新聞情緒如何帶動股價走勢

蘋概股三雄：台積電、鴻海、大立光

組別：伍陸柒捌

工管四 廖守三

會計三 吳佳安

地理二 紀德鑫

圖資一 張以柔

目錄

壹、研究動機

貳、主題簡介

參、資料取徑

肆、分析流程架構

伍、資料清洗與特徵工程

陸、成果評估

壹、研究動機

近年來，金融商品不斷地推陳出新，例如：股票、基金、期貨、ETF……等，使投資人在理財管道擁有多樣化的選擇，而透過股票進行交易則是大眾最廣為熟悉的理財工具之一，因此，如何選擇股票標的以獲得最大利益，即為投資人最關注的議題。

常見的股票分析方法分為兩種：基本分析與技術分析，基本分析主要考量上市公司的營運、財務狀況與分析各項財務指標，藉以預測企業未來營運狀況與股東獲利能力，作為股東選擇股票的依據；技術分析則著重歷史股價變動與金融市場資訊，從中分析股價漲跌幅趨勢間的特徵，藉以預測股價未來的漲跌幅狀況，以作為股票買賣的依據。

基本分析與技術分析仍亦有其弱點，那便是忽略投資決策者容易受到外部環境的影響而改變標的選擇，當今網路與社群媒體發達，消費者經常藉由網路上取得的資訊來決定其消費行為，以股市交易為例，與股票相關的網路新聞經常能左右投資者的決策，然隨著數據與資訊量以極快的速度增長，投資者難以消化來自各方豐富且龐大的資料，其中，質化的資料中經常含有更大量的資訊，因此藉由文字探勘可以有效分析質化資訊並轉化成對於投資者最有價值的資訊。

貳、主題簡介

在標的選擇上，考量標的討論度不能太低否則訓練資料過少，同時該標的股價須易受國際局勢波動，漲跌的比例相當。結合上述條件，本組發現蘋概股非常適合作為研究對象。

蘋概股即為蘋果概念股，定義為上下游與蘋果公司密切合作的廠商。這類型公司非常容易受到政策、景氣、高層、技術的影響。台灣為蘋果供應鏈中不可或缺的一個角色，台灣的蘋概股又以台積電、大立光與鴻海最具代表性，亦為台灣股市中具有顯著影響力的企業，本組以上述三家企業自中美貿易戰（2018年3月自2019年11月）至中美兩國達成協議前的時期，蒐集相關新聞與股價，以文字探勘進行分析與對比，並驗證在事件

