料庫

資料頻率 與範圍

日資料
2018/01/02-
2023/10/30

因子

市場溢酬因子 (MKT)、
市值規模因子 (SMB)、
帳面市值比因子 (HML)、
獲利能力因子 (RMW)、
投資策略因子 (CMA)

Step1：依規模分組

以「在外流通股數X收盤價」計算上市公司之市值，並將樣本股票按規模分為大規模(B)及小規模(S)。

	證券代碼	年月日	收盤價(元)	股價淨值比-TEJ	流通在外股數(千股)	日報酬率 %	市值	市值分組
0	1101 台泥	2018-01-02	20.4447	1.3800	4246509.0	0.2743	8.681860e+07	B
928	1101 台泥	2018-01-03	20.8922	1.4100	4246509.0	2.1888	8.871892e+07	B
1856	1101 台泥	2018-01-04	20.8083	1.4100	4246509.0	-0.4016	8.836263e+07	B
2784	1101 台泥	2018-01-05	20.8642	1.4100	4246509.0	0.2688	8.860001e+07	B
3712	1101 台泥	2018-01-08	21.3676	1.4500	4246509.0	2.4129	9.073771e+07	B
...
1368401	9958 世紀鋼	2023-10-25	170.0000	5.4541	235967.0	-1.1628	4.011439e+07	B
1369395	9958 世紀鋼	2023-10-26	164.0000	5.2616	235967.0	-3.5294	3.869859e+07	B
1370389	9958 世紀鋼	2023-10-27	167.5000	5.3739	235967.0	2.1341	3.952447e+07	B
1371383	9958 世紀鋼	2023-10-30	167.0000	5.3578	235967.0	-0.2985	3.940649e+07	B
1372377	9958 世紀鋼	2023-10-31	150.5000	4.8285	235967.0	-9.8802	3.551303e+07	B

Step2：形成投資組合

依股價淨值比分為低淨值市價比(L),中淨值市價比(M)，高淨值市價比(H)三組，對樣本進行雙向排序，形成六個投資組合S/L、S/M、S/H、B/L、B/M、B/H，如下表。

	證券代碼	年月日	收盤價(元)	股價淨值比-TEJ	流通在外股數(千股)	日報酬率 %	市值	市值分組	股價淨值比分組	組別
0	1101 台泥	2018-01-02	20.4447	1.3800	4246509.0	0.2743	8.681860e+07	B	M	BM
928	1101 台泥	2018-01-03	20.8922	1.4100	4246509.0	2.1888	8.871892e+07	B	M	BM
1856	1101 台泥	2018-01-04	20.8083	1.4100	4246509.0	-0.4016	8.836263e+07	B	M	BM
2784	1101 台泥	2018-01-05	20.8642	1.4100	4246509.0	0.2688	8.860001e+07	B	M	BM
3712	1101 台泥	2018-01-08	21.3676	1.4500	4246509.0	2.4129	9.073771e+07	B	M	BM
...
1368401	9958 世紀鋼	2023-10-25	170.0000	5.4541	235967.0	-1.1628	4.011439e+07	B	H	BH
1369395	9958 世紀鋼	2023-10-26	164.0000	5.2616	235967.0	-3.5294	3.869859e+07	B	H	BH
1370389	9958 世紀鋼	2023-10-27	167.5000	5.3739	235967.0	2.1341	3.952447e+07	B	H	BH
1371383	9958 世紀鋼	2023-10-30	167.0000	5.3578	235967.0	-0.2985	3.940649e+07	B	H	BH
1372377	9958 世紀鋼	2023-10-31	150.5000	4.8285	235967.0	-9.8802	3.551303e+07	B	H	BH

Step3：加權

對於每日所形成之六個投資組合，分別計算 t 日 之加權平均報酬率- $R(S/L)$ 、 $R(S/M)$ 、 $R(S/H)$ 、 $R(B/L)$ 、 $R(B/M)$ 、 $R(B/H)$ 。

Step4：估計SMB、HML因子

依下列公式估計單日 SMB 和 HML 因子：

$$SMB = (R(S/L) + R(S/M) + R(S/H)) / 3 - (R(B/L) + R(B/M) + R(B/H)) / 3$$

$$HML = (R(B/H) + R(S/H)) / 2 - (R(S/L) + R(B/L)) / 2$$

Step5：估計RMW（獲利能力）因子

由於TEJ資料庫無RMW相關資料，因此我們以上市公司之ROE作為獲利能力指標，計算公式為(營業毛利-營業費用)/ (營業收入淨額-營業費用)。（張柯，2017）

證券代碼	年月	月份	營業費用	營業毛利	營業成本	營業收入淨額	年月日	年份	ROE	
39133	1101 台泥	2018/3/31	3	1139949.0	5790640.0	18388574.0	24179214.0	2018-03-31	2018	0.201859
37383	1101 台泥	2018/6/29	6	2429709.0	16059356.0	41440890.0	57500246.0	2018-06-29	2018	0.247494
35698	1101 台泥	2018/9/28	9	3688295.0	24664574.0	64992780.0	89657354.0	2018-09-28	2018	0.243998
33904	1101 台泥	2018/12/28	12	5410638.0	33591539.0	91003063.0	124594602.0	2018-12-28	2018	0.236449
32205	1101 台泥	2019/3/29	3	1454902.0	6487864.0	18868464.0	25356328.0	2019-03-29	2019	0.210572
...	
9032	9962 有益	2022/9/30	9	75427.0	260060.0	2268983.0	2529043.0	2022-09-30	2022	0.075249
7228	9962 有益	2022/12/30	12	111499.0	362567.0	3168241.0	3530808.0	2022-12-30	2022	0.073427
5417	9962 有益	2023/3/31	3	24616.0	78916.0	833372.0	912288.0	2023-03-31	2023	0.061171
3613	9962 有益	2023/6/30	6	47448.0	136788.0	1573072.0	1709860.0	2023-06-30	2023	0.053741
1802	9962 有益	2023/9/28	9	70519.0	190638.0	2366081.0	2556719.0	2023-09-28	2023	0.048314

後續步驟如建立投資組合、計算加權等同前述SMB、HML因子，在此不贅述。

Step6：估計CMA（投資策略）因子

由於TEJ資料庫無CMA相關資料，因此我們以上市公司之資產成長率作為投資模式指標，計算公式為：資產成長率=(當季度資產總額 - 前一季度資產總額)／前一季度資產總額。

	證券代碼	年月日	年份	月份	月份分組	月份_組別	資產總額	年月	上個月年月	上個月年	上個月月	上個月月份分組	年份_上個月	月份_上個月	資產總額_上個月	資產成長率
0	1101 台泥	2018-01-02	2018	1	3	3.0	283088584.0	2018-01-01	2017-12-01	2017	12	12	2017.0	12.0	272557049.0	0.03864
1	1101 台泥	2018-01-03	2018	1	3	3.0	283088584.0	2018-01-01	2017-12-01	2017	12	12	2017.0	12.0	272557049.0	0.03864
2	1101 台泥	2018-01-04	2018	1	3	3.0	283088584.0	2018-01-01	2017-12-01	2017	12	12	2017.0	12.0	272557049.0	0.03864
3	1101 台泥	2018-01-05	2018	1	3	3.0	283088584.0	2018-01-01	2017-12-01	2017	12	12	2017.0	12.0	272557049.0	0.03864
4	1101 台泥	2018-01-08	2018	1	3	3.0	283088584.0	2018-01-01	2017-12-01	2017	12	12	2017.0	12.0	272557049.0	0.03864

後續步驟如建立投資組合、計算加權等同前述SMB、HML因子，在此不贅述。

04

資料：回歸分析結果

04

回歸分析結果

本章節分為以下二部分：

(1)

Fama French 回歸模型

(2)

時間序列分析

(1)

模型使用



m:盤中 ; M:盤後

Fama French factor model (盤中版本)

Note: 選擇0050以及0051作為標的

01

單因子模型

$$R_t - Rf_t = a_1 + b_1 * m_t + \epsilon_{1,t}$$

Define it to be excess return

02

三因子模型

$$R_t - Rf_t = a_2 + b_2 * m_t + c_2 * MKT_t + d_2 * SMB_t + e_2 * HML_t + \epsilon_{2,t}$$

03

五因子模型

$$R_t - Rf_t = a_3 + b_3 * m_t + c_3 * MKT_t + d_3 * SMB_t + e_3 * HML_t + f_3 * RMW_t + g_3 * CMA_t + \epsilon_{3,t}$$

Fama French factor model (盤後版本)

Note: 選擇0050作為標的

01

單因子模型

$$R_t - Rf_t = a_4 + b_4 * M_t + \epsilon_{4,t}$$

02

三因子模型

$$R_t - Rf_t = a_5 + b_5 * M_t + c_5 * MKT_t + d_5 * SMB_t + e_5 * HML_t + \epsilon_{5,t}$$

03

五因子模型

$$R_t - Rf_t = a_6 + b_6 * M_t + c_6 * MKT_t + d_6 * SMB_t + e_6 * HML_t + f_6 * RMW_t + g_6 * CMA_t + \epsilon_{6,t}$$

檢驗方式

全時間	樣本區間：2018/06/01-2023/10/30
漲跌時間段區分	<p>上漲區間(Bullish) : 2019/01/14-2020/01/20, 2020/03/23-2022/01/21, 2022/11/07-2023/10/30</p> <p>下跌區間(Bearish) : 2018/06/01-2019/01/11, 2020/01/30-2020/03/20, 2022/01/24-2022/11/04</p>
波動區分	<p>small: excess return在0.5%以上</p> <p>middle: excess return在1%以上</p> <p>big: excess return在2%以上</p>

(2)

0050檢驗

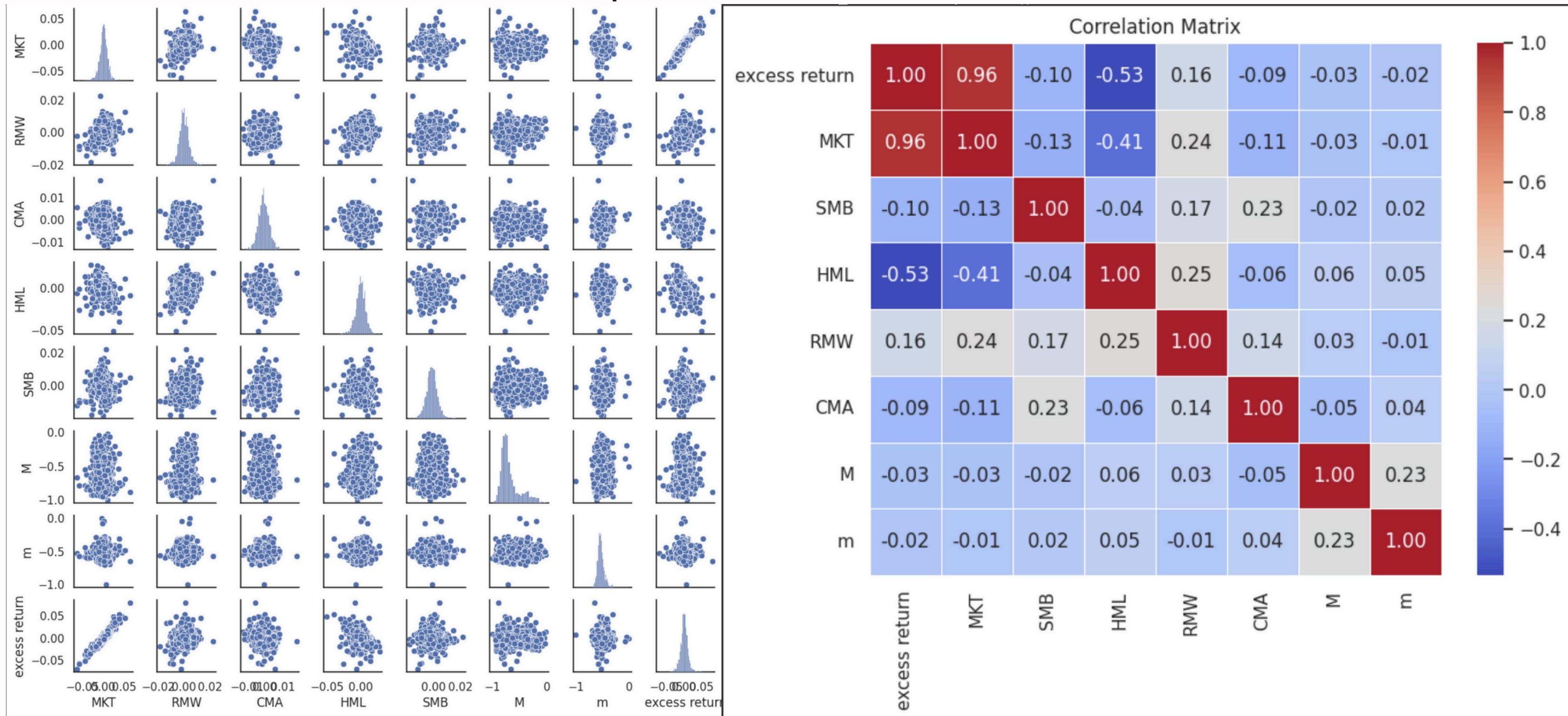
基本描述(M:盤後 ; m:盤中)

1320 row x 8 columns

	excess return	MKT	SMB	HML	CMA	RMW	M	m
mean	0.00050	0.00034	-0.00108	-0.00003	-0.00058	-0.00004	-0.68090	-0.53132
std	0.01172	0.01069	0.00424	0.00076	0.00303	0.00378	0.18015	0.06734
min	-0.07029	-0.06314	-0.01784	-0.00509	-0.01211	-0.01836	-1.00000	-1.00000
25%	-0.00541	-0.00496	-0.00361	-0.00044	-0.00249	-0.00238	-0.80097	-0.57175
50%	0.00052	0.00069	-0.00109	0.00000	-0.00064	-0.00009	-0.73778	-0.53886
75%	0.00687	0.00639	0.00144	0.00447	0.00137	0.00225	-0.62321	-0.50043
max	0.07949	0.06366	0.02186	0.02752	0.01773	0.02230	-0.02457	-0.00533

Pairplot & Correlation Matrix

<註> FAMA FRENCH 認為 predictors 之間係數小於0.75則不影響迴歸進行



Fama French factor model (盤中)

結果發現盤中情緒指標與超額報酬無關，即使以漲跌時間區間作為區分亦然

盤中		m_t	MKT_t	SMB_t	HML_t	RMW_t	CMA_t	a
All	Model(1)	-0.0034 (-0.718)	-	-	-	-	-	-0.0013 (-0.516)
	Model(2)	-0.0010 (-0.819)	0.9739*** (119.669)	0.0344* (1.832)	-0.2662*** (-23.432)	-	-	-0.0004 (-0.613)
	Model(3)	-0.0009 (-0.803)	0.9777*** (106.849)	0.0487** (2.466)	-0.2605*** (-20.462)	-0.0335 (-1.360)	-0.0460* (-1.687)	-0.0004 (-0.617)
	Model(1)	0.0015 (0.294)	-	-	-	-	-	0.0022 (0.777)
	Model(2)	-0.0009 (-0.610)	0.9511*** (83.508)	0.0143 (0.665)	-0.2814*** (-20.859)	-	-	-0.0003 (-0.431)
	Model(3)	-0.0008 (-0.553)	0.9482*** (74.261)	0.0191 (0.841)	-0.2851*** (-18.389)	0.0047 (0.170)	-0.0473 (-1.577)	-0.0003 (-0.405)
Bullish	Model(1)	-0.0076 (-0.757)	-	-	-	-	-	-0.0056 (-1.061)
	Model(2)	-0.0008 (-0.379)	0.9952*** (80.952)	0.0700* (1.801)	-0.2548*** (-11.037)	-	-	-0.0002 (-0.218)
	Model(3)	-0.0009 (-0.416)	1.0094*** (70.389)	0.1059** (2.572)	-0.2362*** (-9.597)	-0.1236** (-2.248)	-0.0520 (-0.828)	-0.0003 (-0.252)

Fama French factor model (盤中)

同樣的，結果證實盤中情緒指標與超額報酬無關，即使以漲跌波動度作為區分亦然

盤中 (波動度)		m_t	MKT_t	SMB_t	HML_t	RMW_t	CMA_t	a
Small	Model(1)	-0.0040 (-0.485)	-	-	-	-	-	-0.0013 (-0.301)
	Model(2)	-0.0004 (-0.227)	0.9944*** (109.059)	0.0229 (0.889)	-0.2759*** (-19.481)	-	-	-3.027e-05 (-0.035)
	Model(3)	-0.0004 (-0.216)	1.0003*** (93.635)	0.0356 (1.329)	-0.2675*** (-16.770)	-0.0456 (-1.341)	-0.0299 (-0.816)	-2.86e-05 (-0.033)
	Model(1)	-0.0046 (-0.296)	-	-	-	-	-	-0.0018 (-0.220)
	Model(2)	0.0006 (0.225)	1.0137*** (97.458)	0.0014 (0.036)	-0.2836*** (-16.315)	-	-	0.0006 (0.488)
	Model(3)	0.0005 (0.185)	1.0207*** (78.625)	0.0139 (0.346)	-0.2744*** (-13.775)	-0.0534 (-1.044)	-0.0074 (-0.132)	0.0006 (0.447)
	Model(1)	0.0703 (1.322)	-	-	-	-	-	0.0378 (1.355)
	Model(2)	-0.0057 (-0.881)	1.0434*** (66.529)	0.0463 (0.562)	-0.3239*** (-9.796)	-	-	-0.0022 (-0.656)
	Model(3)	-0.0042 (-0.643)	1.0567*** (50.435)	0.0682 (0.800)	-0.3053*** (-8.119)	-0.1285 (-1.203)	-0.0145 (-0.121)	-0.0015 (-0.447)

Fama French factor model (盤後)

結果發現盤後情緒指標與超額報酬無關，即使以漲跌時間區間作為區分亦然

盤後		M_t	MKT_t	SMB_t	HML_t	RMW_t	CMA_t	\hat{a}
All	Model(1)	-0.0009 (-0.491)	-	-	-	-	-	-0.0001 (-0.080)
	Model(2)	0.0001 (0.310)	0.9738*** (119.573)	0.0342* (1.818)	-0.2668*** (-23.497)	-	-	0.0002 (0.700)
	Model(3)	0.0001 (0.340)	0.9775*** (106.826)	0.0485** (2.455)	-0.2613*** (-20.556)	-0.0329 (-1.339)	-0.0469* (-1.721)	0.0002 (0.687)
Bullish	Model(1)	-0.0002 (-0.080)	-	-	-	-	-	0.0013 (0.823)
	Model(2)	8.771e-05 (0.155)	0.9507*** (83.586)	0.0141 (0.655)	-0.2819*** (-20.932)	-	-	0.0002 (0.475)
	Model(3)	0.0001 (0.340)	0.9775*** (106.826)	0.0485** (2.455)	-0.2613*** (-20.556)	-0.0329 (-1.339)	-0.0469* (-1.721)	0.0002 (0.687)
Bearish	Model(1)	0.0008 (0.236)	-	-	-	-	-	-0.0011 (-0.503)
	Model(2)	0.0001 (0.199)	0.9951*** (80.874)	0.0697* (1.793)	-0.2557*** (-11.057)	-	-	0.0003 (0.554)
	Model(3)	0.0002 (0.323)	1.0094*** (70.382)	0.1057** (2.567)	-0.2372*** (-9.657)	-0.1242** (-2.251)	-0.0519 (-0.824)	0.0003 (0.673)

Fama French factor model (盤後)

