

BDA GROUP12

MIDTERM PROJECT

徐嫵鎔 廖文豪 陳宜君 鄭皓 陳牧忠 藤田教譽



CONTENTS

MAIN TOPICS

關鍵步驟

文章集分析

雜訊過濾

權值股討論熱度排名

股票漲跌定義

特徵選擇

訓練模型

事件偵測

出手頻率

預測結果

關鍵步驟

1

篩後文章籍

切出訓練和測試文章集
剔除雜訊文章



2

找出特徵字

挑選預測個股
分出漲跌文章集



4

測試模型

推測出手時機



3

訓練模型

算出特徵權重
切出K Fold



文章集分析

新聞當訓練和判斷

- 量大
- BBS和論壇的幫助不顯著

類別	總篇數
新聞*	232207
BBS.	46158
論壇	20025

熱絡度分析



雜訊過濾

目的(過濾無法標記的文章)

- 漲跌在同一篇出現的文章
- 多支股票出現的文章

利用在篩選新聞文章的關鍵字

- keyword_list= ['盤後','盤中','盤前','交易概況','上市認購','農訊','各報要聞','報價簡訊','台北股市','海外存託憑證','國內匯市','證交所','y早報','y晚報','焦點新聞','投顧','晨間解析','集中市場']
- paradox_a= ['買進','賣出']
- paradox_b = ['上漲','下跌']
- paradox_c = ['跌破','衝上']
- paradox_d = ['買超','賣超']

權值股熱度討論

新聞

(['台積電', '鴻海', '大立光', '南亞', '統一', '台塑', '華新', '國泰金', '聯發科', '友達',
'國巨', '富邦金', '聯電', 華新科', '中華電', '可成', '中信金', '群創', '中鋼', '南亞科',
'新光金', '華邦電', '兆豐金', '第一金', '大同', '旺宏', '台達電', '開發金', '台塑化', '玉山金',
'元大金', '長榮', '台灣大', '台泥', '和碩', '台化', '永豐金', 'GIS', '巨大', '遠東新',
'英業達', '上銀', '華碩', '台新金', '智邦', '欣興', '大成鋼', '臻鼎', '亞泥', '微星',
'台勝科', '緯創', '統一超', '廣達', '儒鴻', '遠傳', '華航', '仁寶', '力成', '長榮航',
'華南金', '亞德客', '中租', '光寶科', '臺企銀', '鴻準', '日月光投控', '合庫金', '台灣高鐵',
'聚陽', '中壽', '瑞昱', '瑞儀', '研華', '聯詠', '致茂', '豐泰', '彰銀', '矽力', '興富發',
'大聯大', '潤泰全', '健鼎', '上海商銀', '寶成', '美利達', '正新', '和泰車', '福懋', '聯強',
'東元', '群光', '潤泰新', '億豐', '台中銀', '裕日車', '佳格', '台肥', '旭隼', '宏碁'],