日前若有相關新聞的發布，是否對於近日股價有所影響，探討股價是否有提前反應之效果。

參、資料取徑

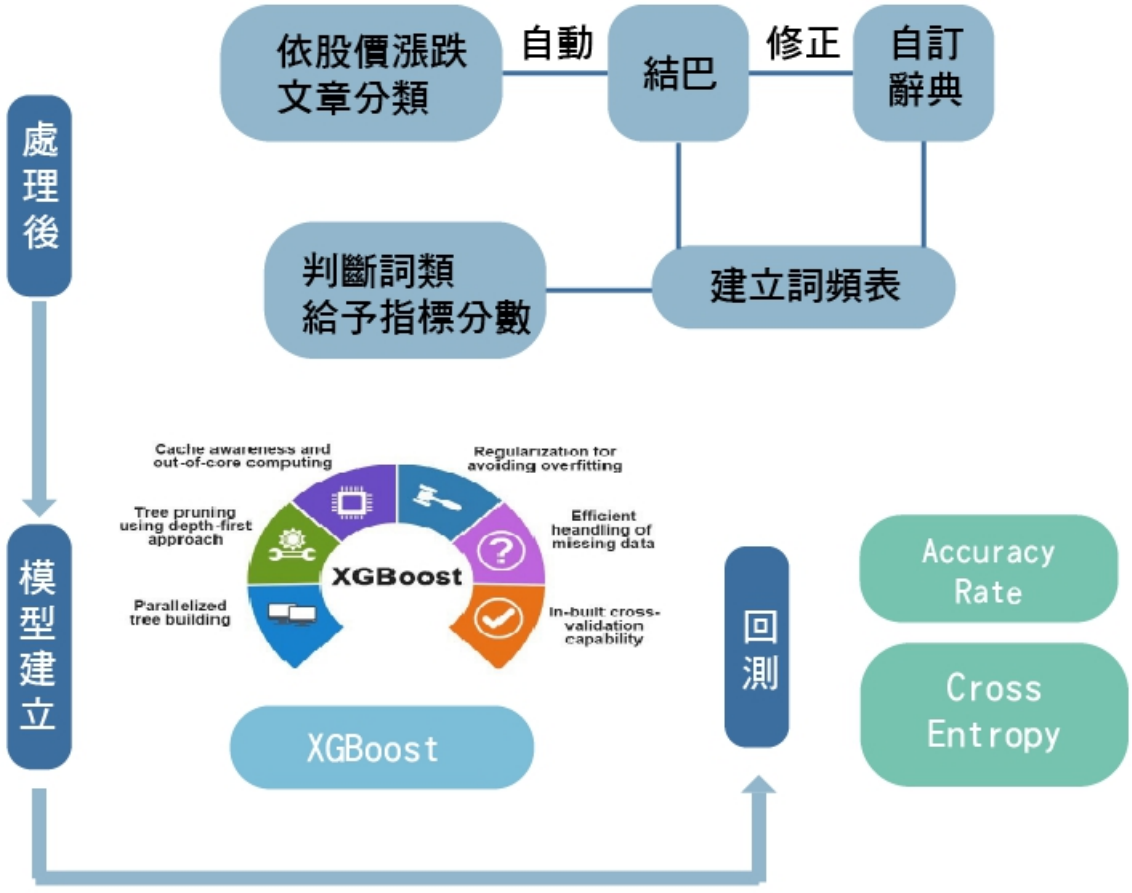
一、新聞資料取得

利用爬蟲套件，蒐集中美兩國達成協議前，中美貿易戰（2018年3月至2019年11月）時期網路（udn聯合新聞網、ptt等）台積電、大立光與鴻海的相關新聞。

二、股價資訊取得

利用爬蟲套件，蒐集相對應時期上述三家企業的股價，並與前一日收盤價漲跌幅超過2%作為標準，篩選出較有顯著性的股價漲跌趨勢。

肆、分析流程架構



伍、資料清洗與特徵工程

關於當天股價究竟是漲還是跌，我們給此一個核心定義：

- 漲：若當天股價的收盤價比前一日的最高價還要高出2%，那麼我們有足夠信心稱當天為股價飆漲日

- 跌：若當天股價的收盤價比前一日的最低價還要低2%，那麼我們有足夠信心稱當天為股價慘跌日

- 不漲不跌：由於本組對特殊事件要求嚴格，因此不符上述條件的其餘交易日都被歸類為正常波動

根據上述條件，在每日收盤後給予該日「漲」「跌」「正常」的訊號標籤，接著我們定義何謂「有影響力」的產經新聞。以下是我們的定義：假設8/22收盤後，該日被認定是漲訊，則8/22往前推3日(8/21、8/20、8/19)的新聞被歸為漲訊文章；相反地，若8/22收盤後，該日被認定是跌訊，則8/22往前推3日(8/21、8/20、8/19)的新聞被歸為跌訊文章不選擇8/22當天的新聞的，是因為我們發現投資人的反應並未那麼靈敏，多半會希望多抱幾天才脫手。

處理完每篇新聞的漲跌訊號之後，我們分別對正向新聞和負向新聞切詞(使用結巴內建字典及本組自定義字典，因此在對公司名稱、社會人物的切詞會較準確)並且使用多個指標篩出真正能代表正向負向文章的字詞。指標包含td、df、td-idf、Mutual Information、tfChi-Square、dfChi-Square、Lift值等等。