同樣的，結果證實盤後情緒指標與超額報酬無關，即使以漲跌波動度作為區分亦然

盤中 (波動度)		m_t	MKT_t	SMB_t	HML_t	RMW_t	CMA_t	a
Small	Model(1)	-0.0040 (-0.485)	-	-	-	-	-	-0.0013 (-0.301)
	Model(2)	-0.0004 (-0.227)	0.9944*** (109.059)	0.0229 (0.889)	-0.2759*** (-19.481)	-	-	-3.027e-05 (-0.035)
	Model(3)	-0.0004 (-0.216)	1.0003*** (93.635)	0.0356 (1.329)	-0.2675*** (-16.770)	-0.0456 (-1.341)	-0.0299 (-0.816)	-2.86e-05 (-0.033)
	Model(1)	-0.0046 (-0.296)	-	-	-	-	-	-0.0018 (-0.220)
	Model(2)	0.0006 (0.225)	1.0137*** (97.458)	0.0014 (0.036)	-0.2836*** (-16.315)	-	-	0.0006 (0.488)
	Model(3)	0.0005 (0.185)	1.0207*** (78.625)	0.0139 (0.346)	-0.2744*** (-13.775)	-0.0534 (-1.044)	-0.0074 (-0.132)	0.0006 (0.447)
Middle	Model(1)	0.0703 (1.322)	-	-	-	-	-	0.0378 (1.355)
	Model(2)	-0.0057 (-0.881)	1.0434*** (66.529)	0.0463 (0.562)	-0.3239*** (-9.796)	-	-	-0.0022 (-0.656)
	Model(3)	-0.0042 (-0.643)	1.0567*** (50.435)	0.0682 (0.800)	-0.3053*** (-8.119)	-0.1285 (-1.203)	-0.0145 (-0.121)	-0.0015 (-0.447)

(3)

時間序列分析
(針對加權指數作為標的)

Vector Autoregressive model(3)

簡介：VAR 模型是一種時間序列模型分析方法，用於解釋變數之間是否有相關性。其基本思想是將所有變數都視為同時獨立地受到自身和其他變數的影響。（選擇 lag =3 ）

$$R_t = a_1 + \sum_{j=1}^3 \beta_{1,j} * R_{t-j} + \sum_{j=1}^3 \gamma_{1,j} * M_{t-j} + \epsilon_{1,t}$$

盤後

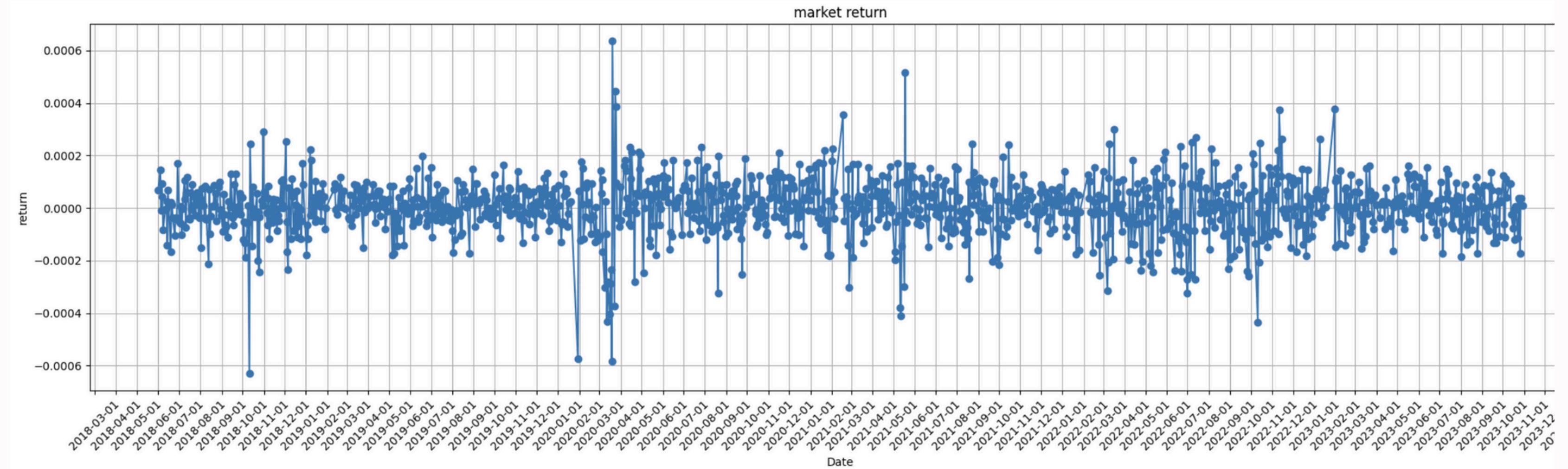
$$M_t = a_2 + \sum_{j=1}^3 \beta_{2,j} * R_{t-j} + \sum_{j=1}^3 \gamma_{2,j} * M_{t-j} + \epsilon_{2,t}$$

$$R_t = a_3 + \sum_{j=1}^3 \beta_{3,j} * R_{t-j} + \sum_{j=1}^3 \gamma_{3,j} * m_{t-j} + \epsilon_{3,t}$$

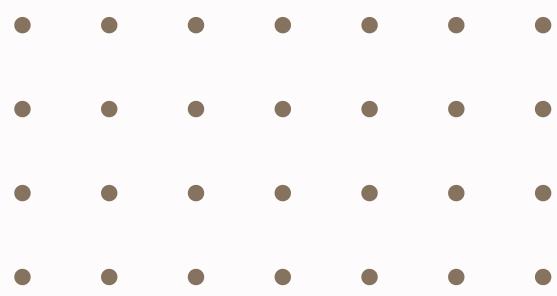
盤中

$$m_t = a_4 + \sum_{j=1}^3 \beta_{4,j} * R_{t-j} + \sum_{j=1}^3 \gamma_{4,j} * m_{t-j} + \epsilon_{4,t}$$

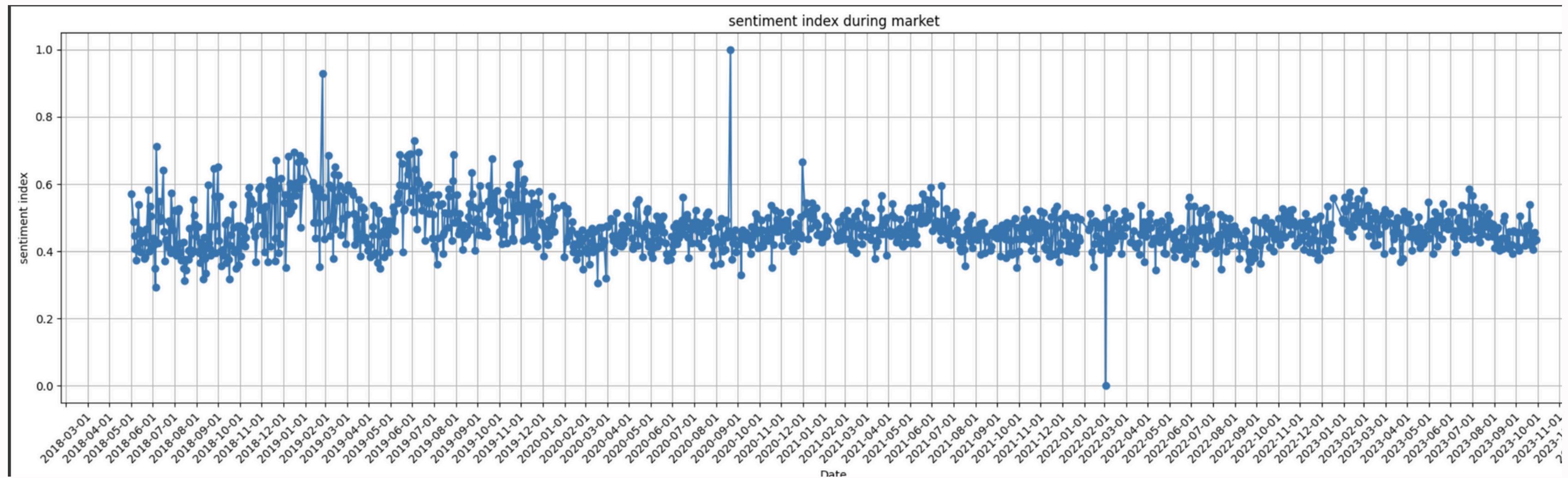
報酬(stationary)



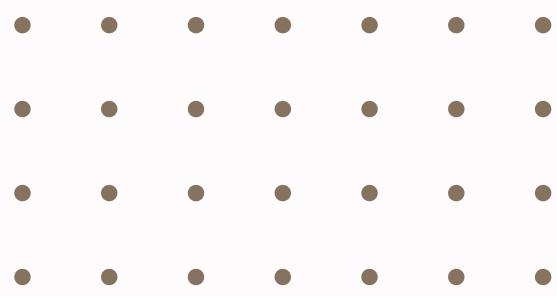
p-value -0.000000



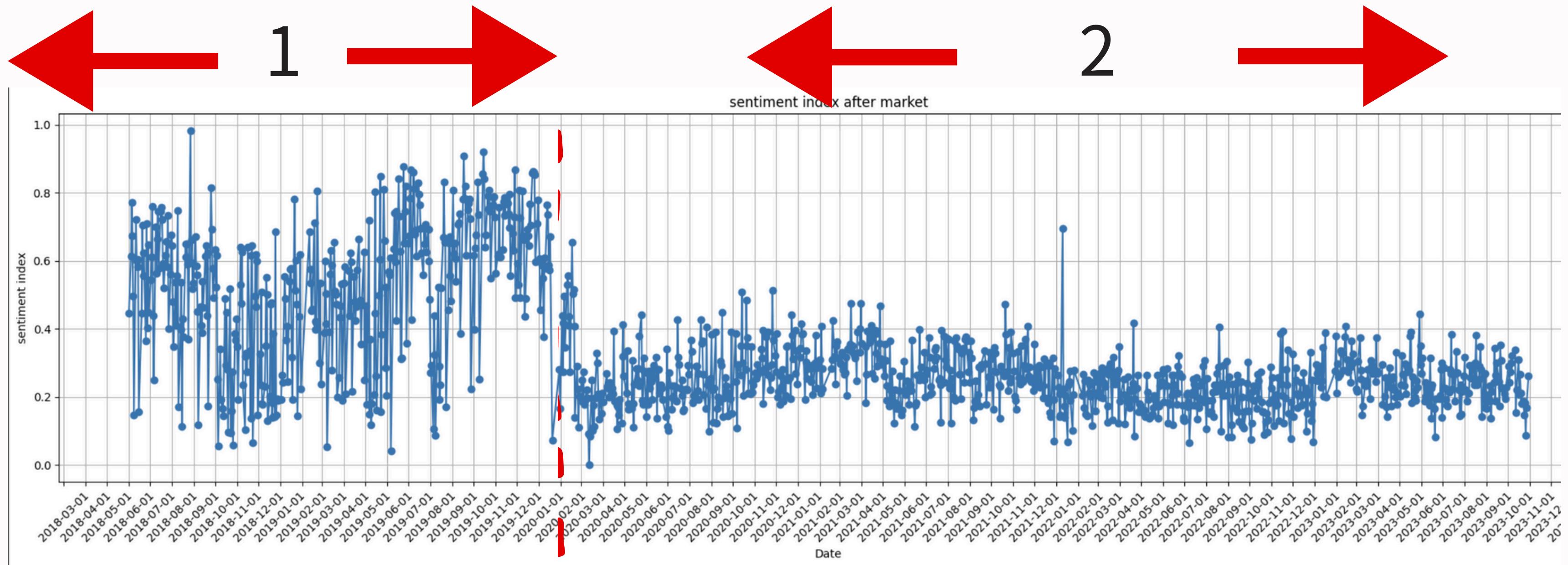
盤中情緒指數(stationary)



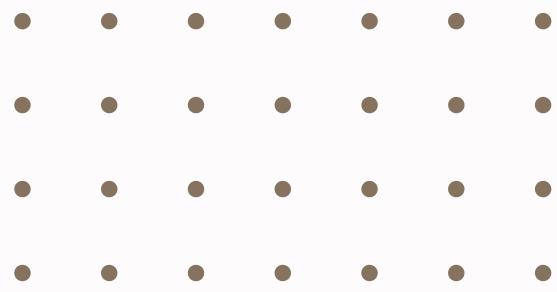
p-value: 0.000011



盤後情緒指數 (non-stationary)



p-value : 0.1078 ; 2020/01/20 出現一個顯著的drop
區間1 p-value: -0.043938 ; 區間2 p-value: 0.000001



Vector Autoregressive model(盤中)

盤中區間 1	m_{t-1}	m_{t-2}	m_{t-3}	R_{t-1}	R_{t-2}	R_{t-3}
R_t	0.007106*	0.004342	-0.002323	-0.088485**	0.071877*	-0.005215
	(1.397)	(0.830)	(-0.463)	(-1.740)	(1.403)	(-0.102)
m_t	0.245948***	0.066073	0.195004***	1.250259***	0.343253	0.532735
	(4.949)	(1.293)	(3.975)	(2.517)	(0.686)	(1.068)
盤中區間 2	m_{t-1}	m_{t-2}	m_{t-3}	R_{t-1}	R_{t-2}	R_{t-3}
R_t	0.004197	-0.000144	-0.005025	0.017137	0.089283***	0.004048
	(0.533)	(-0.018)	(-0.645)	(0.511)	(2.697)	(0.122)
m_t	0.124497***	0.101199***	0.062400**	0.399778***	0.054282	0.073158
	(3.719)	(3.025)	(1.885)	(2.804)	(0.386)	(0.518)

<註> 上方數值為係數 (coefficient) ；下方括號內為 t 值，星號 * , ** , *** 分別對應 90% , 95% , 99% 顯著性

Vector Autoregressive model(盤後)

盤後區間 1	M_{t-1}	M_{t-2}	M_{t-3}	R_{t-1}	R_{t-2}	R_{t-3}
R_t	0.002960*	0.001028	0.001352	-0.086124**	0.063492	-0.020520
	(1.330)	(0.460)	(0.623)	(-1.704)	(1.233)	(-0.397)
M_t	0.223064***	0.149429***	0.169688***	0.223064***	0.149429***	0.169688
	(4.481)	(2.991)	(3.497)	(3.507)	(2.912)	(0.207)
盤後區間 2	M_{t-1}	M_{t-2}	M_{t-3}	R_{t-1}	R_{t-2}	R_{t-3}
R_t	0.006316	-0.004072	0.003520	0.017046	0.088690***	-0.001153
	(1.281)	(-0.819)	(0.731)	(0.512)	(2.700)	(-0.035)
M_t	0.226684***	0.147500***	0.137677***	0.563296***	0.171226	0.216227
	(6.879)	(4.438)	(4.279)	(2.532)	(0.780)	(0.981)

<註> 上方數值為係數 (coefficient) ；下方括號內為 t 值，星號 * , ** , *** 分別對應 90% , 95% , 99% 顯著性

05

機器學習：預測模型

05

預測模型能力

本章節分為以下四部分：

(1)

使用資料說明

(2)

Logistic
Regression

(3)

Random
Forest

(4)

Support Vector
Machine

(1)

特徵資料說明

特徵資料說明-新增資料

除了使用原本的五因子、情緒分數，我們加入了其它我們認為可能會影響到台灣大盤走勢的變數，分別是：

1. 大盤相關波動度

台灣VIX波動度

2. 商品 (return %)

Gold、Crude Oil

3. 匯率

美元兌台幣匯率

4. 美股四大指數 (開、高、低、收、交易量、return %)

DJI(道瓊)、GSPC(標普)、IXIC(那斯達克)、SOX(費城)

特徵資料說明-資料預處理

我們將所有資料做了一些處理，讓模型能更好的利用全方面的資料：

1. 針對美股四大指數，新增：

- (1). 最高價/開盤價 (High/Open)
- (2). 最低價/開盤價 (Low/Open)
- (3). 最高價/最低價 (High/Low)
- (4). 最高價/收盤價 (High/Close)
- (5). 最低價/收盤價 (Low/Close)
- (6). 交易量與前一天的變化百分比 (Volume_change)

2. 針對美元兌台幣匯率、台灣VIX波動、盤中/盤後情緒分數，新增：

與 t 時除以前一天 $t-1$ 時的變動百分比

特徵資料說明-資料預處理

3. 資料normalize：

為了避免不同資料的尺度不同造成差異，我們將所有特徵資料全部標準化至 [-1 , 1] 的區間

4. return二元分類化：

由於我們使用的3個模型都是以分類為主要功能的模型，因此我們將大盤的報酬率return轉成二元分類的形式，若當天 $\text{return} > 0$ 則歸類為1，若當天 $\text{return} < 0$ 則歸類到0。

預測目標：大盤在第 $t+1$ 天的漲跌

特徵資料說明-不同形式預測

1. 刪除return在 正負0.5% 區間的資料：

由於在實際應用上過小的報酬在實際投資操作上並沒有被正確分類出來的價值，因此我們將報酬漲幅過小的數據刪除，同時也可以避免因為在0周圍過小的漲幅對分類型模型的影響。

2. 刪除return在 正負1% 區間的資料（剩餘資料量352天）

3. 刪除return在 正負2% 區間的資料（剩餘資料量**75**天）

(2)

Logistic Regression

Logistic Regression-模型公式

我們將模型套用到Logistic Regression的模型內，模型公式如下：

$$p(Y_{t+1}) = \frac{e^{\beta_0 + \beta_1 MKT_t + \beta_2 SMB_t + \dots + \beta_{44} Jeiba_M_change_t + \beta_{45} TVIX_change_t}}{1 + e^{\beta_0 + \beta_1 MKT_t + \beta_2 SMB_t + \dots + \beta_{44} Jeiba_M_change_t + \beta_{45} TVIX_change_t}}$$