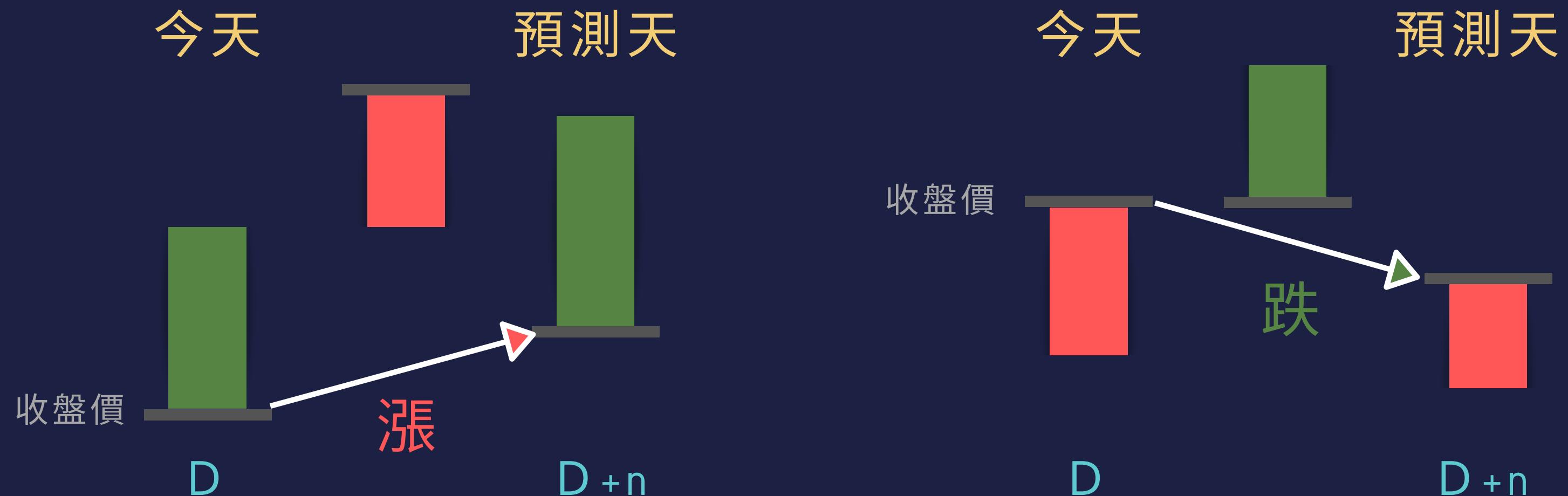
BBS

(['台積電', '鴻海', '南亞', '國巨', '中鋼', '中信金', '大立光', '統一', '華新', '友達',
'群創', '聯電', '兆豐金', '台塑', '新光金', '聯發科', '國泰金', '玉山金', '第一金', '中華電',
'長榮', '富邦金', '華新科', '大同', '旺宏', '台新金', '華邦電', '巨大', '永豐金', '元大金',
'宏碁', '開發金', '可成', '緯創', '南亞科', '微星', '和碩', '台灣大', '仁寶', '台泥',
'華航', '英業達', '智邦', '華南金', '長榮航', '廣達', '台達電', '合庫金', 'GIS', '台化',
'中壽', '力成', '華碩', '欣興', '彰銀', '上銀', '亞泥', '遠東新', '遠傳', '臺企銀', '臻鼎',
'中租', '正新', '台塑化', '大成鋼', '寶成', '光寶科', '儒鴻', '瑞儀', '興富發', '鴻準',
'潤泰全', '大聯大', '台灣高鐵', '聯詠', '聯強', '瑞昱', '台勝科', '聚陽', '統一超', '豐泰',
'東元', '潤泰新', '健鼎', '台中銀', '致茂', '研華', '億豐', '上海商銀', '亞德客', '日月光投控',
'美利達', '裕日車', '和泰車', '福懋', '群光', '台肥', '矽力', '佳格', '旭隼'],

論壇

(['鴻海', '台積電', '中華電', '統一', '中信金', '第一金', '中鋼', '大立光', '巨大', '南亞',
'國巨', '聯電', '台塑', '聯發科', '玉山金', '友達', '大同', '台新金', '兆豐金', '國泰金',
'台灣大', '仁寶', '長榮', '群創', '永豐金', '富邦金', '華新', '旺宏', '元大金', '遠傳',
'宏碁', '上銀', '華新科', '微星', '和碩', '台泥', '台中銀', '華航', '力成', '合庫金',
'開發金', '新光金', '台達電', '中壽', '華碩', '統一超', '彰銀', '廣達', '緯創', '南亞科',
'華邦電', '可成', '亞泥', '興富發', '鴻準', '華南金', '寶成', 'GIS', '臻鼎', '台灣高鐵',
'英業達', '長榮航', '潤泰全', '正新', '智邦', '欣興', '中租', '瑞儀', '儒鴻', '台勝科',
'潤泰新', '聚陽', '遠東新', '東元', '大成鋼', '台塑化', '臺企銀', '上海商銀', '瑞昱', '大聯大',
'光寶科', '台化', '聯強', '研華', '和泰車', '豐泰', '福懋', '佳格', '致茂', '台肥', '聯詠',
'亞德客', '美利達', '矽力', '群光', '健鼎', '日月光投控', '裕日車', '億豐', '旭隼'],