- Mutual Information

$$MI = \log \frac{P(x,y)}{P(x)P(y)} = \log \frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N}\frac{f(y)}{N}} = \log \frac{f(x,y)}{f(x)f(y)}$$

P : probability

N : size of the corpus

f(x) : the occurrences of term x in the corpus

f(y) : the occurrences of term x in the corpus

f(x,y) : the co-occurrences of term x and y in the corpus

x 該類別之事件
y 該詞之事件
x,y 該詞在該類別之事件

			f(x,y)	f(x)		f(y)	f(x,y) / f(x)f(y)		
no	term	tf	df	N	tf-idf	df in all corpus	MI	tf-idf * MI	Rank
1	歐晉德	93	12	421	15.8	114	0.0003	0.0039	5
2	以色列	28	3	421	15.5	62	0.0001	0.0018	9
3	小白兔	26	3	421	15.2	16	0.0004	0.0068	3
4	能源	39	6	421	15.0	285	0.0001	0.0008	10
5	金門	53	9	421	14.8	150	0.0001	0.0021	8
6	不分區立委	27	4	421	14.6	15	0.0006	0.0093	1
7	假釋	15	1	421	14.5	6	0.0004	0.0057	4
8	募兵捐	34	6	421	14.4	29	0.0005	0.0071	2
9	李登輝	26	4	421	14.4	36	0.0003	0.0038	6
10	禁閉	26	4	421	14.4	60	0.0002	0.0023	7

- Use χ^2 (chi square)

IDF會受到極端稀有詞影響，例如某單篇才有的用法，其IDF值大

$$\chi^2 = \frac{\sum(O-E)^2}{E}$$

O: observed value　　E: expected value

別意義，其為(200-100)/100.

χ2常用在類別檢定。假設每一篇文章都當成是「一類」，若「馬英九」一詞總共出現1000次，出 現在10篇中，每篇的期望次數是1000/10=100次 。此時若有一篇出現「馬英九」 200次，代表這 個詞對這篇有特

$$\text{lift} = \frac{\frac{\text{該詞出現在該類別之篇數}}{\text{該類別篇數}}}{\frac{\text{該詞出現之篇數}}{\text{總篇數}}}$$

以上是三種我們有加入參考的指標。在多項指標中，本組最終選擇 **tf-idf*MI*df卡方值**的前1000名字詞並使用CountVertorizer 將每一篇文章轉化為1x2000的稀疏矩陣向量。以下是本組利用 tf-idf*MI*df的標準所挑出的各企業關鍵詞

鴻海

排名	正向詞	df 卡方*tf-idf＊MI	負向詞	df 卡方*tf-idf＊MI
1	格力	218.675917	ndy	79.490587
2	六房	154.359471	umsci	73.375926
3	nice	154.359471	楊瓊瓔	73.375926
4	澎思	147.927827	壽山	48.917284
5	烏日區	147.927827	郭王會	39.996418
6	調和式	141.496182	dollars	36.687963
7	董明珠	136.997170	曾健	36.687963
8	科睿	178.408193	yahoo	31.305485
9	失智症	126.614503	逐鹿	30.573303
10	李香蘭	115.769603	kuso	30.573303

大立光

排名	正向詞	df 卡方*tf-idf＊MI	負向詞	df 卡方*tf-idf＊MI
1	大立光	334.235147	健全	87.428739
2	台股	295.362808	碎鏈	32.785777
3	蘋果	291.613853	院長	27.321481
4	指數	238.200849	釋單	22.764765
5	億元	207.531354	交法	21.857185
6	表現	194.256004	葉子菁	21.857185
7	市場	186.433270	科創	21.857185
8	台積電	178.408193	富比士	21.857185
9	手機	165.762658	陸客	21.857185
10	華為	165.237405	喘息	21.510727