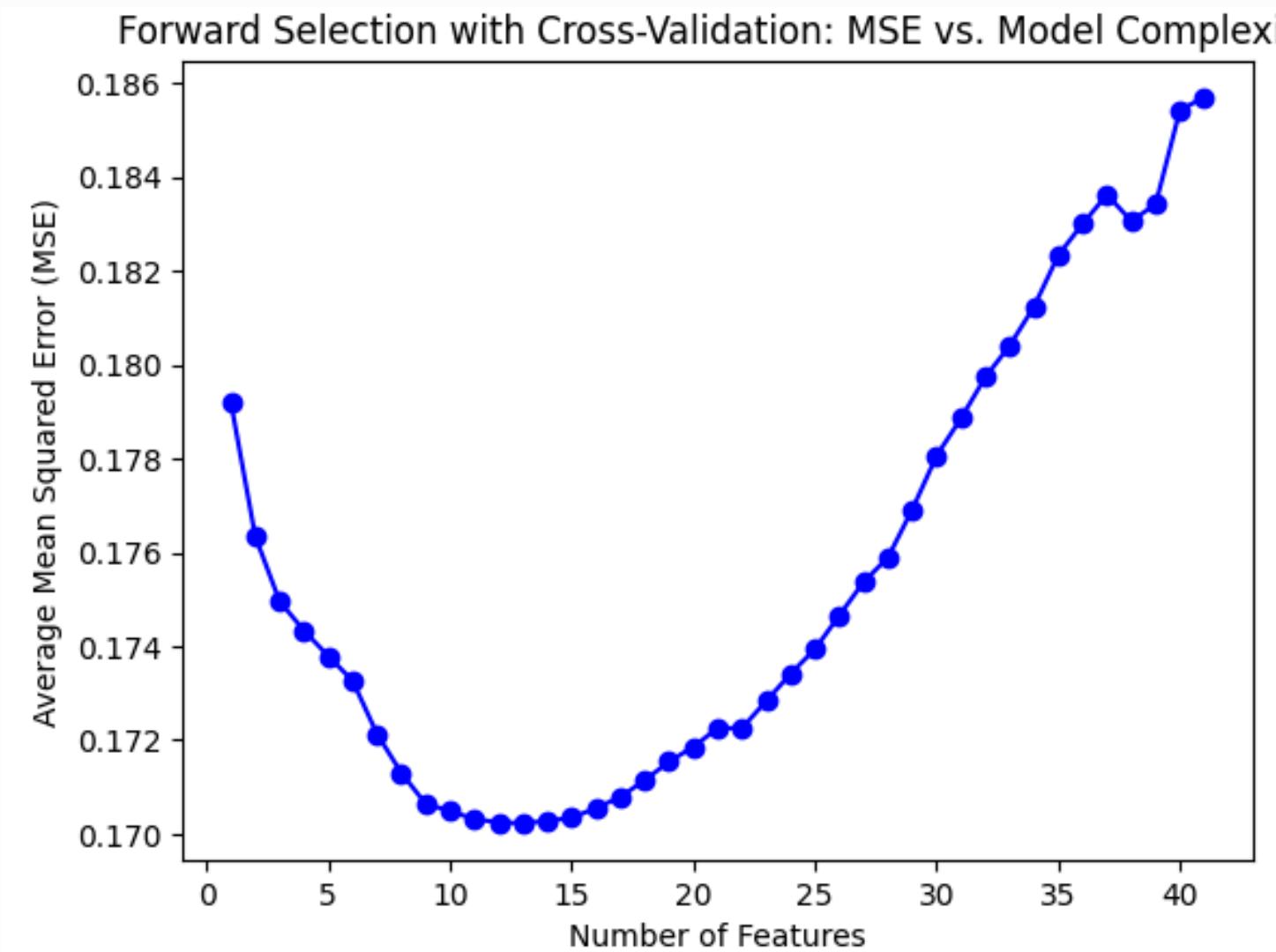
$p(Y_{t+1})$: 大盤在t+1天會是漲的機率

$\beta_0, \beta_1, \beta_2, \dots, \beta_{44}, \beta_{45}$: 截距項及各特徵變數的係數

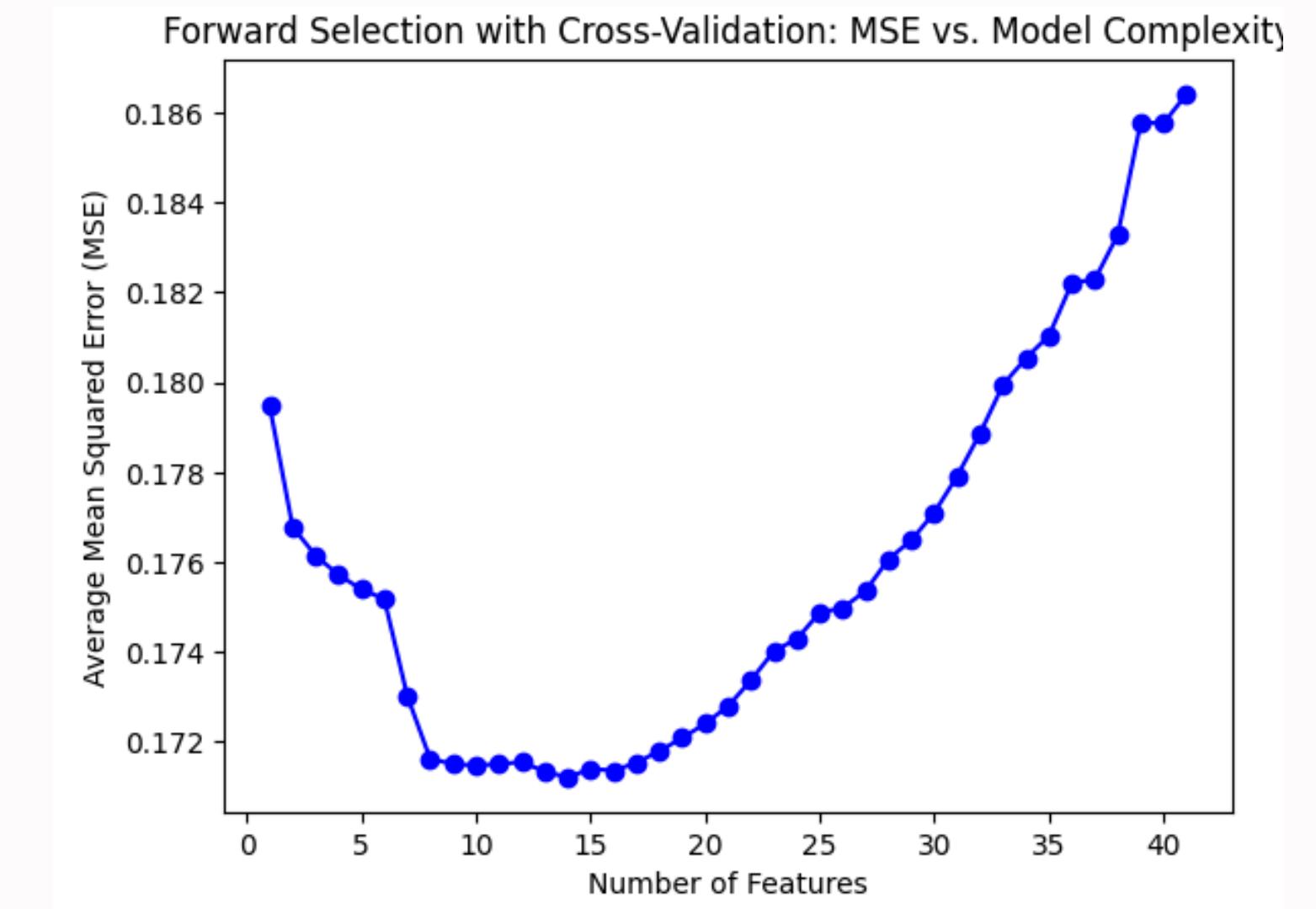
$MKT_t, SMB_t, \dots, Jeiba_M_change_t, TVIX_change_t$: 特徵變數在時間t的數值

Logistic - 特徵挑選- 刪除- 0.5~+0.5版本

由於Logistic模型若有太多沒有幫助的特徵變數會影響到模型的效能，因此我們結合forward selection以及cross validation來檢測模型在選取變數上的順序性並找尋最適合的變數組合，而我們使用了在cross validation中常用的5 Fold和10 Fold來檢測：



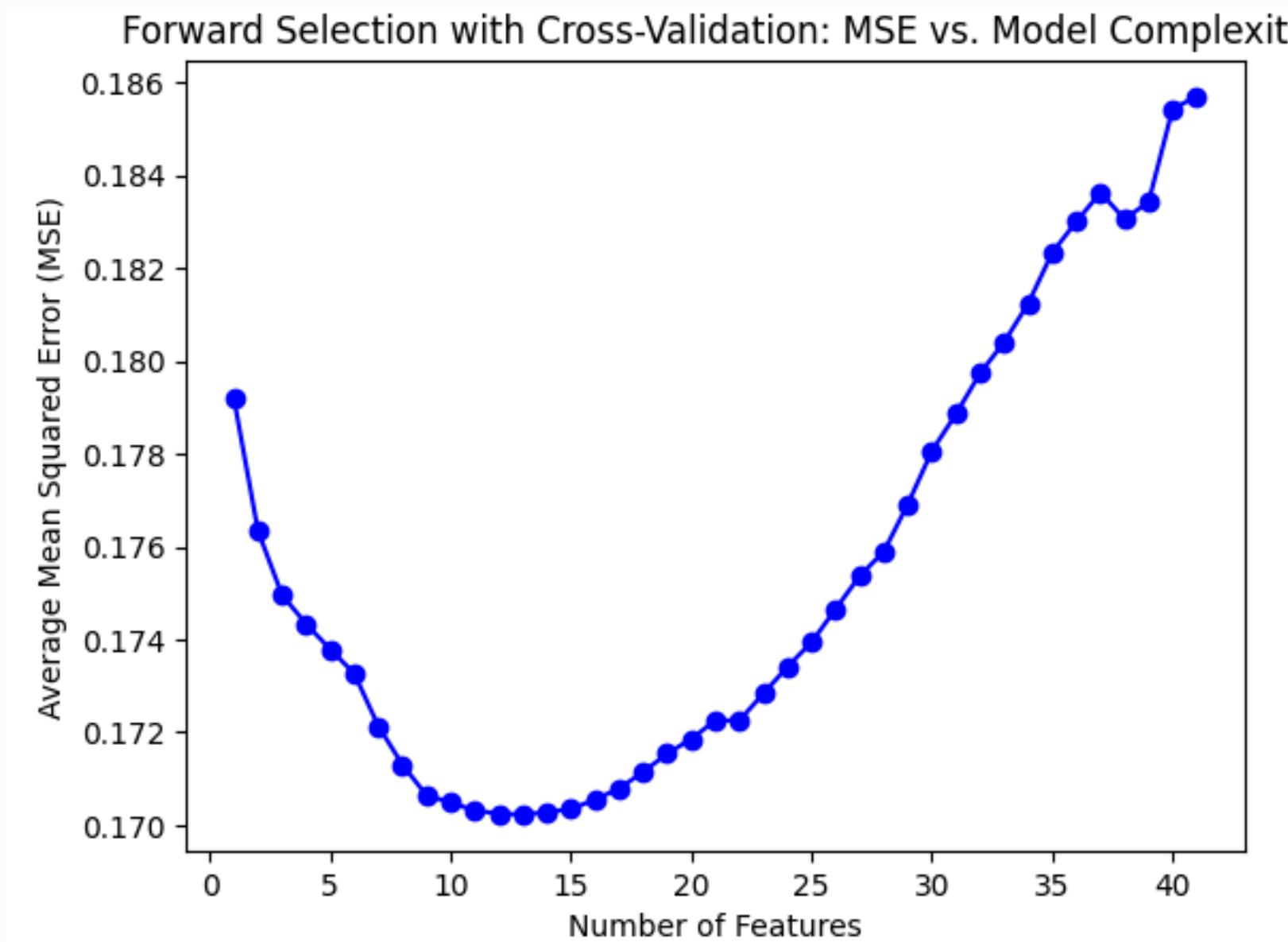
5 Fold



10 Fold

Logistic - 特徵挑選 - 刪除 - 0.5~+0.5 版本

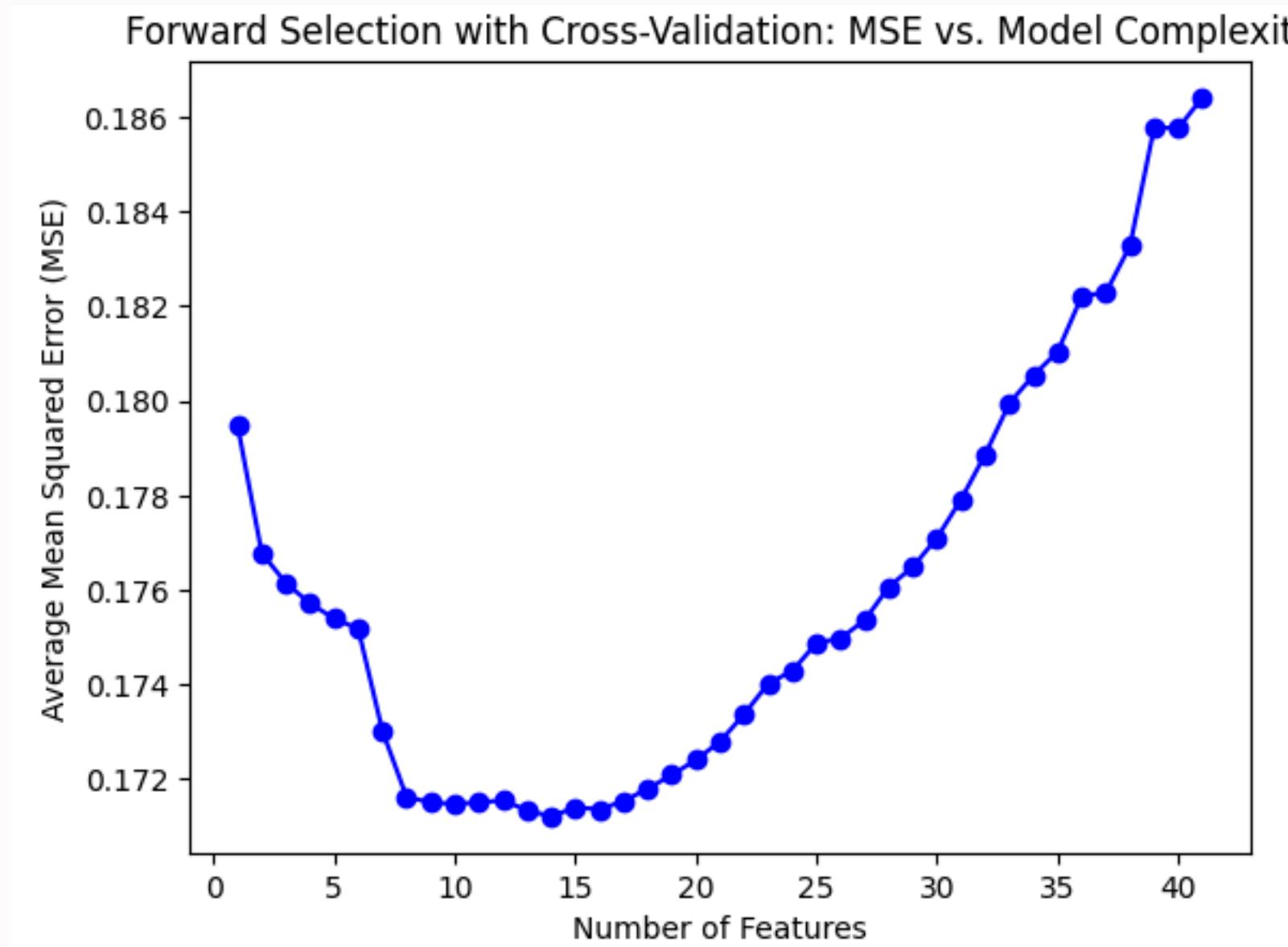
以5 Fold為例：



變數排序	
1. SOX_Return	7. SOX_Low/Close
2. IXIC_Low/Open	8. GSPC_High/Open
3. SOX_High/Close	9. SMB
4. Oil_change	10. DJI_volume_change
5. 台灣VIX	11. Jeiba_當天盤後
6. SOX_High/Low	

Logistic - 特徵挑選 - 刪除 - 0.5~+0.5 版本

以10 Fold為例：



變數排序	
1. SOX_Return	7. SOX_Low/Close
2. IXIC_Low/Open	8. DJI_Low/Close
3. SOX_High/Close	8. GSPC_High/Open
4. Oil_change	9. Jeiba_當天盤後
5. RMW	
6. SOX_High/Low	

Logistic-模型實測 刪除-0.5~+0.5版本

在經過forward selection的挑選之後，我們使用在5Fold和10Fold之中前13組有效變數當作我們要放入logistic regression的特徵，並計算相關的模型衡量指標：

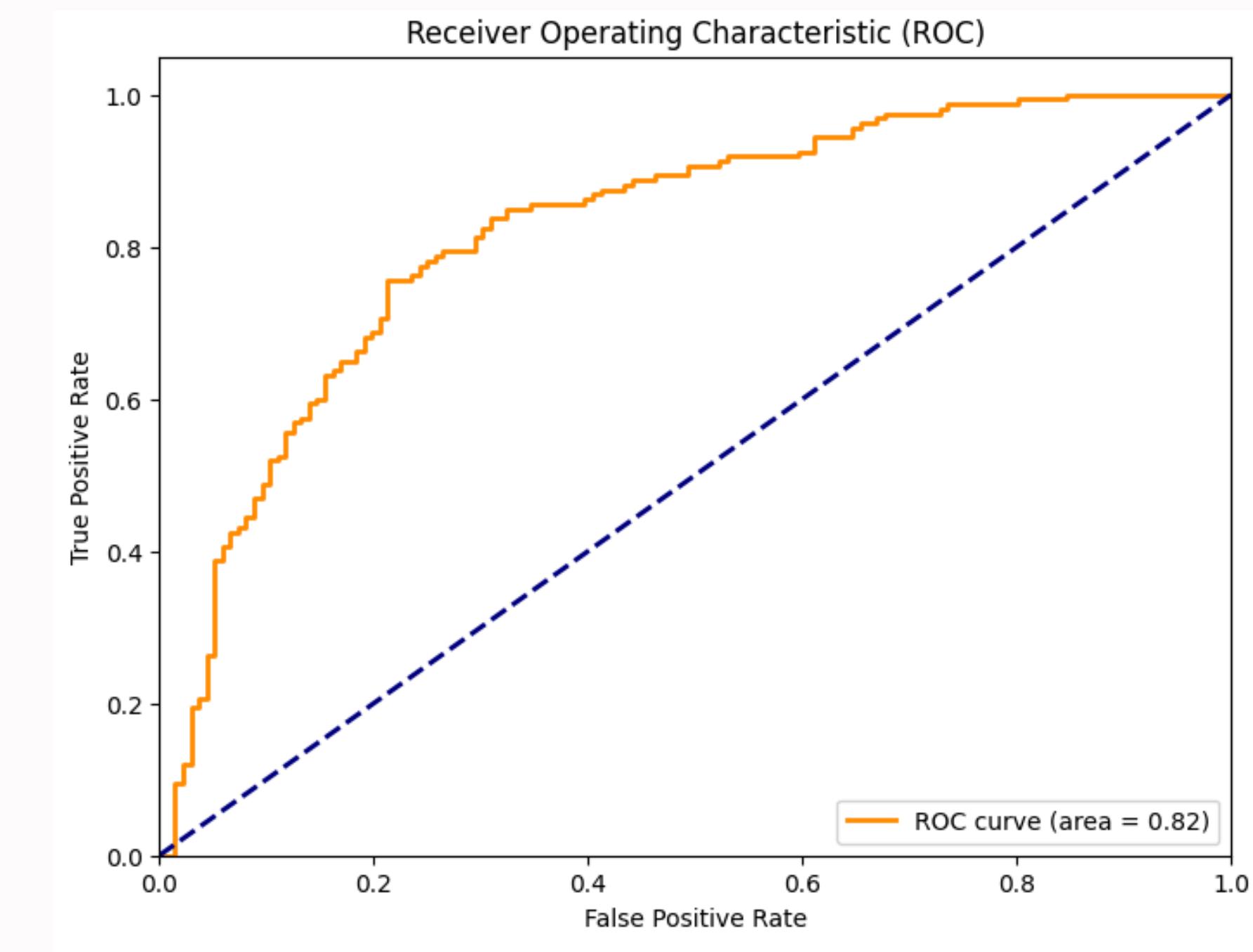
使用特徵	
	7. SOX_High/Low
1. SOX_Return	8. SOX_Low/Close
2. IXIC_Low/Open	9. DJI_Low/Close
3. SOX_High/Close	10. GSPC_High/Open
4. Oil_change	11. SMB
5. RMW	12. DJI_volume_change
6. 台灣VIX	13. Jeiba_當天盤後