股票漲跌定義





特徵選擇

挑出個股文章(利用個股名字當關鍵字) $h=1$

- $\delta = 0.0$

選取方法

- TF-IDF、TFDF2、Chi-Square 跑的速度最快結果最好

總數

- 5000、10000 對於某些個股才有差異

台積電(漲)特徵

1 chi2_up_features[0:100]

```
array(['端午', '法國', 'ic', '反彈', '年線', '開紅盤', '台揚', '德意志', 'nm', '券商', '股市', '脫歐', '公投', 'etf', '零股', '翻揚', '計算', '設計', '液化', '開市', '格羅方德', '英文', '三', '重建', '台苯', '鈦', '最長', '奈米', '宏達電', 'rs', '英國', '歐洲', '新政府', '元月', '網友', '大利', '南光', '中秋', '巨頭', '銀行', '程', '克宏', '土壤', '獨家', '反向', '交棒', '簽約', '園區', '活水', '來台', '展訊', '打臉', '六月份', '財金', '輪', '奈', '落腳', '工程', '缺口', '希望', '股納入', '重返', '曆', '會議', '股東會', '生', '米', '法國大選', '南海', '台南', '義', '季減', '歐系', '興富發', '南', '買', '下周', '格里', '建漢', '馬', '康樂', '當家', '歐派', '國際', '仲琦', '欣技', '大昌', '善意', 'ipc', '籌資', '重罰', '麥', '池塘', '新廠', '奈米廠', '未過', '矽品', '晶圓', '睽', '中間'], dtype='<U17')
```

鴻海(漲)特徵

1 chi2_up_features[0:100]

```
array(['公投', '脫歐', '電視', '海信', '封關', '面板', '英國', '印度', '郭台銘', '威斯康辛', 'robohon', '新政府', '股東', '仟元', '千股', '股東會', 'ps', '端午', '寬頻', '飛鷹', 'aptt', '華亞科', '州', '退休', '江申', '除息', 'msci', '漲', '職演', '麥格理', '向上', '光法', '公平', '本土', '英脫', '量', '開紅盤', '負債', 'oled', '東會', '台聚', '配合', '權重', '今晚', '漲點', '轉強', '造', '遊戲', '誠意', '歐派', '有線', '擴張', '開高', '走揚', '千億量', 'tbc', 'ubi', 'smart', '泛鴻海', '激勵', '新春', '群創', '下周一', '攻克', '聯詠', '橡', '匯流', '夏普股', '生效', '德宏', '尾牙', '網站', '赴美', '缺口', '挑戰前', '說', '日經連', '建漢', '備貨量', '往上走', '景氣', '突破', '恢', '英國脫', '冷軋', '稅後', '留歐', '小股', '販售', '預計', '脫', '集團', '吸引力', '娶', '往前', 'sumc', '點序', '蔡', '阿里', 'sony'], dtype='<U15')
```

大立光(漲)特徵

chi2_up_features[0:100]

```
array(['恩平', '美元', '林', '觀望', '說', '航運股', '同致', '量', '反彈', '星歐', 'etf', '聯發', '曆', '新政府', '醞釀', '星巴克', '記錄', '年減', '砸', '支出', '懷特', '股史', '虹膜', '砍價', '端午', '台半', '瞄準', '以下', '明光', '反向', '榮海', '瑞聲', '週線連', '一季', '揚', '法國', '股東會', '軋空', '元月', '下台', '輪', '生效', '巴菲特', '通訊', '財務', '股新', '年前', '辭', '南滋', '池塘', '贖', '金居', '低', '受', '領軍下', '級法', '朋程', '型態', '魔咒', '以外', 'vr', '不好', '雙鏡頭', '攻抵', '萬戶', '灣藍籌', 'sunedison', '點後拉回', '地盤', '配', '專家', '農', '今收', '陽明', '神盾', '智易', '施祖成', '結盟', '席', '爆氣彈', '瞬間', '中小', '記憶體', '糾結', '榮成', '日以', '光法', '態', '友通', '難漲', '認售', '大降價', '敵台', '仍連', '龍燈', '關將', '季減', '太', '榮', '措施'], dtype='<U11')
```