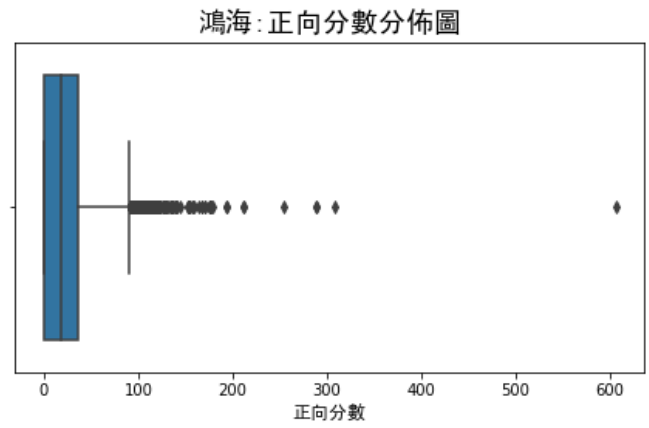
台積電

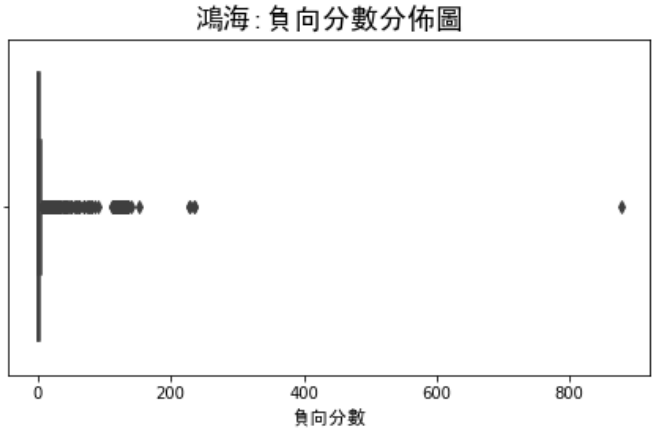
排名	正向詞	df 卡方*tf-idf＊MI	負向詞	df 卡方*tf-idf＊MI
1	停班	44.594174	rec	83.557167
2	Rtec	39.019902	tvbs	33.422867
3	林錫銘	27.871359	Google	27.852389
4	米塔	27.871359	李錦記	22.281911
5	往元	27.871359	郭偉政	22.281911
6	童年	22.297087	柏亨	22.281911
7	Azor	22.297087	借錢	22.281911
8	均華	20.897230	Rrazer	22.281911
9	vocs	16.722815	蕭惠中	16.711433
10	app	16.722815	熱區	16.711433

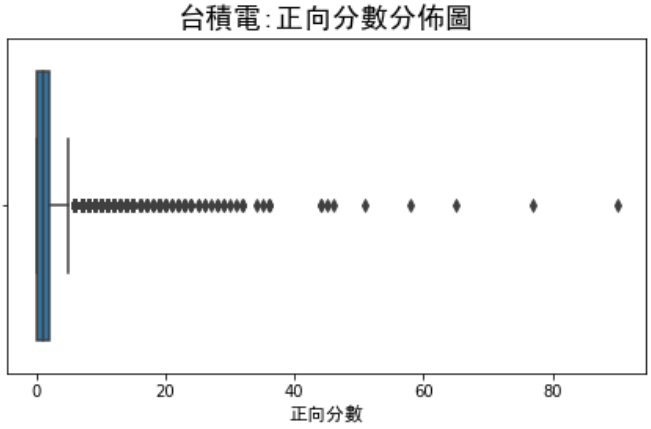
根據文章有的正負向字詞豐富度，可以加權算出該篇文章的正負向情感分數和整理情感分數。(該篇文章每出現1個正向詞，正向分數就加+1分，每出現一個負向詞，負向分數就+1)