Logistic - 模型實測- 刪除-0.5~+0.5版本

在使用我們要的特徵變數後，以下是模型實測的結果：

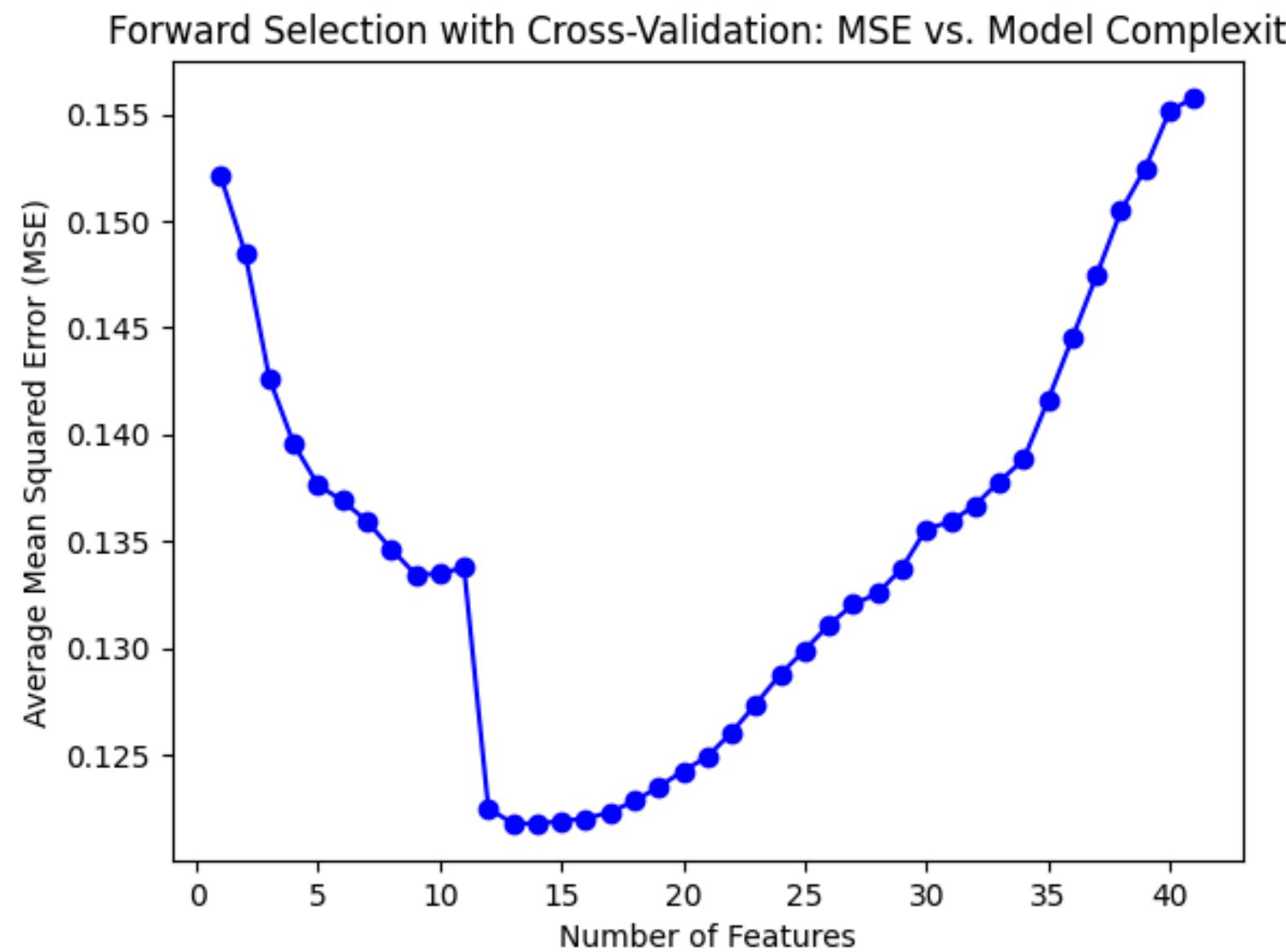
整體準確率：76.014%

	precision	recall
0	0.79	0.65
1	0.74	0.86
	f1-score	
0	0.71	
1	0.79	



Logistic - 特徵挑選 - 刪除- 1 ~ +1 版本

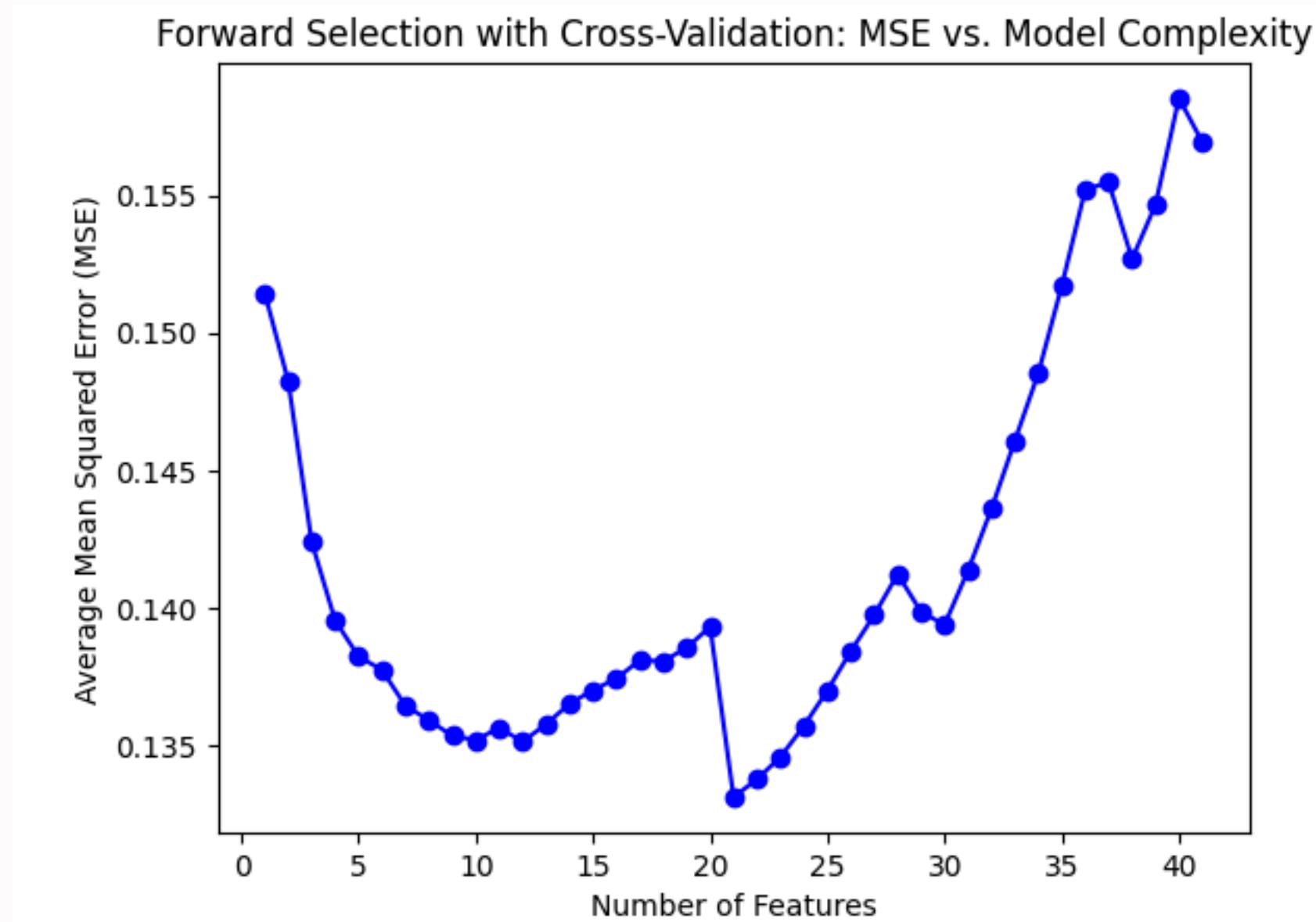
以5 Fold為例：



變數排序	
1. SOX_Return	7. IXIC_High/Low
2. GSPC_Low/Open	8. RMW
3. 台灣VIX	9. 台灣VIX_change
4. SOX_High/Close	10. GSPC_Low/Close
5. DJI_volume_change	11. SOX_High/Low
6. IXIC_volume_change	12. SOX_Low/Close

Logistic -特徵挑選- 刪除- 1 ~ +1 版本

以10 Fold為例：



變數排序	
1. SOX_Return	7. RMW
2. GSPC_Low/Open	8. IXIC_volume_change
3. 台灣VIX	9. GSPC_Low/Close
4. SOX_High/Close	10. 台灣VIX_change
5. DJI_volume_change	11. Jeiba_當天盤中
6. IXIC_High/Low	

Logistic-模型實測 刪除-1~+1 版本

在經過forward selection的挑選之後，我們使用在5Fold和10Fold之中前13組有效變數當作我們要放入logistic regression的特徵，並計算相關的模型衡量指標：

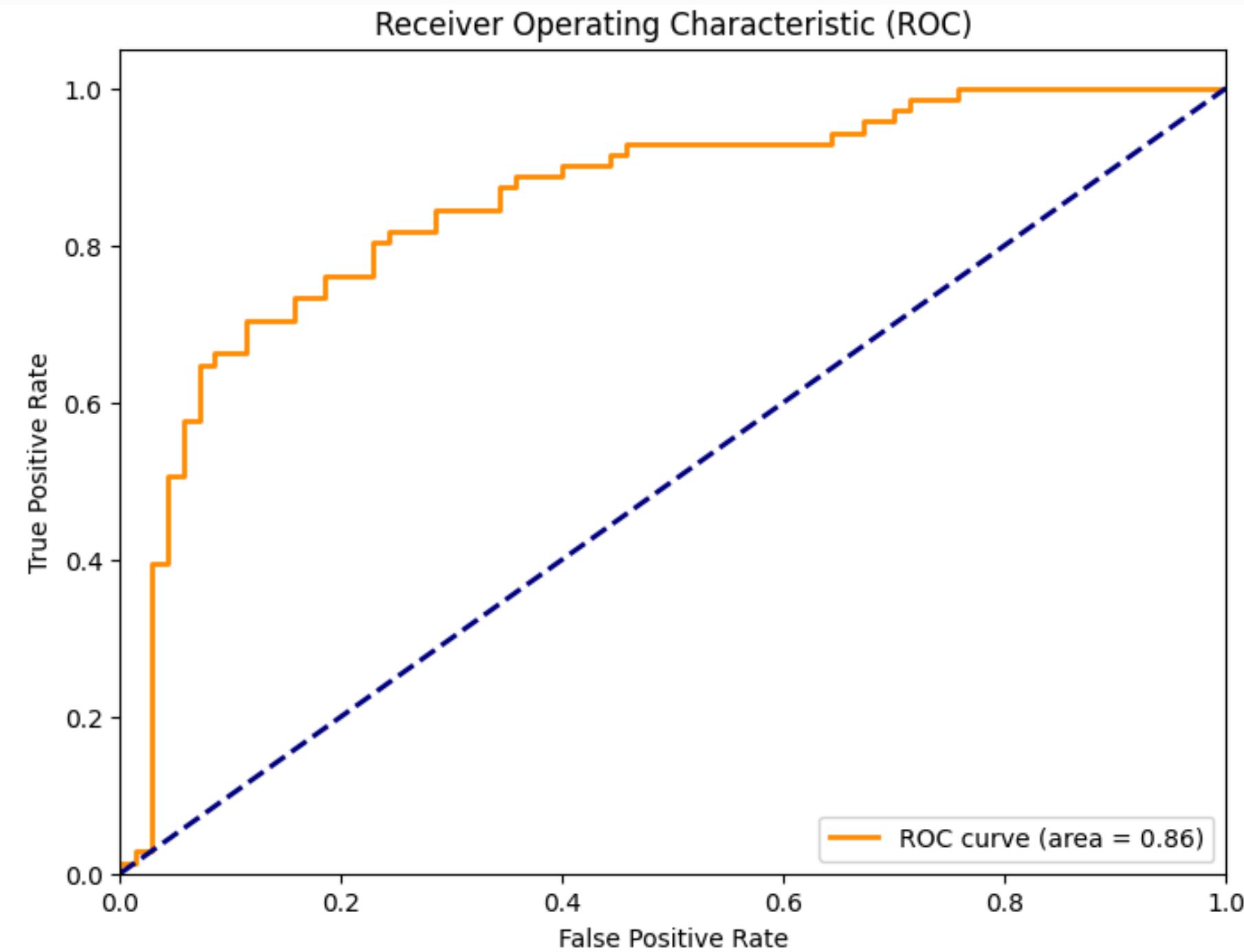
使用特徵	
	7. SOX_High/Low
1. SOX_Return	8. SOX_Low/Close
2. GSPC_Low/Open	9. GSPC_Low/Close
3. SOX_High/Close	10. 台灣VIX_change
4. IXIC_High/Low	11. IXIC_volume_change
5. RMW	12. DJI_volume_change
6. 台灣VIX	13. Jeiba_當天盤中

Logistic - 模型實測- 刪除- 1~+1 版本

在使用我們要的特徵變數後，以下是模型實測的結果：

整體準確率：76.596%

	precision	recall
0	0.84	0.66
1	0.72	0.87
	f1-score	
0	0.74	
1	0.79	



[附錄：甚麼是ROC曲線？](#)

[附錄：甚麼是Precision、recall、f1-score？](#)

Logistic - 模型實測- 刪除-2~+2版本

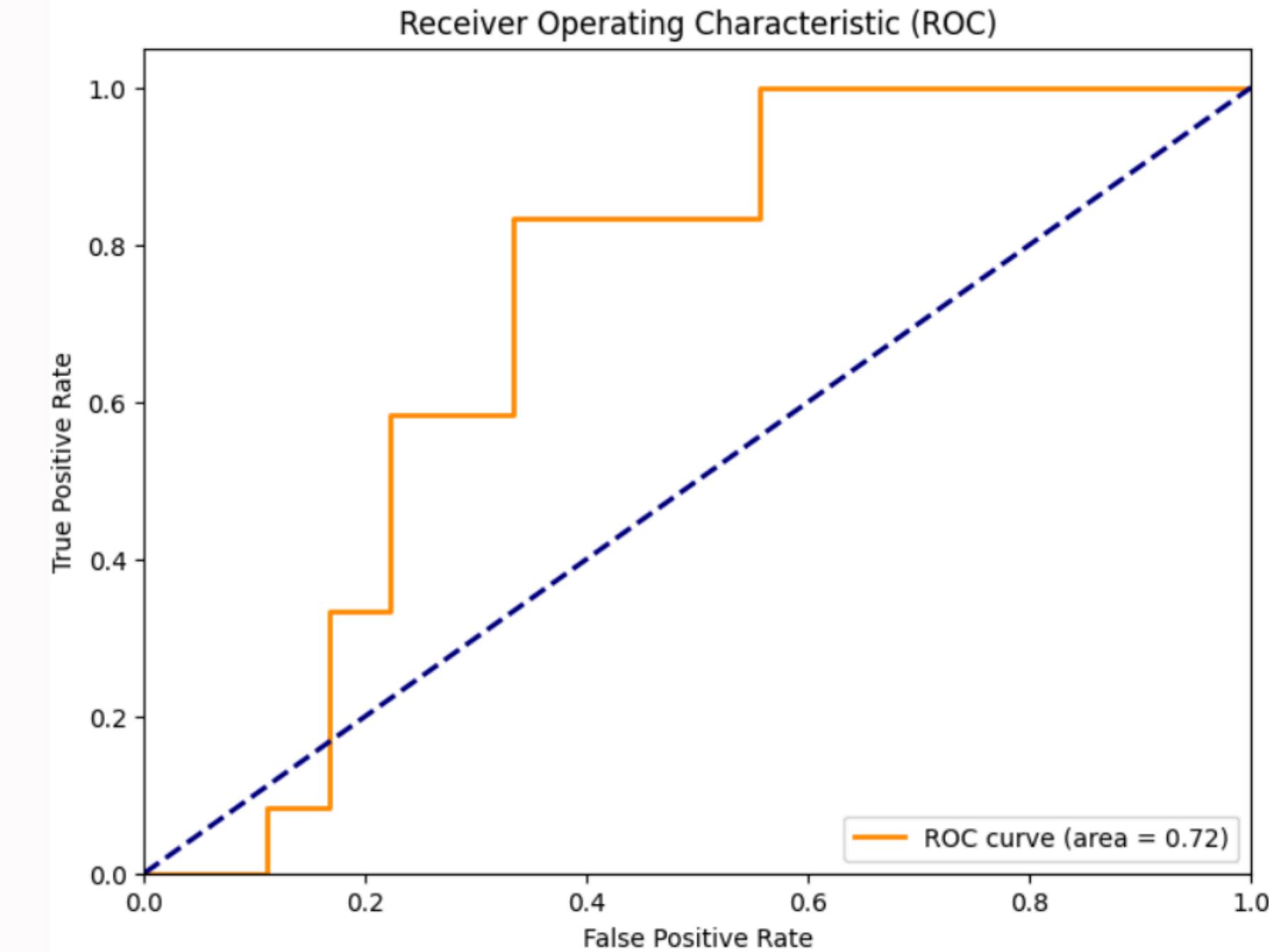
由於在刪除retun介於-2~+2之間的資料，原本的資料量只剩下75天，在使用cross validation以及forward selection時會出現over fitting的問題，因此對於這個版本我們將會直 4 一

Logistic - 模型實測- 刪除-2~+2版本

在使用我們要的特徵變數後，以下是模型實測的結果：

整體準確率：70%

	precision	recall
0	0.8	0.67
1	0.6	0.75
	f1-score	
0	0.73	
1	0.67	

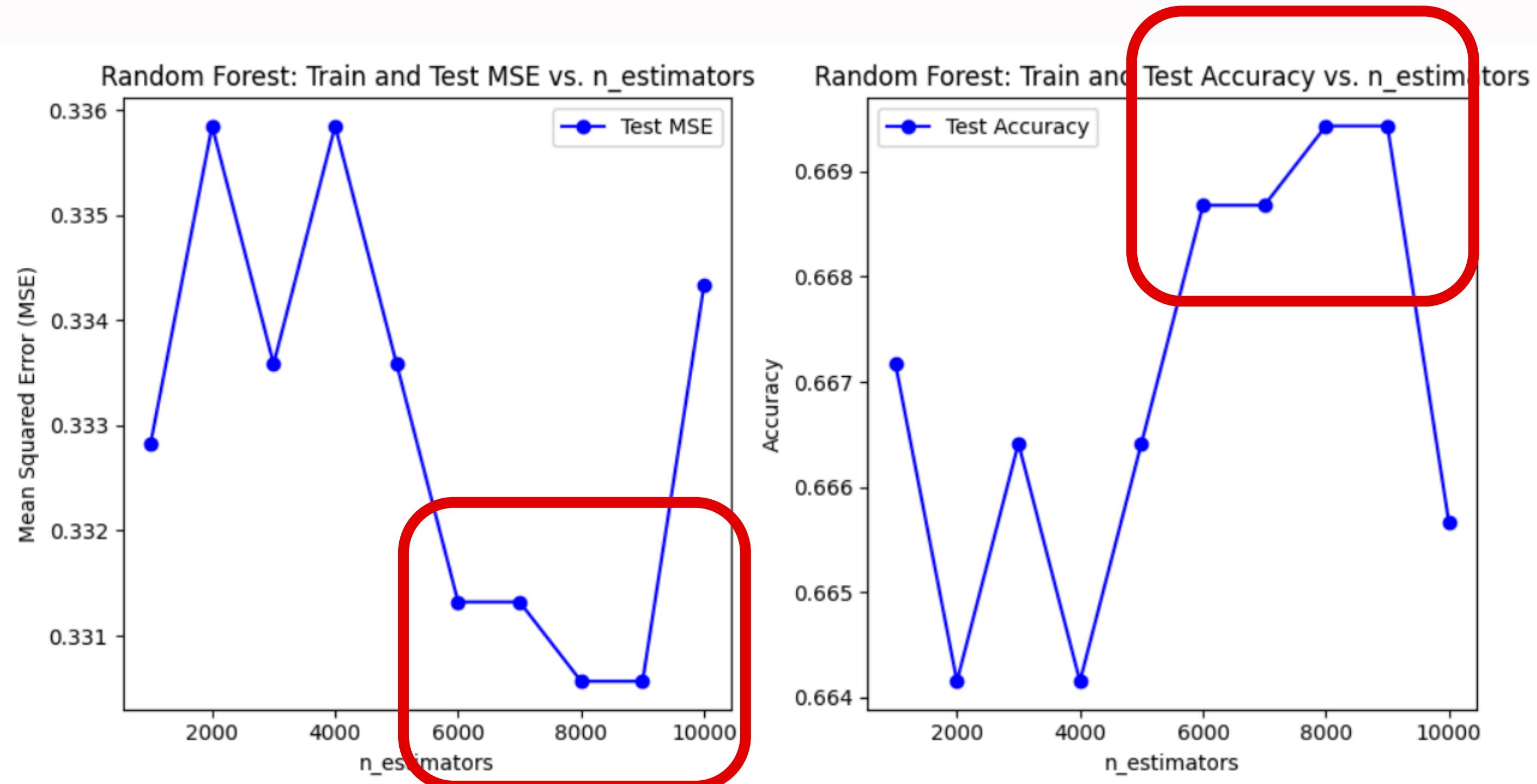


(3)