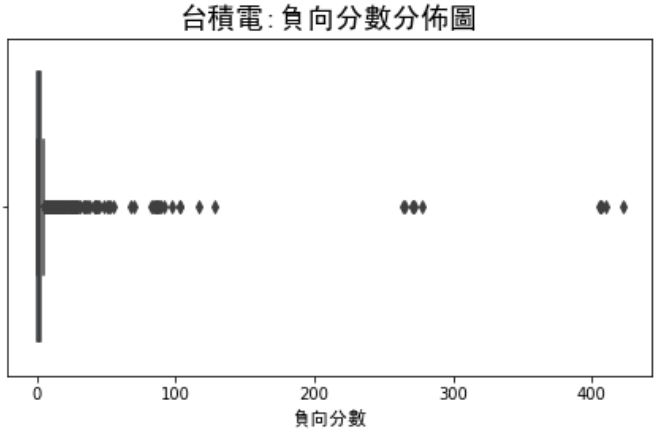
舉例來說：一篇2017-05-16所發佈的新聞「到法拉利上班不是夢！ 城市科大畢業生成超跑技師」，因為內文出現6個正面詞彙，因此正向分數=6，同理可得出負向分數，並且情緒分數 = 正向分數-負向分數。

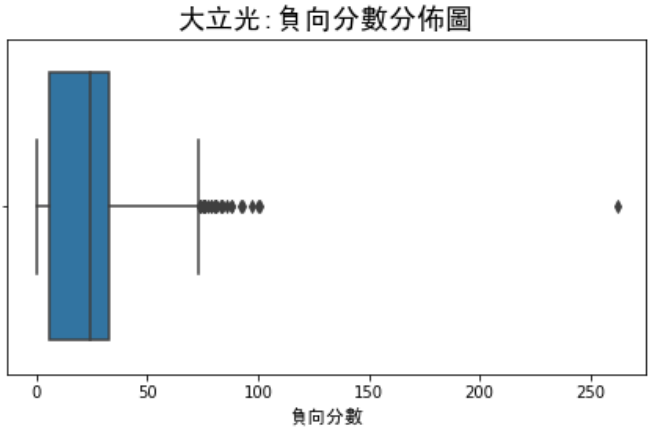
以下為三家企業的正負向分數盒狀圖：

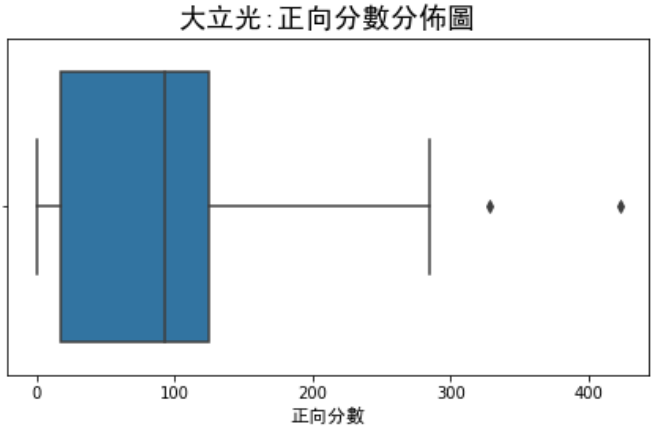

鴻海:正向分數分佈圖
0
100
200