Random Forest

Random Forest-分支數量挑選

由於沒有資訊貢獻量的變數對Random Forest影響不大，因此我們僅使用cross validation去檢驗對於模型幫助的最多分支數量：

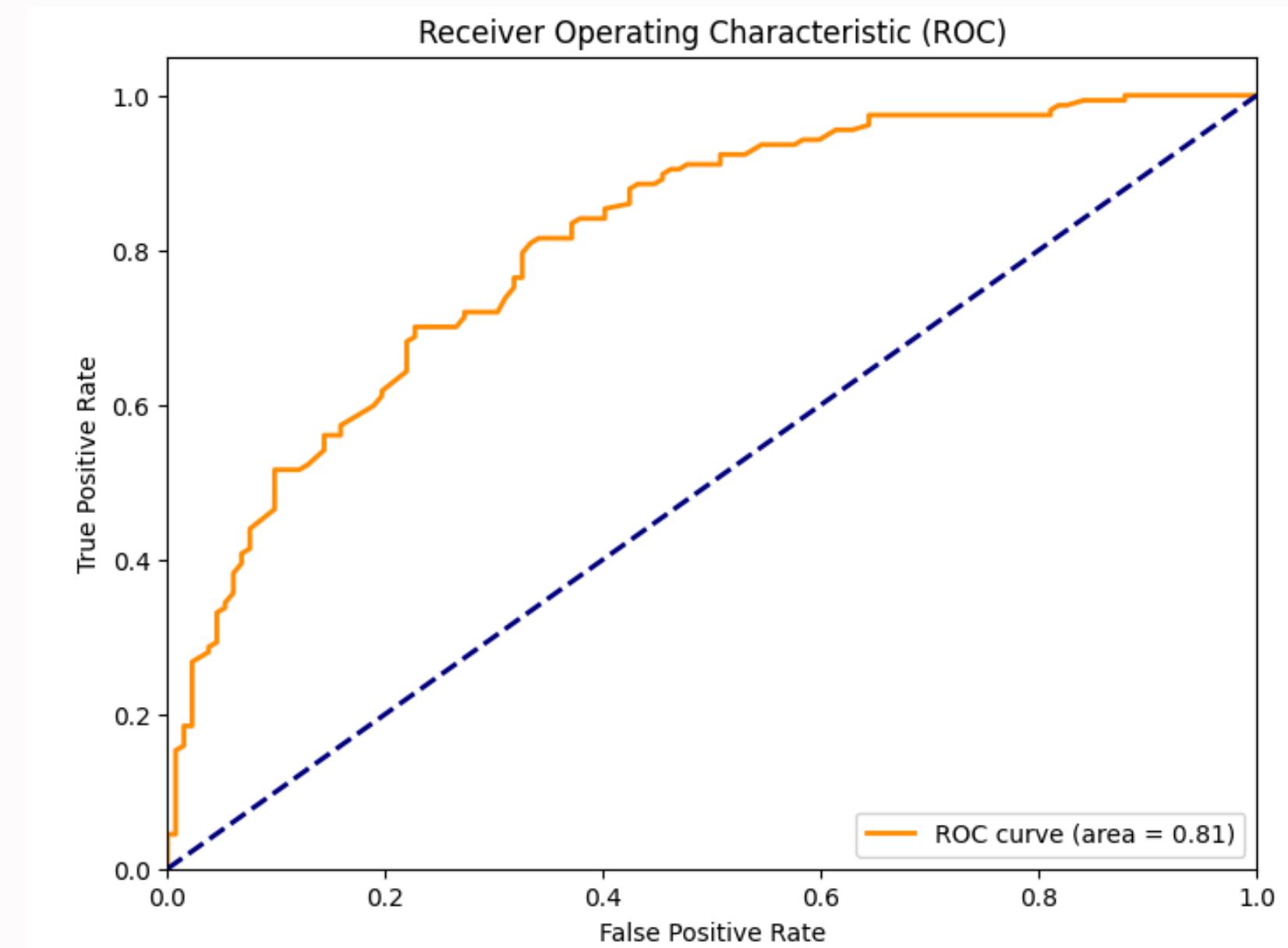


隨機森林-模型實測 刪除-0.5~+0.5版本

我們使用做分支樹數量為 240 的模型做實測，結合前面 forward selection 得到的有效變數，得到以下結果：

整體準確率：74.048%

	precision	recall
0	0.73	0.63
1	0.73	0.83
	f1-score	
0	0.69	
1	0.78	

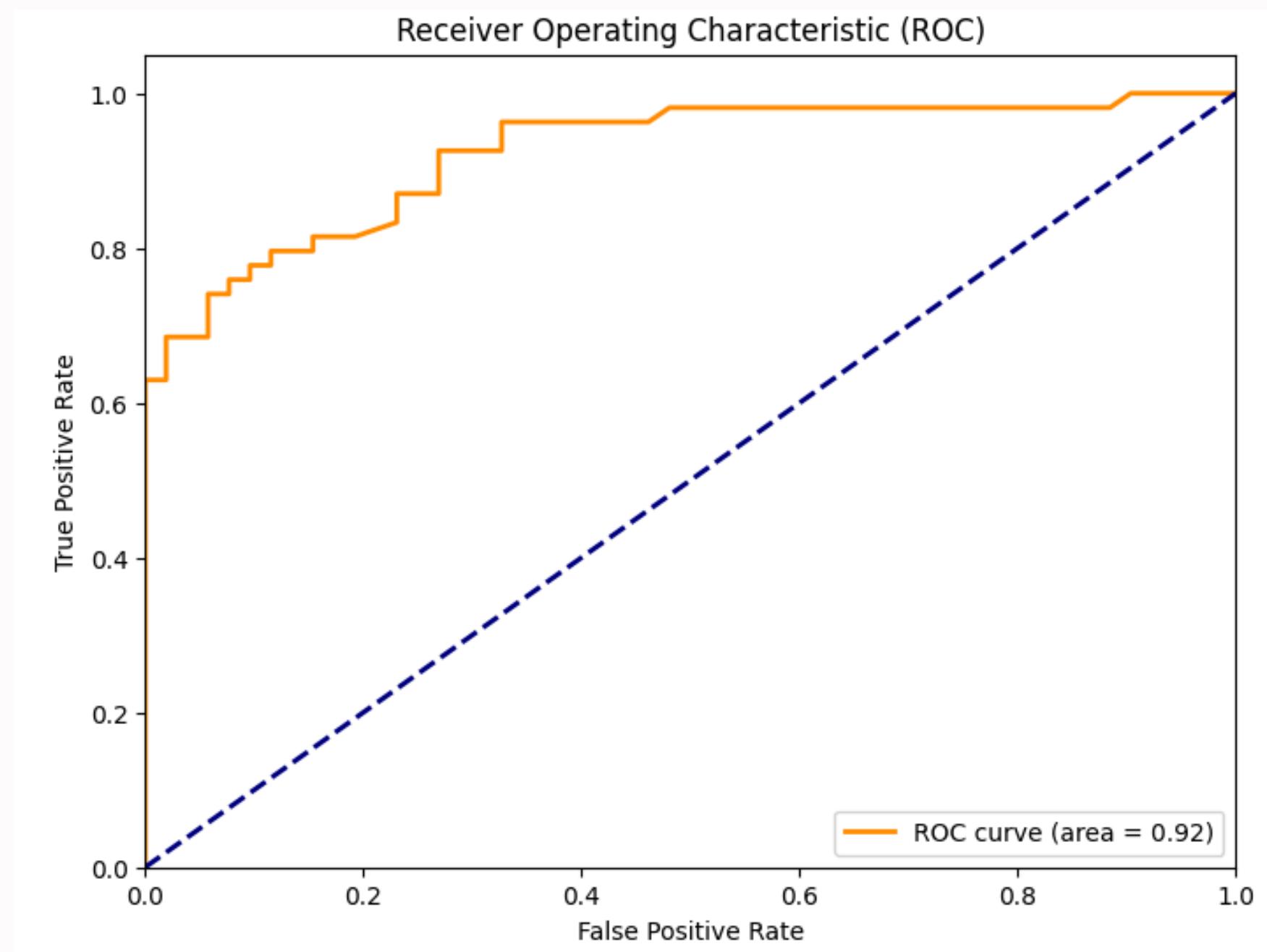


隨機森林-模型實測 刪除-1~+1 版本

我們使用做分支樹數量為 1250 的模型做實測，結合前面 forward selection 得到的 13 個有效變數，得到以下結果：

整體準確率：80.189%

	precision	recall
0	0.82	0.77
1	0.79	0.83
	f1-score	
0	0.79	
1	0.81	

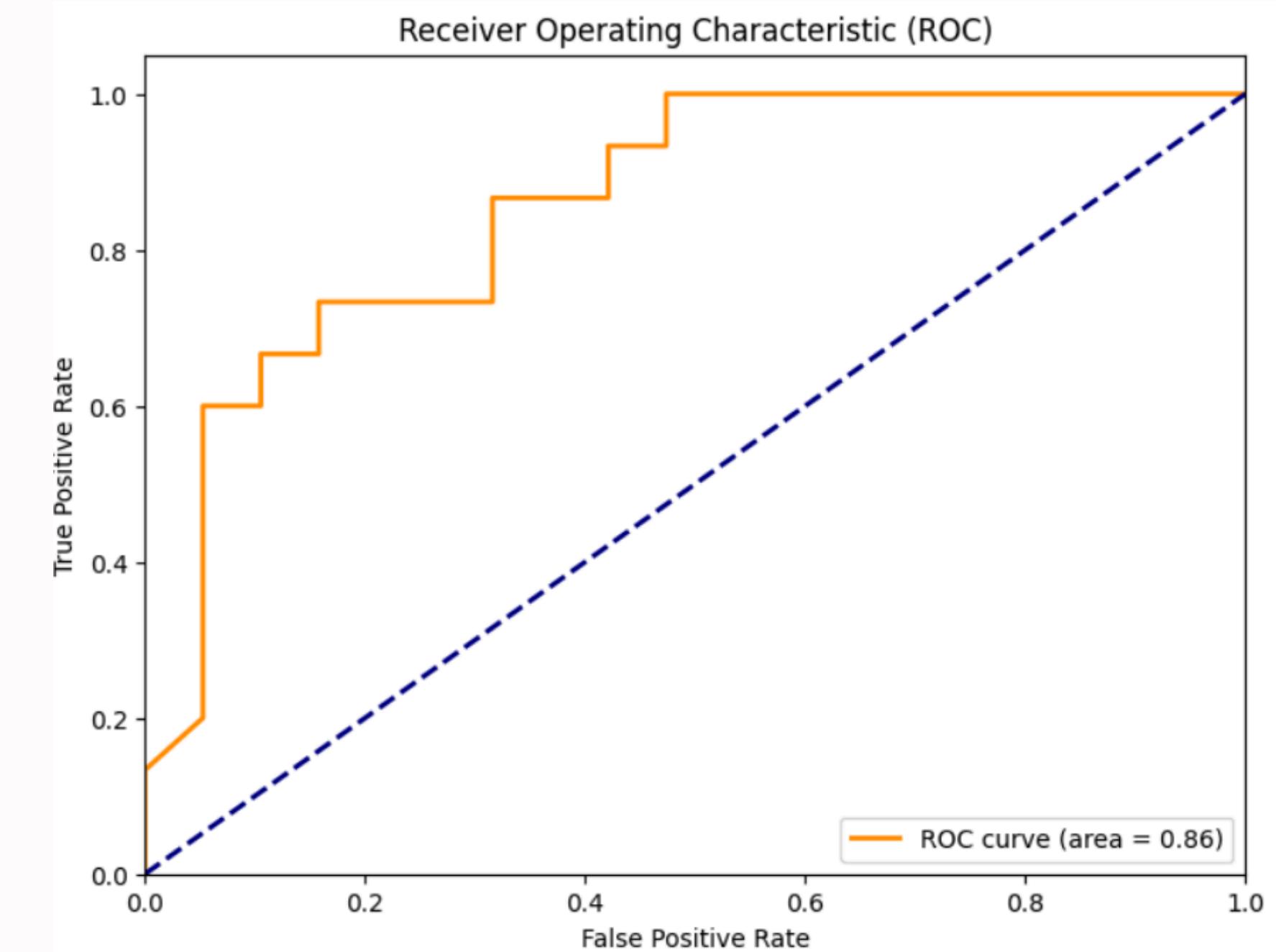


隨機森林-模型實測 刪除-2~+2 版本

我們使用做分支樹數量為150的模型做實測，得到以下結果：

整體準確率：76.471%

	precision	recall
0	0.87	0.68
1	0.68	0.87
f1-score		
0	0.76	
1	0.76	



(4)

Support Vector Machine

Support Vector Machine-模型公式

我們將資料套用到SVM模型，同時選擇可以處理非線性資料的 RBF 核函數來幫助處理資料，公式如下：

$$SVM(X) = sign \left(\sum_{n \in SV}^{\square} \alpha_n y_n K(x, x_n) + b \right)$$

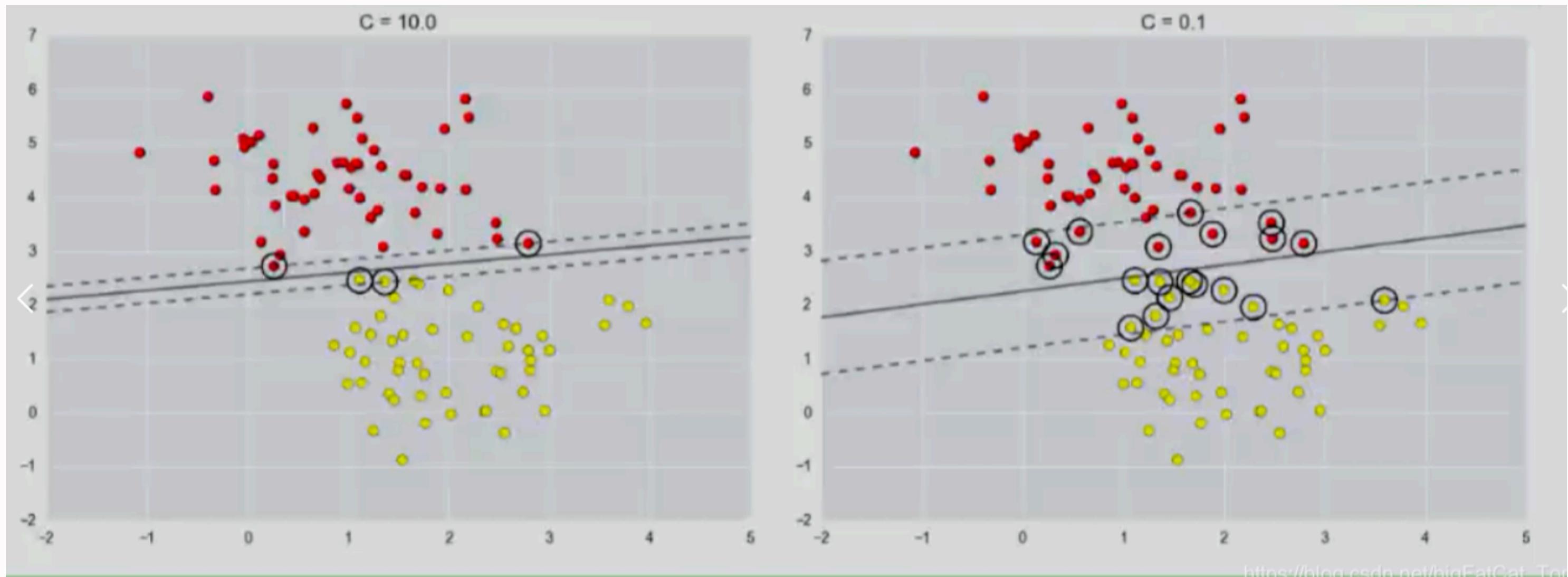
其中 $K(x, x_n) = e^{-\gamma \|x - x_n\|^2}$ 為RBF 核函數

$$x = [MKT_t, SMB_t, \dots, Jeiba_M_change_t, TVIX_change_t]$$

Support Vector Machine-模型公式

其中在SVM RBF模型之下有兩個可以調整的參數，分別為C懲罰項以及gamma：

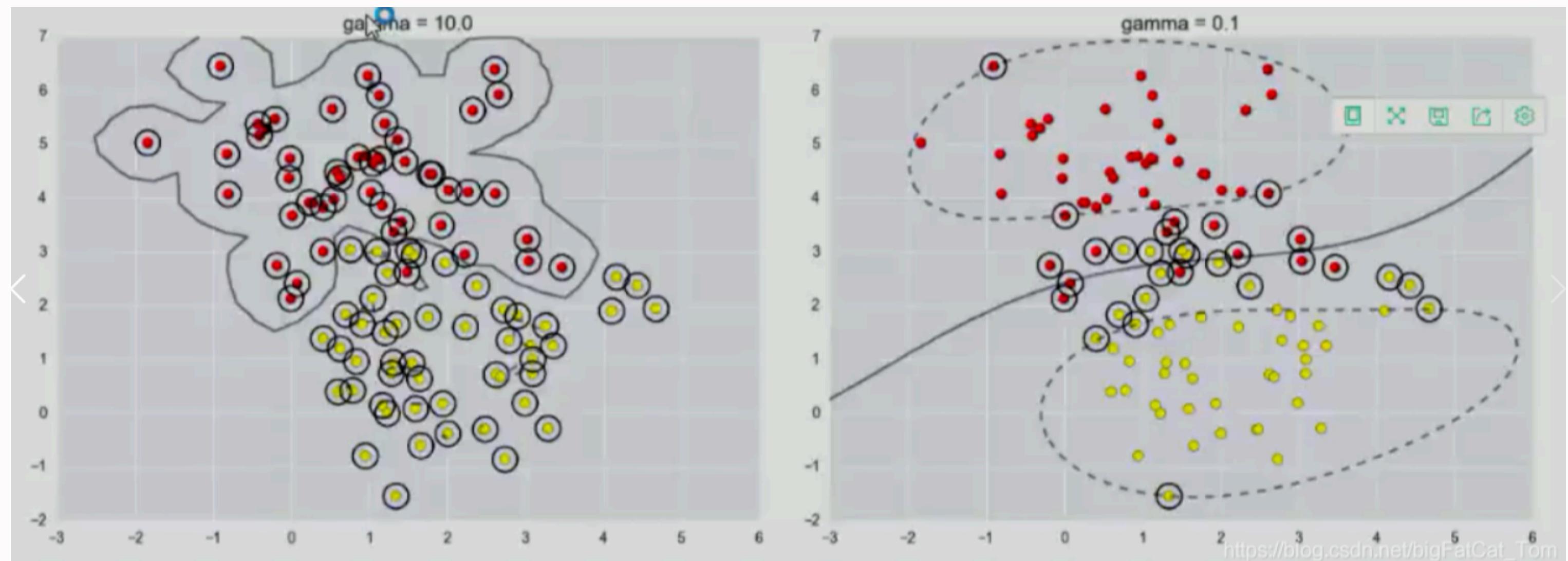
以懲罰項c的不同舉例：



Support Vector Machine-模型公式

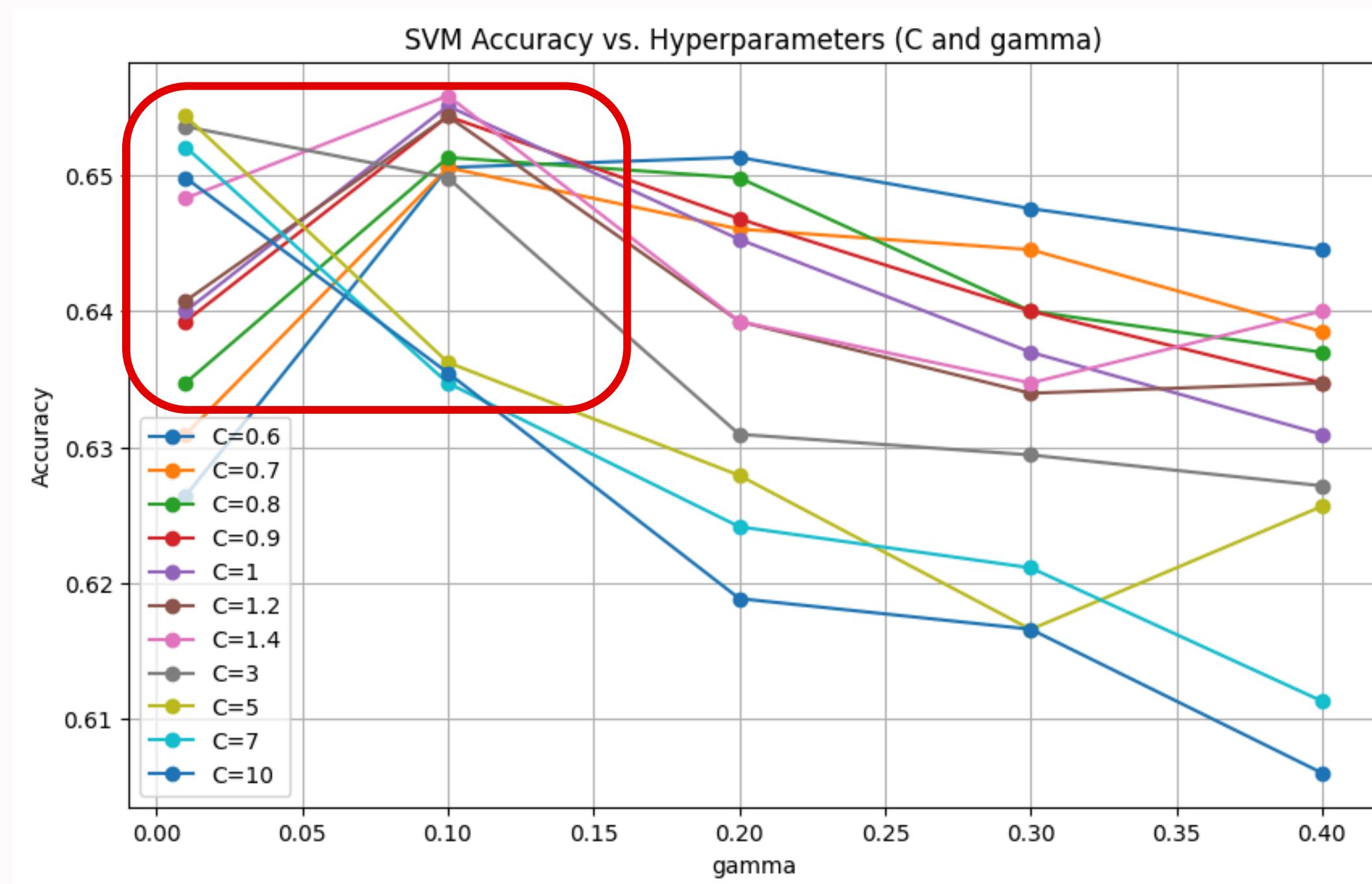
其中在SVM RBF模型之下有兩個可以調整的參數，分別為C懲罰項以及gamma：

以gamma的不同舉例：

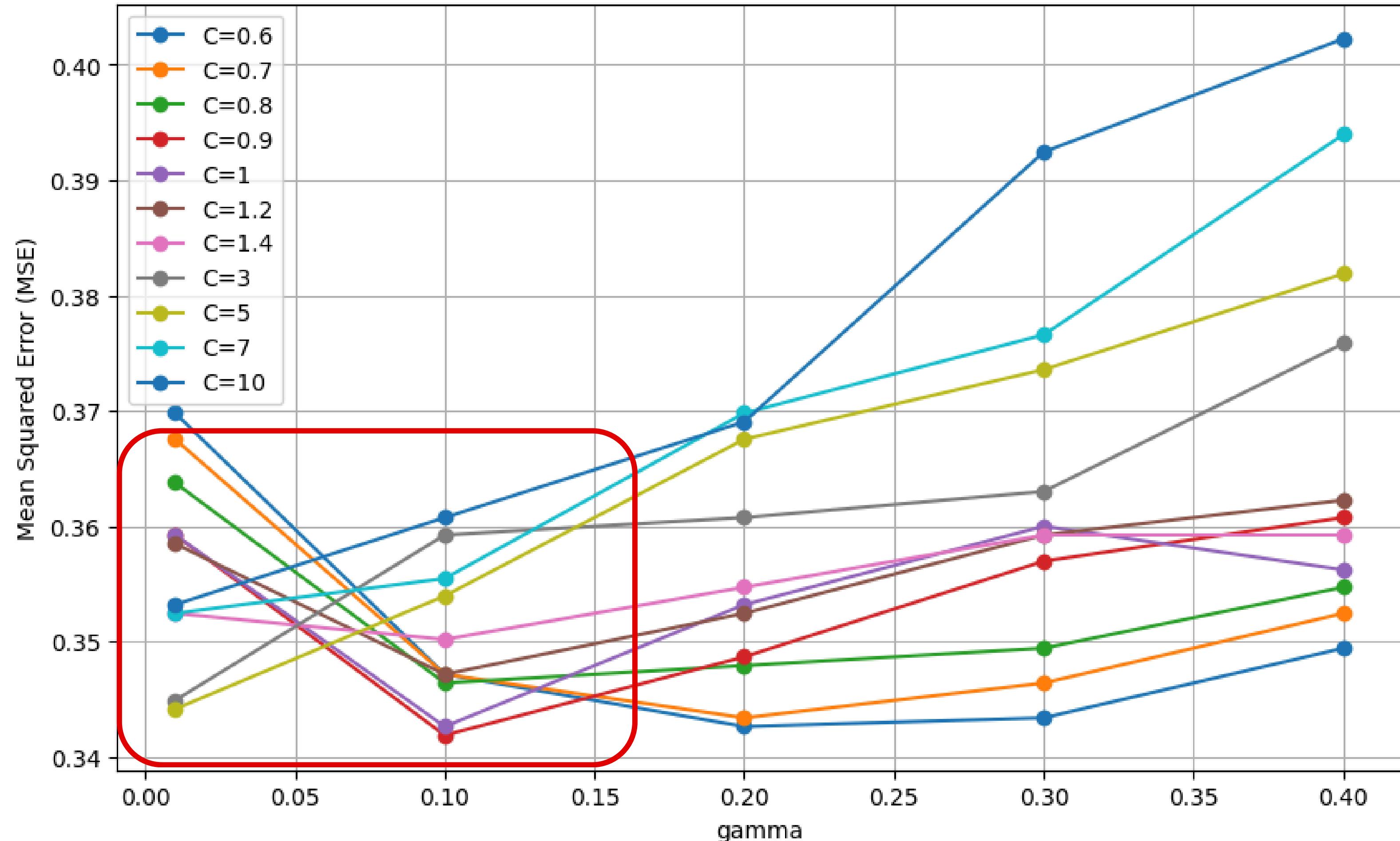


Support Vector Machine-參數挑選

為了找出表現交加的參數組合，我們結合corss validation並將懲罰項c以及參數gamma作為變動的變數，來觀測在甚麼範圍內的c和gamma可以使得SVM模型有較佳的表現



SVM MSE vs. Hyperparameters (C and gamma)

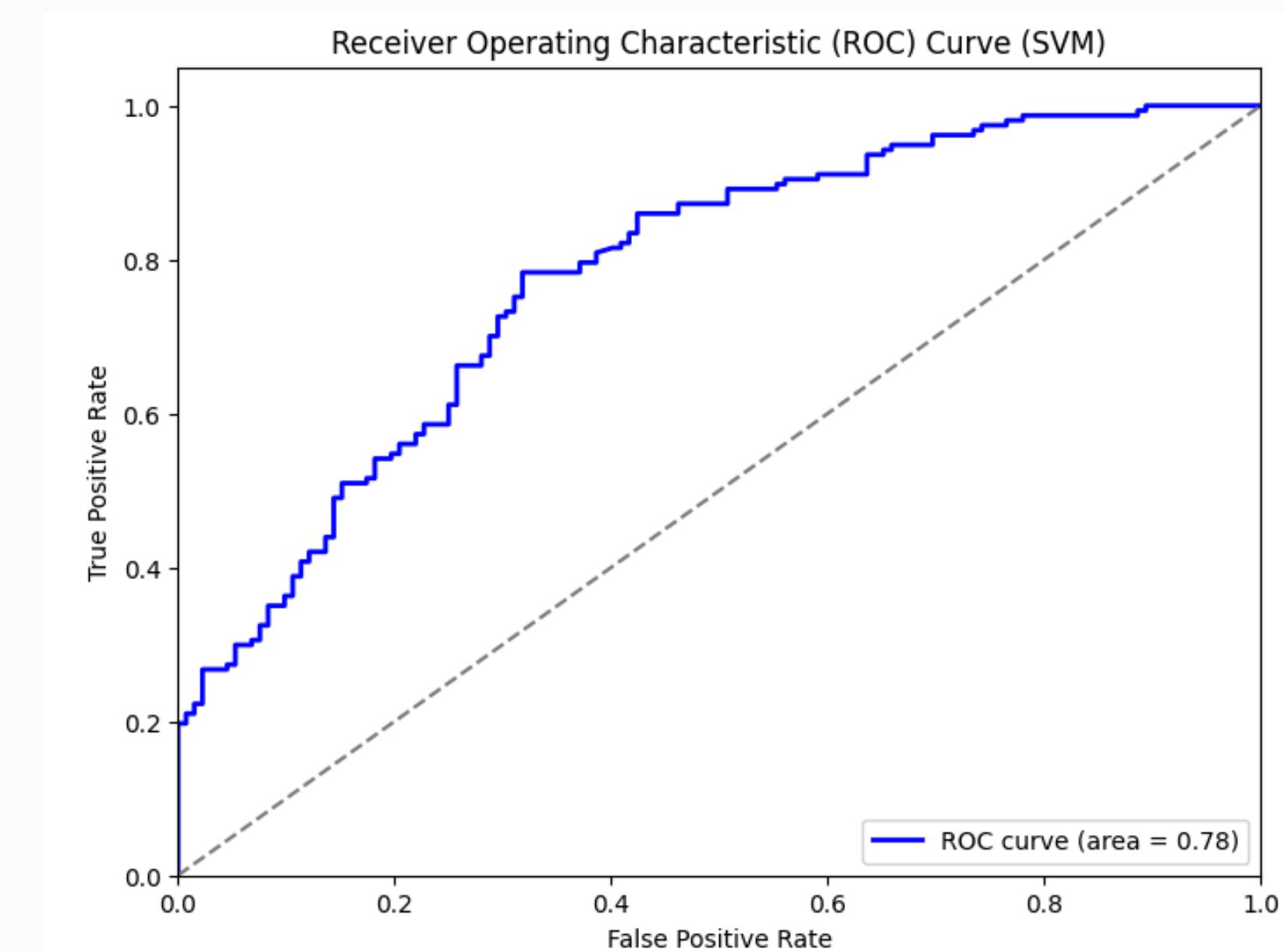


SVM-模型實測-刪除-0.5~+0.5版本

最後經過篩選之後，我們認為gamma為0.2而c=7.5的時候模型不論在準確度還是MSE的表現上都會較好，因此我們設置gamma=0.2、c=7.5作為我們的模型參數，以下為結果：

整體準確率：71.626%

	precision	recall
0	0.74	0.59
1	0.70	0.82
f1-score		
0	0.66	
1	0.76	

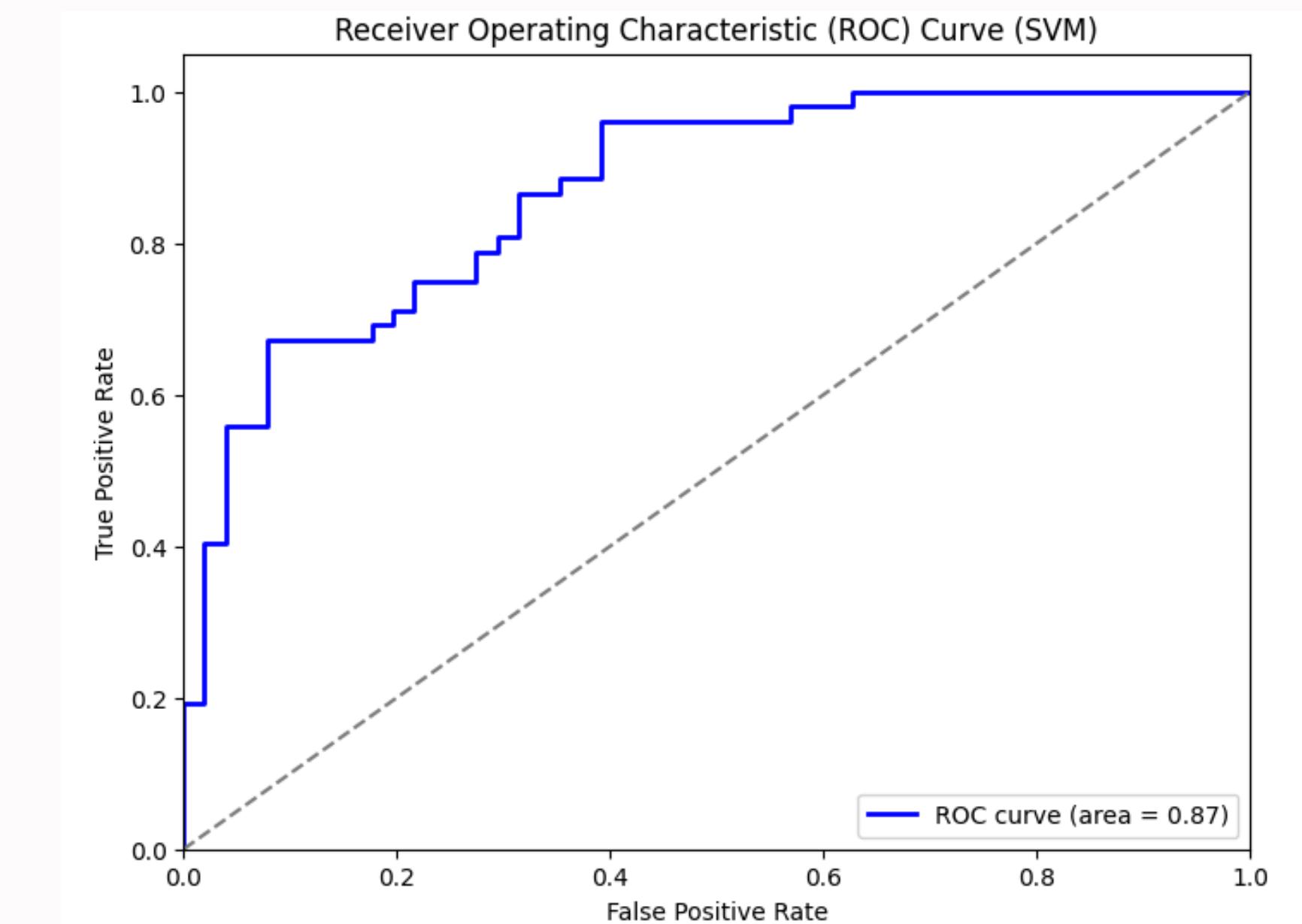


SVM-模型實測-刪除-1~+1 版本

最後經過篩選之後，我們認為gamma為0.02而c=14.5的時候模型不論在準確度還是MSE的表現上都會較好，因此我們設置gamma=0.02、c=14.5作為我們的模型參數，以下為結果：

整體準確率：77.67%

	precision	recall
0	0.94	0.59
1	0.70	0.96
f1-score		
0	0.72	
1	0.81	

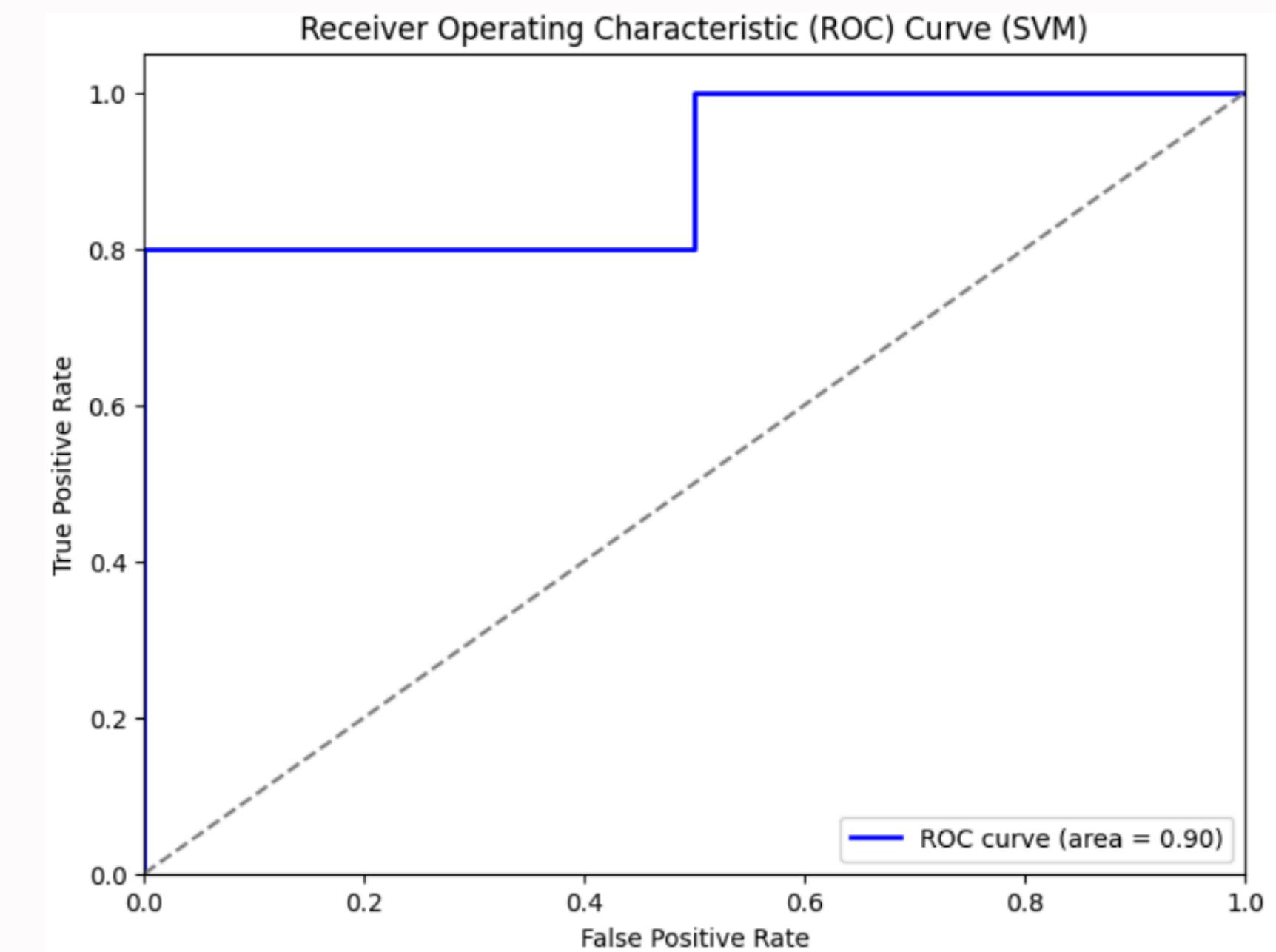


SVM-模型實測-刪除-2~+2版本

最後經過篩選之後，我們認為gamma為0.07而c=8的時候模型不論在準確度還是MSE的表現上都會較好，因此我們設置gamma=0.02、c=14.5作為我們的模型參數，以下為結果：