300
400
500
600
正向分數


鴻海:負向分數分佈圖
0
200
400
600
800
負向分數


台積電:正向分數分佈圖
0
20
40
60
80
正向分數


台積電:負向分數分佈圖
0
100
200
300
400
負向分數


大立光:負向分數分佈圖
0
50
100
150
200
250
負向分數

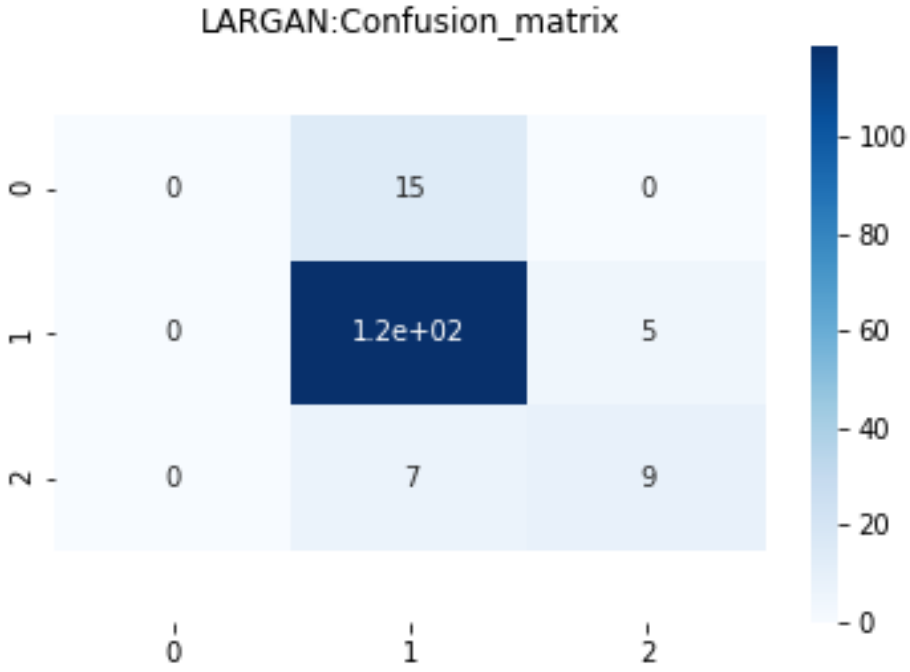

大立光:正向分數分佈圖
0
100
200
300
400
正向分數