整體準確率：77.778%

	precision	recall
0	0.75	0.75
1	0.8	0.8
f1-score		
0	0.75	
1	0.8	



06

總結

專題研究總結

1. 運用 PTT 爬蟲，Jieba 切詞法以及詞嵌入向量等技術後，再透過 PCA 以及反距離加權等方法，我們得到字詞的情緒分數，並以之最後得到每日盤中以及盤後的情緒分數。
2. 運用該情緒分數，在單因子模型、Fama French 三因子以及五因子模型中，與0050報酬並無顯著關係，且不論牛熊市及不同漲跌情境皆然；然而，若使用時間序列的向量自我回歸模型(VAR)於解釋加權指數報酬，則可發現情緒指數之間具有強自我相關，而報酬自身則無。另外，盤中分數能被前一日報酬解釋而報酬則能被前一日盤後情緒解釋。
3. 該情緒分數在Logistic預測模型上經由forward selection驗證後確認情緒分數是能提供預測力的，不過隨著預測目標的篩選越來越嚴格，資料量越來越少，也因而導致預測模型的能力也逐漸下降，經由測試模型在分類隔日return大於正負1%時的效能會最好。

07

參考資料

參考資料(論文)

1. 因子模型於台灣股票之實證檢驗。(劉晉翰，2016)
2. 財金新聞情緒分類實證研究-以鉅亨網財金新聞為例。(吳錦文、王昭文、黃振聰、田高銘，2020)
3. 三大法人未平倉量、前十大交易人未平倉量、三大法人買賣超市值對台指期之影響。(顏鈺庭，2014)
4. 五因子模型 A 股實證研究。(王柯，2017)
5. 中文情感分析應用於 PTT 之研究。(劉昱函，2017)
6. A text mining based study of investor sentiment and its influence on stock returns. (Yong Shi, Ye-ran Tang, Ling-xiao Cui, Wen Long, 2018)
7. An introduction to statistical learning. (Gareth James, Daniela Witten, Trevor Hastie, 2013)
8. Investor Sentiment and Return Predictability in agriculture future markets. (Changyun Wang, 2001)
9. UTCNN: a Deep Learning Model of Stance Classification on Social Media Text. (Wei-Fan Chen, Lun-Wei Ku, 2016)

參考資料(網路)

1. 【自然語言處理 — 概念篇】來認識一下詞向量(Word Embedding or Word Vector)
2. Inverse weighted method: Wiki
3. Fama-French 三因子模型 (Python 實現)
4. SVM Classifier and RBF Kernel — How to Make Better Models in Python
5. 線性迴歸的變數選取方法 - IBM
6. 模型選取- SAS Taiwan

特別感謝老師：
星華老師！



附錄

高斯分佈的機率密度函數：

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (\text{X 為隨機變量, } \sigma \text{ 為變異數, } \mu \text{ 為平均})$$

因此原本高維的數據可以這樣表示：

$$p_{ij} = \frac{\exp\frac{(-\|x_i - x_j\|^2)}{2\sigma^2}}{\sum \frac{\exp(-\|x_k - x_l\|^2)}{2\sigma^2}}$$

低維的數據用 t 分布的機率密度函數可以這樣表示(自由度為 1)：

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum (1 + \|y_k - y_l\|^2)^{-1}}$$

x 為高維當中的數據， y 為低維當中的數據， P, Q 分別代表機率分佈。

為什麼會使用 t 分佈來近似低維的數據呢？主要是因為轉換成低維之後一定會丟失許多訊息，所以為了不被異常值影響可以使用 t 分佈。t 分佈在樣本數較少時，可以比較好模擬母體分布的情形，不容易被異常值所影響。

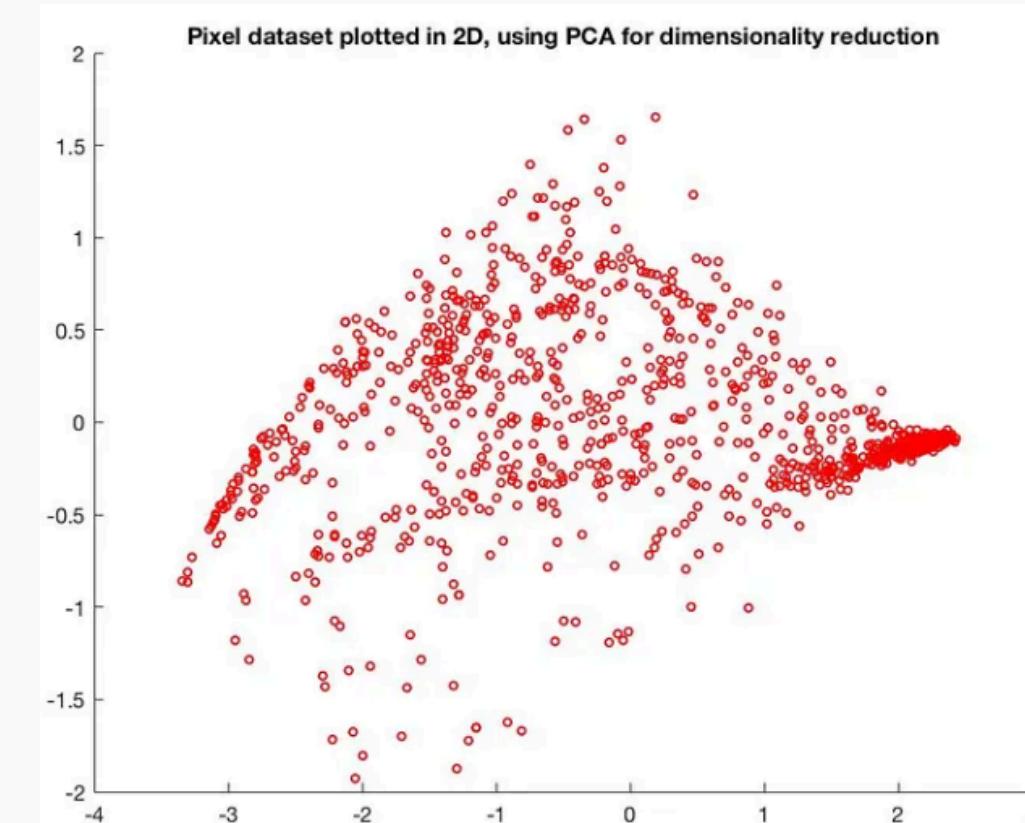
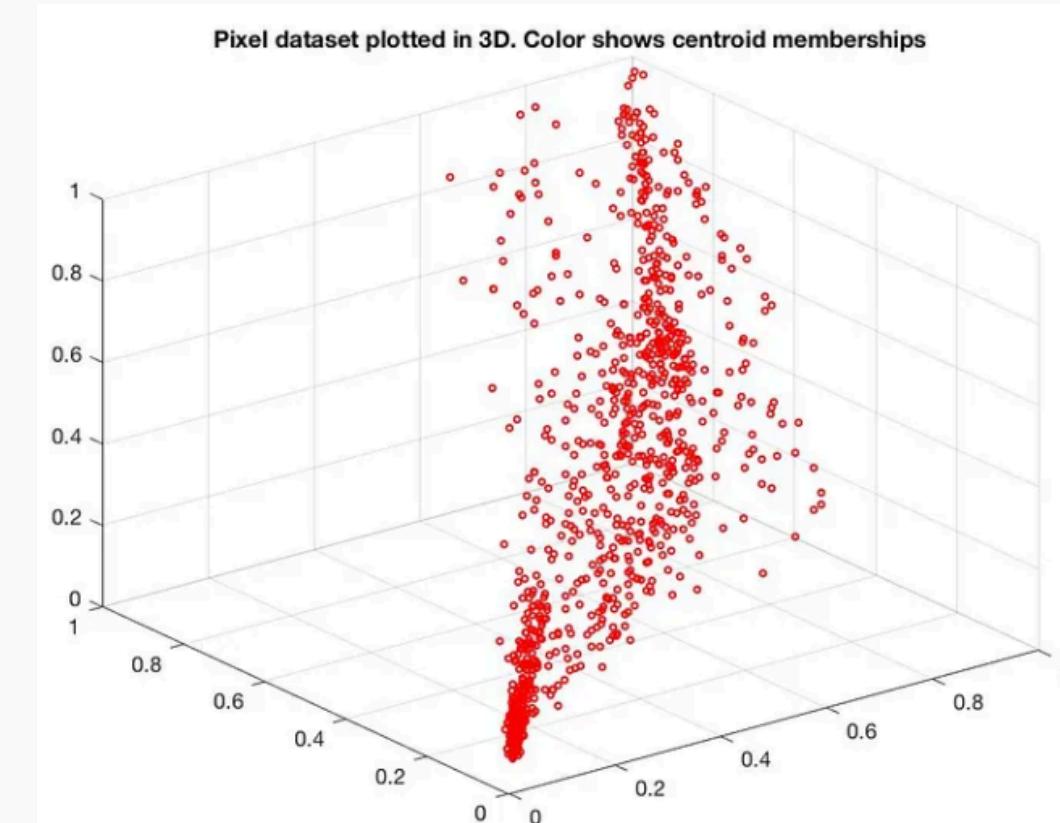
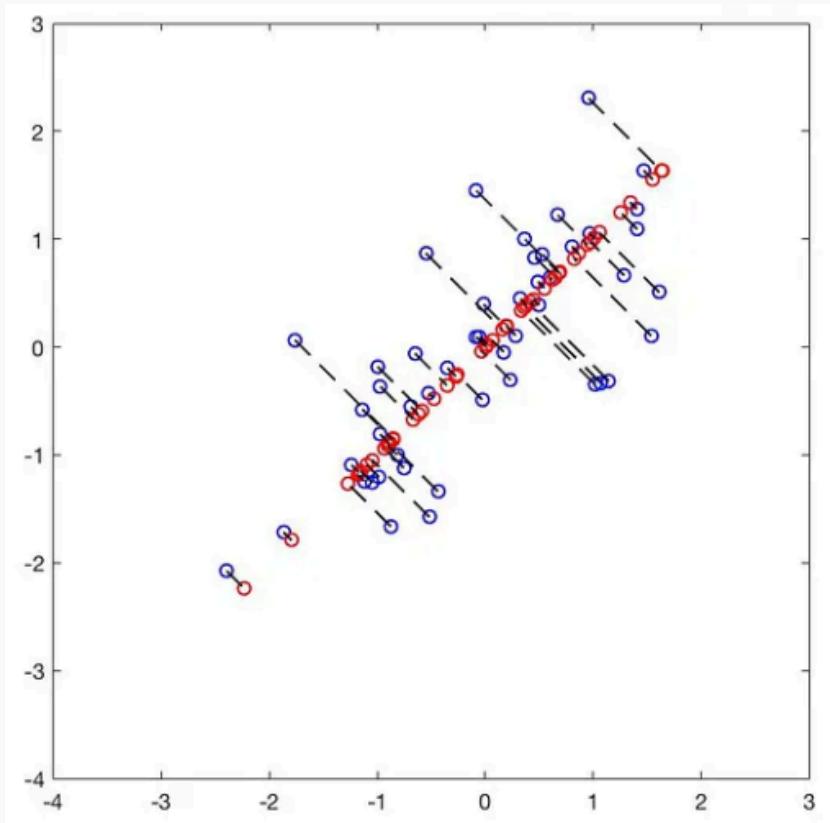
奇異值分解: $A = U\Sigma V^T$

A 為一個 $m \times n$ 的矩陣， U 跟 V 都為正交矩陣， Σ 為奇異值矩陣。奇異值矩陣為矩陣 A 對應的特徵值，在 PCA 當中又叫做主成份，代表對保存訊息的重要程度，通常由大到小遞減排列在對角中

A 我們通常使用共變異數矩陣 (covariance matrix)

$$\Sigma = \frac{1}{m} \sum_{i=1}^m X^{(i)} \times X^{(i)T} = \frac{1}{m} \times X \times X^\top$$

因此如果要降維，我們可以用 U 的前 k 列乘上對應 Σ 當中的特徵向量，就可以得出新的特徵了，運算在幾何當中，其實是將 X 投影到 U 的前 k 個向量



STATIONARY

- **Definition:** A random process whose joint distribution does not change over time.
- **Characteristics:**
 1. μ is constant
 2. σ is constant
 3. no seasonality or periodic behavior
- With such quality, therefore, we could use Gauss-Markov to estimate its parameters and confidence level.

Testing Stationary-Dickey Fuller Test

Consider a sequence y_t .

With lagged 1 assumption, We have,

$$y_t = a + b_1 y_{t-1} + \epsilon_t$$

and hypothesis, $H_o : b_1 = 1$; $H_1 : b_1 < 1$. This is equivalent to,

$$\Delta y_t = a + (b_1 - 1)y_{t-1} + \epsilon_t$$

or

$$\Delta y_t = a + B y_{t-1} + \epsilon_t$$

with hypothesis, $H_o : B = 0$; $H_1 : B < 0$. If we reject the hypothesis, then the sequence is stationary.

Testing Stationary-Augmented Dickey Fuller Test(ADF test)

Also, with higher lagged model, written as

$$y_t = a + \sum_{i=1}^L b_i y_{t-i} + \epsilon_t$$

One can write as form,

$$\Delta y_t = a + B y_{t-1} + \sum_{i=1}^L b_i \Delta y_{t-i} + \epsilon_t$$

Also, with hypothesis, $H_0 : B = 0$; $H_1 : B < 0$.. If we reject the hypothesis, then the sequence is stationary.

- Notice that the t-statistic here follows by Dickey Fuller distribution.

因子建構---加權方式

- SMB、HML

```
SL = sum(data_new_small_low['日報酬率 %']*data_new_small_low['市值'])/sum(data_new_small_low['市值'])
SM = sum(data_new_small_medium['日報酬率 %']*data_new_small_medium['市值'])/sum(data_new_small_medium['市值'])
SH = sum(data_new_small_high['日報酬率 %']*data_new_small_high['市值'])/sum(data_new_small_high['市值'])
BL = sum(data_new_big_low['日報酬率 %']*data_new_big_low['市值'])/sum(data_new_big_low['市值'])
BM = sum(data_new_big_medium['日報酬率 %']*data_new_big_medium['市值'])/sum(data_new_big_medium['市值'])
BH = sum(data_new_big_high['日報酬率 %']*data_new_big_high['市值'])/sum(data_new_big_high['市值'])
```

- RMW

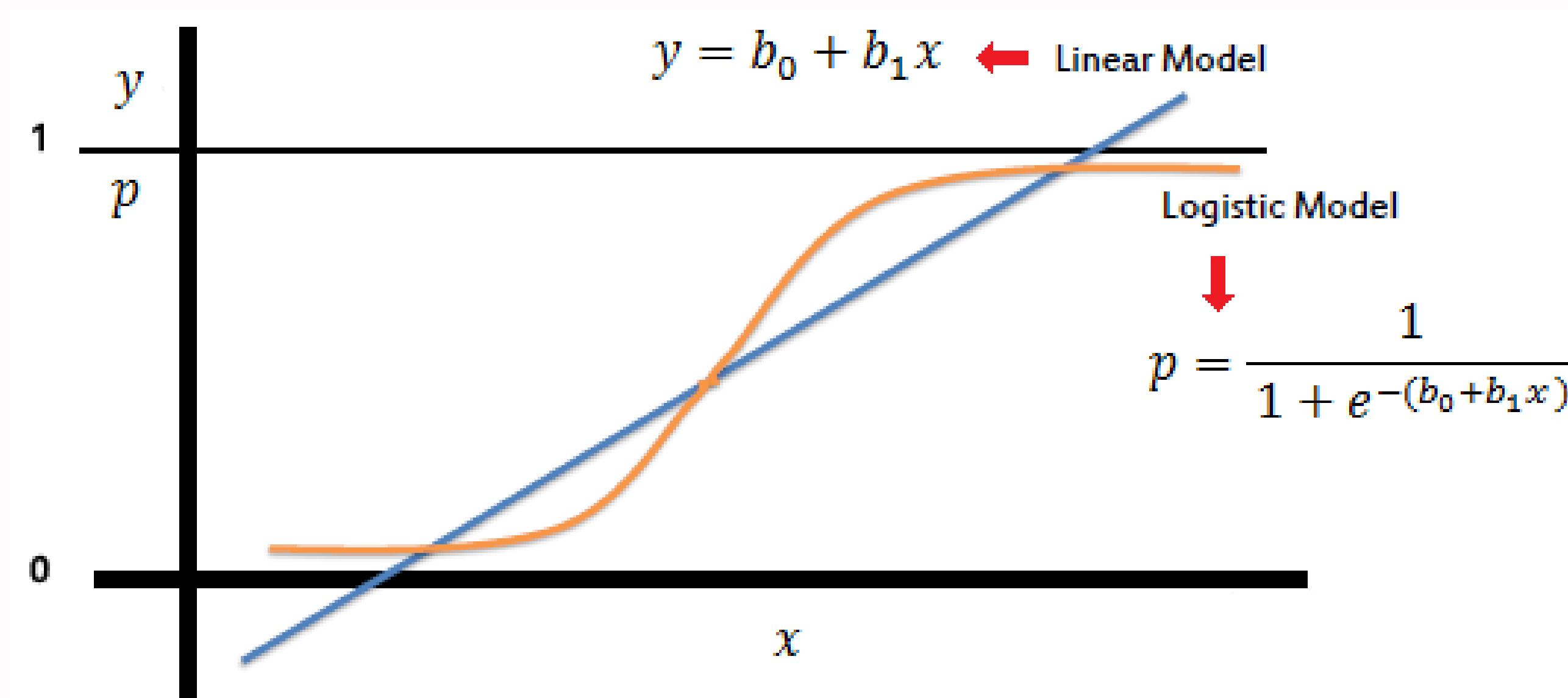
```
BR = sum(merged_data_big_high['日報酬率 %']*merged_data_big_high['市值'])/sum(merged_data_big_high['市值'])
BF = sum(merged_data_big_medium['日報酬率 %']*merged_data_big_medium['市值'])/sum(merged_data_big_medium['市值'])
BW = sum(merged_data_big_low['日報酬率 %']*merged_data_big_low['市值'])/sum(merged_data_big_low['市值'])
SR = sum(merged_data_small_high['日報酬率 %']*merged_data_small_high['市值'])/sum(merged_data_small_high['市值'])
SW = sum(merged_data_small_low['日報酬率 %']*merged_data_small_low['市值'])/sum(merged_data_small_low['市值'])
```

- CMA

```
SC = sum(merged_data_CMA_small_high['日報酬率 %']*merged_data_CMA_small_high['市值'])/sum(merged_data_CMA_small_high['市值'])
SA = sum(merged_data_CMA_small_low['日報酬率 %']*merged_data_CMA_small_low['市值'])/sum(merged_data_CMA_small_low['市值'])
BC = sum(merged_data_CMA_big_high['日報酬率 %']*merged_data_CMA_big_high['市值'])/sum(merged_data_CMA_big_high['市值'])
BA = sum(merged_data_CMA_big_low['日報酬率 %']*merged_data_CMA_big_low['市值'])/sum(merged_data_CMA_big_low['市值'])
```

甚麼是Logistic Regression

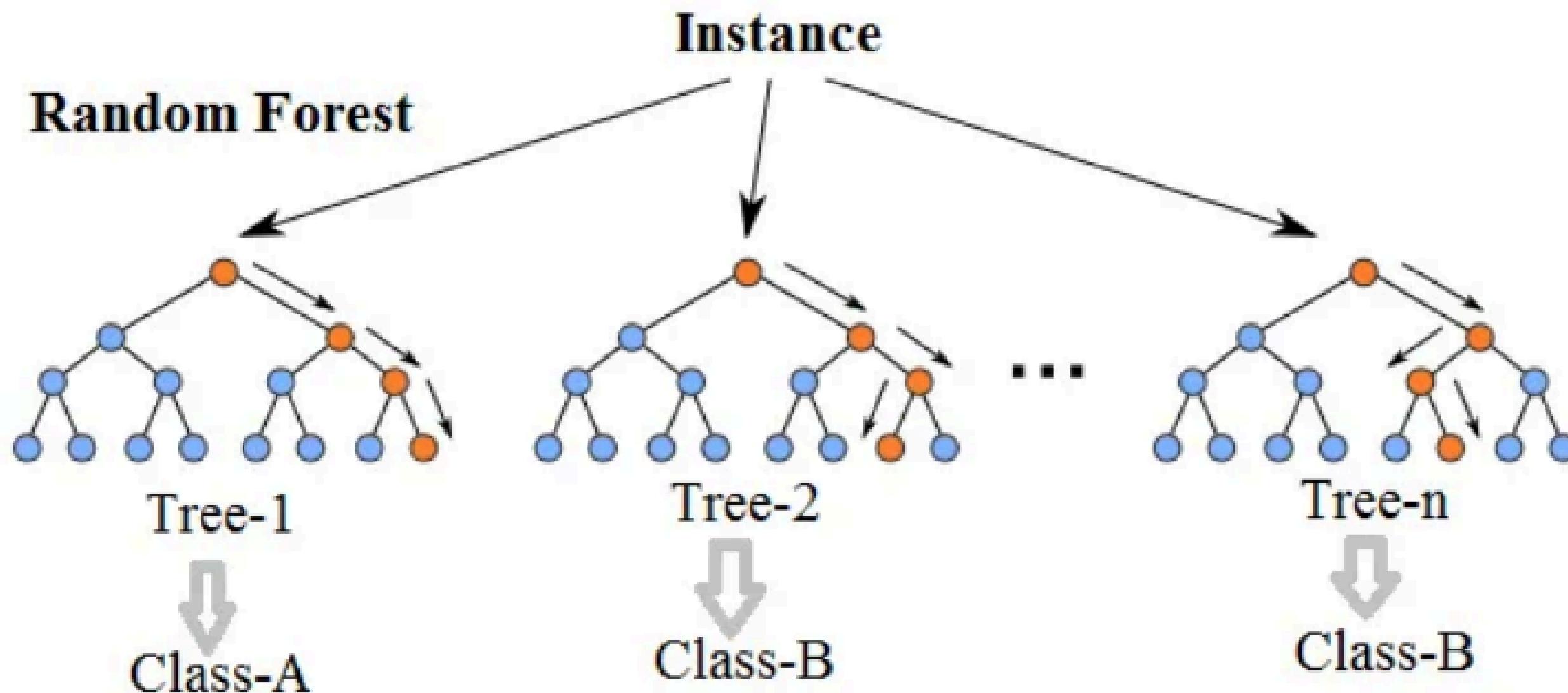
Logistic Regression是一個可以用來作為二元分類器的機器學習模型，其函式來源於簡單線性回歸：



甚麼是Random Forest

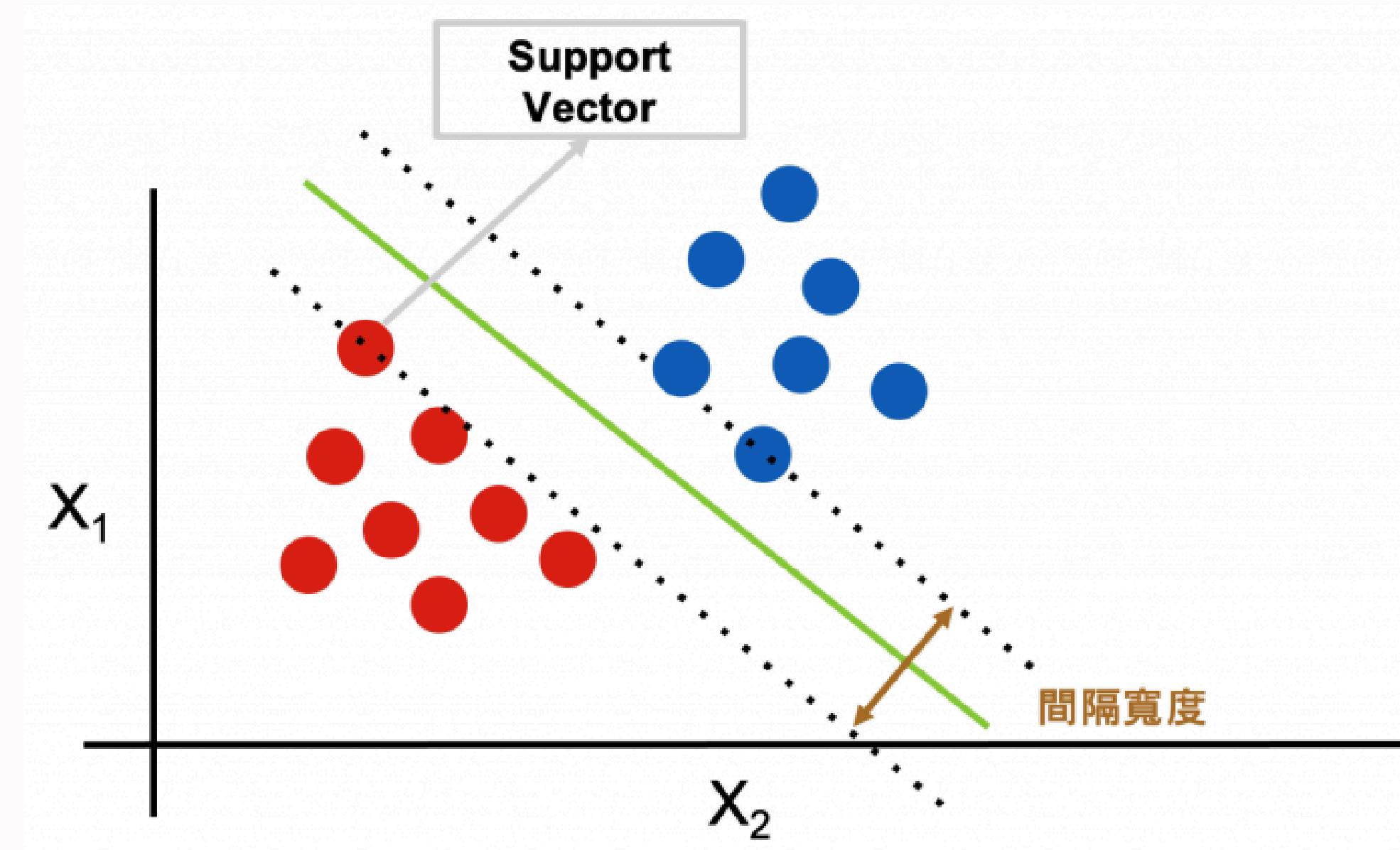
Random Forest是一種決策樹的延伸，其會發展出多條分支數，並以多數決決定測試樣本的歸類，其中分支樹的數量是可以被調整的參數：

Random Forest Simplified



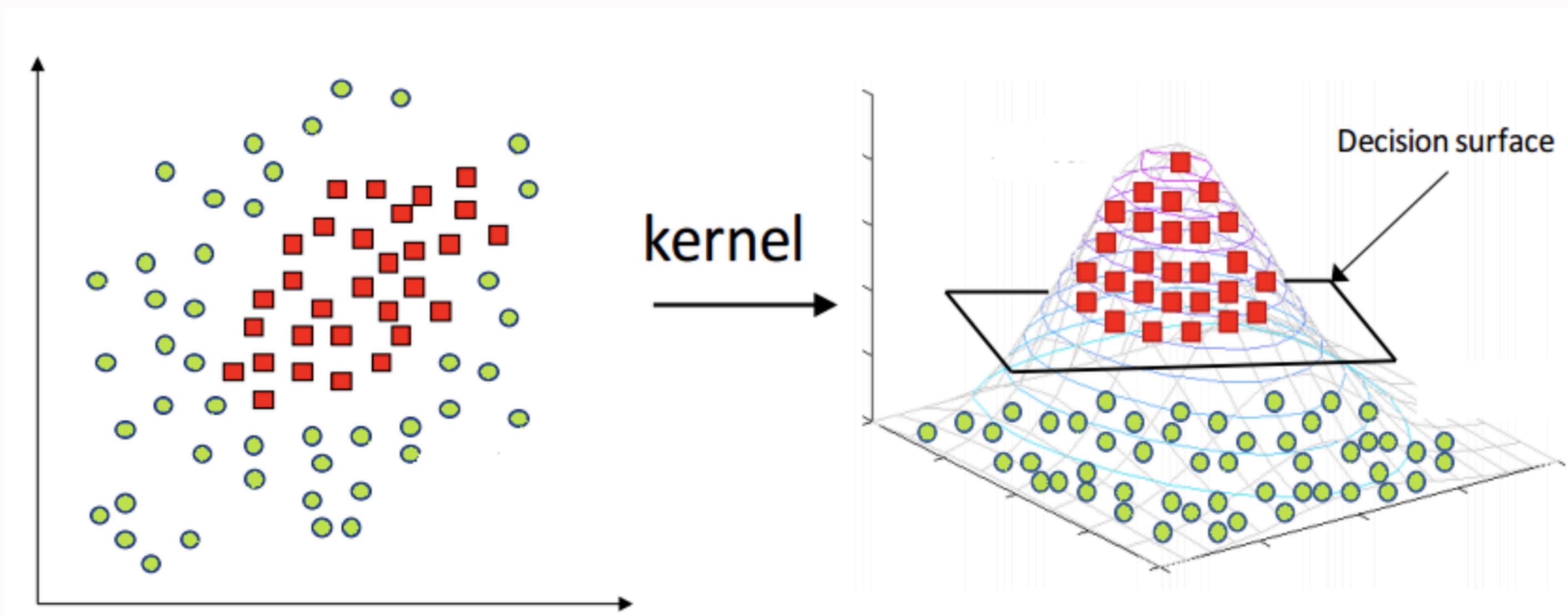
甚麼是Support Vector Machine

SVM是一個用於進行分類的模型，其簡單概念為在不同種類的樣本間畫上界線，並以新樣本與在界線的哪一側決定分類：



甚麼是Support Vector Machine

其中在SVM中的kernel選擇非常重要，當遇上資料無法以純線性切開時，就需要kernel函數將資料進行投影，創造出一個可以被切割開的資料分布



資料說明-資料符號說明

以下為本次預測使用到的特徵變數在說明時的符號表示：

變數	符號		
MKT	MKT_t	IXIC_Return	$IXIC_r_t$
SMB	SMB_t	SOX_Return	SOX_r_t
HML	HML_t	台灣VIX波動度	$TVIX_t$
RMW	RMW_t	Gold_return	$Gold_r_t$
CMA	CMA_t	Oil_Return	Oil_r_t
DJI_Return	DJI_r_t	Jeiba_盤中分數	$Jeiba_m_t$
GSPC_Return	$GSPC_r_t$	Jeiba_盤後分數	$Jeiba_M_t$
		DJI_High/Open	DJI_H/O_t
		DJI_Low/Open	DJI_L/O_t
		DJI_High/Low	DJI_H/L_t
		DJI_High/Close	DJI_H/C_t
		DJI_Low/Close	DJI_L/C_t
		GSPC_High/Open	$GSPC_H/O_t$
		GSPC_Low/Open	$GSPC_L/O_t$
		GSPC_High/Low	$GSPC_H/L_t$
		GSPC_High/Close	$GSPC_H/C_t$

資料說明-資料符號說明

*：變動百分比的計算方式是以 t 時的資料除以 t-1 時的資料做計算

GSPC_Low/Close	$GSPC_L/C_t$
IXIC_High/Open	$IXIC_H/O_t$
IXIC_Low/Open	$IXIC_L/O_t$
IXIC_High/Low	$IXIC_H/L_t$
IXIC_High/Close	$IXIC_H/C_t$
IXIC_Low/Close	$IXIC_L/C_t$
SOX_High/Open	SOX_H/O_t
SOX_Low/Open	SOX_L/O_t

SOX_High/Low	SOX_H/L_t
SOX_High/Close	SOX_H/C_t
SOX_Low/Close	SOX_L/C_t
*	DJI_Volume_change
*	$DJI_V_change_t$
*	GSPC_Volume_change
*	$GSPC_V_change_t$
*	IXIC_Volume_change
*	$IXIC_V_change_t$
*	TWD/USD_change
*	TWD/USD_change_t
*	Jeiba盤中分數變動%
*	$Jeiba_m_change_t$

資料說明-資料符號說明

* : 變動百分比的計算方式是以 t 時的資料除以 t-1 時的資料做計算

*	Jeiba盤後分數變動%	$Jeiba_M_change_t$
*	台灣VIX波動度變動%	$TVIX_change_t$

預測目標	Y_{t+1}	大盤在第t+1天的漲或跌
------	-----------	--------------

模型效能衡量指標說明

		真實狀況	
		事實為真	事實為假
預測狀況	預測為 陽性	tp	fp (Type I error)
	預測為 陰性	fn (Type II error)	tn

精確率 (Precision) : $tp/(tp+fp)$

召回率 (Recall) : $tp/(tp+fn)$

f1-score : $2 / ((1/\text{Precision}) + (1/\text{Recall}))$

模型效能衡量ROC說明

ROC曲線是一個用來衡量模型預測力的指標之一，以FPR為x軸，TPR為y軸，來測試模型的預測能力是否有勝過隨機預測或等同隨機預測

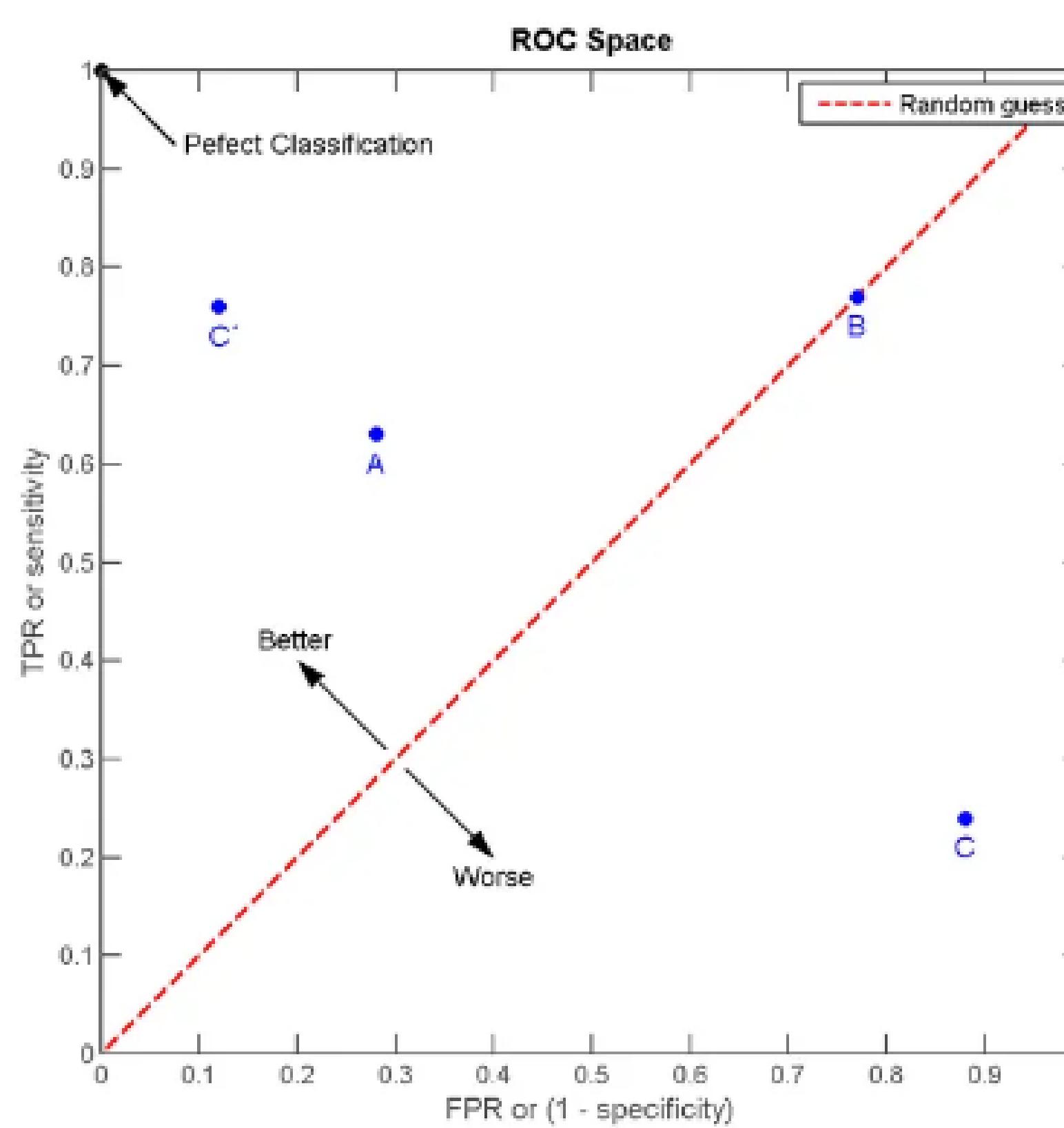
- TPR：在所有實際為陽性的樣本中，被正確地判斷為陽性之比率。

$$TPR = TP / (TP + FN)$$

- FPR：在所有實際為陰性的樣本中，被錯誤地判斷為陽性之比率。

$$FPR = FP / (FP + TN)$$

模型效能衡量ROC說明



模型效能衡量ROC說明