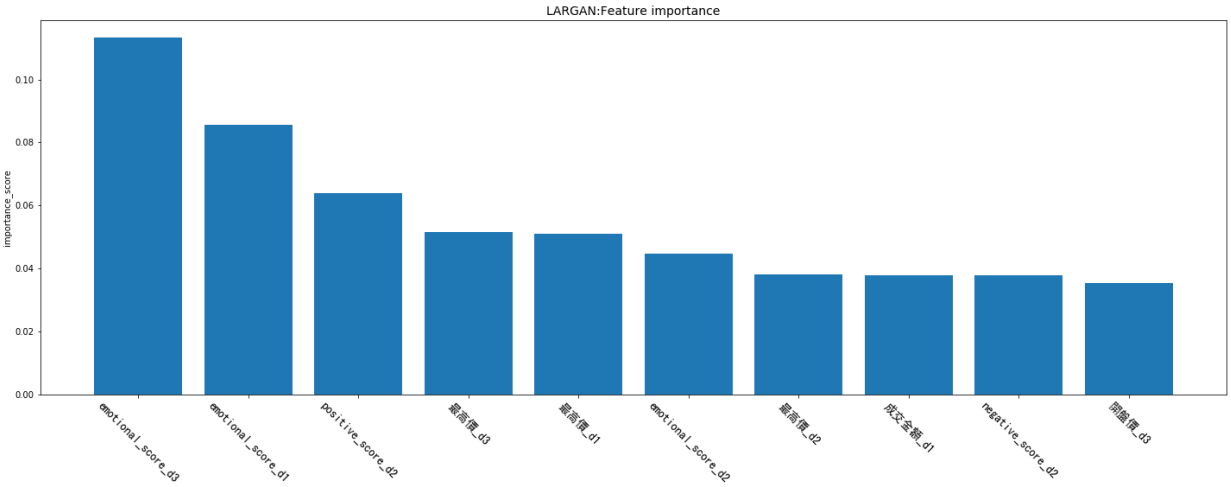
陸、成果評估

一、大立光

1.準確率：83%

2.重要特徵：3日前情緒分數/2日前正向分數/3日前最高價

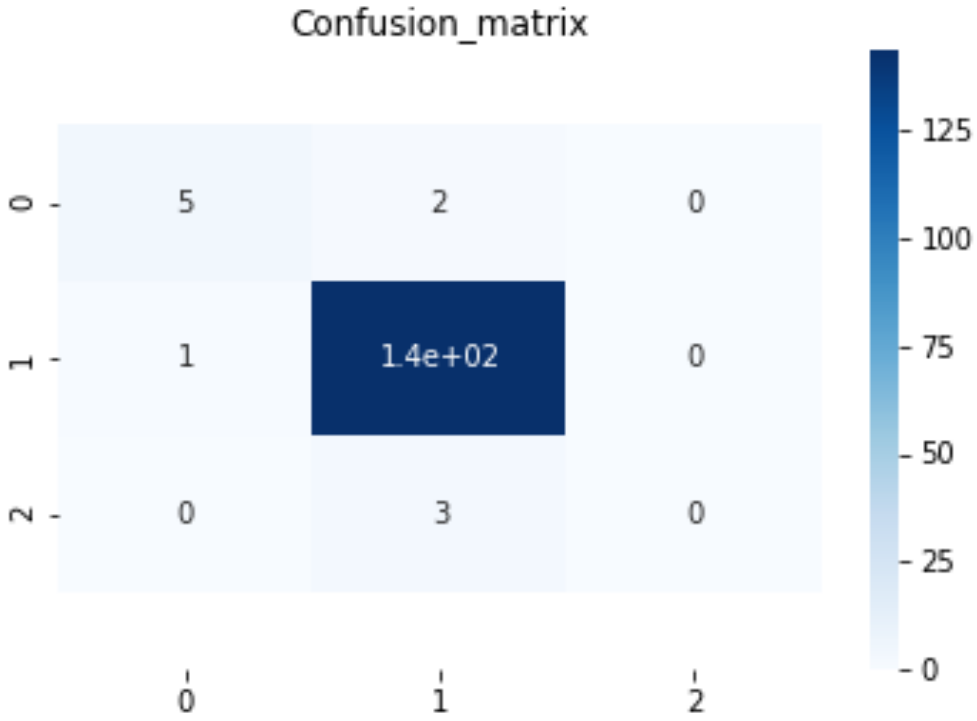


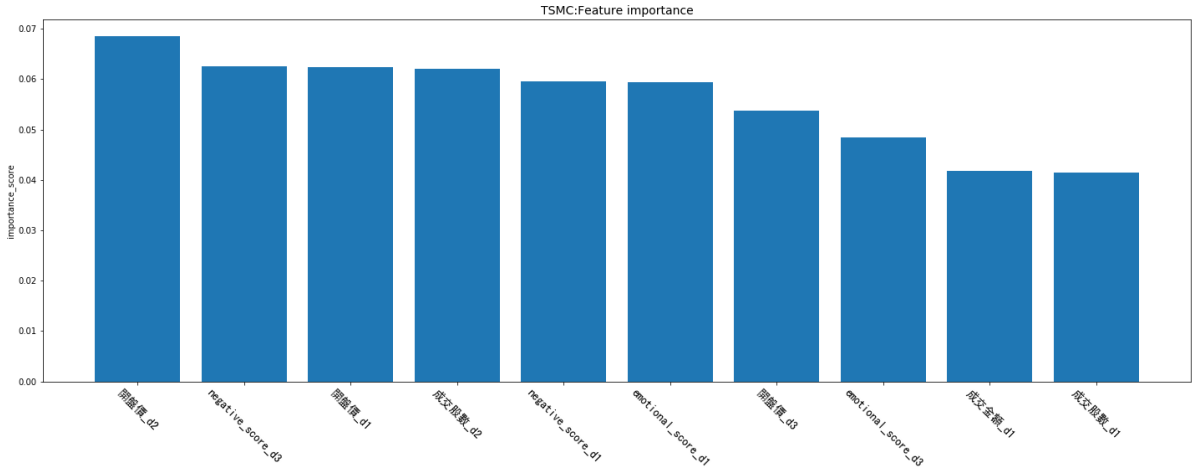


二、台積電

1.準確率：97%

2.重要特徵：2日前開盤價/3日前負向分數/1日前開盤價





三、鴻海

1.準確率：92%

2.重要特徵：3日前正向分數/2日前情緒分數/2日前開盤價