台積電(跌)特徵

1 chi2_down_features[0:100]

```
array(['希拉', '蕊', '重挫', '財報', '熊本', '北韓', '電郵門', '跌停', '新高', '降評', '大跌', '下跌', '開票', 'fbi', '欣銓', '美國大選', '前瞻', '大選', '殭', 'ai', '美大選', '狹幅', '美總統', 'ibiden', '博通', '跌破', '科技股', '盤後股價', '停電', '權重', '摩台', '遭雙降', '嘉聯益', '停工', '延攬', '起訴', '拆解', '電遭', '鎖碼', '失守', '減碼', '目標價', '大摩調', '學', '守月線', '恩智浦', '東芝', '回測', '感恩', '國家級', '加劇', '康控', '指期', '箱型', '正價差', '進機', '劉金標', '屍股', '曾銘宗', '蕭乾祥', 'led', '郭', '耳機', '北韓試', '蔣尚義', '爆氣彈', '不法', '人工智慧', '千興', '步入', '未定', '口至', '周年', '期現', '再創', '壓', '殺盤', '空方', '暫', '真空', '竹北', '再飆', '董', 'sunedison', '大供', '啃', '畸形', '小跌', '鴻海領', '單季', '黨候', '阿富汗', '再來', '繩', 'nand', '一度', '新至', '事業', '震盪', '拖累'], dtype='<U17')
```

鴻海(跌)特徵

1 chi2_down_features[0:100]

```
array(['大跌', '重挫', '下跌', '跌幅', 'fbi', '北韓', '跌破', '希拉', '蕊', '賣超', '砍', '攢', '東芝', 'ces', '熔斷', '富士通', '蘋果', '韓聯盟', '飛彈', '王仲良', '震旦', '繩', '逾百點', 'pc', '電郵門', '去年同期', '雷射', '機制', '大選', '破月線', '地緣', '阿富汗', '機器', '熱賣', 'pro', '暴跌', 'google', 'pixel', '外資則', '起訴', 'jdi', 'ipo', '殺', 'fii', 'socket', '管制', '感恩', '破', '戴正吳', 'purley', '履約價', '日美', '防禦型', '調查局', '減少', '廣達', '虧損', '台廠', 'wwdc', '鞋', '護盤', '營收', '名元', 'robo', '和沛', '尹衍', '超台灣', '出資額', '季季', '魏永祥', '十月', '下挫', 'ai', '萬點', '失守', '結算', 'nb', '陣營', '洛克', '緯創', '股紅', '回夏普', '要鴻海', '油電氣類', '回測', '台新金', 'tmc', '人工智慧', '人民', '兆元', '欣銓', '遭遇', '賣壓', '上市', '勵案', '投資獎', '炸彈', '關島', '翟本喬', '拍檔'], dtype='<U15')
```

大立光(跌)特徵

chi2_down_features[0:100]

```
array(['萬點', '希拉', '蕊', '電郵門', '熊本', '失守', 'mate', '三星', '併', '股票', 'fbi', 'macbook', '塑化', '概況', '重挫', '鴻', '月合', '款', '新機', '環境', '戰萬點', '結算', '熔斷', '機制', '鴻海', '年線', '跌破', '觀測', '光磊', '獎金', '市場鉅額', '起訴', '甦', '高台', '亞聚', '福利', '維持', '轉倉', '年薪', '晶電', '月營', '震天', 'go', '跌幅', '警示', '選人', '收創', '川習', '銀', '明日', '最壞', '作業員', '當日', 'pro', '旺旺', '六個', '營收', '潤泰', '降評', '沙烏地', '全額', '天王', '阿拉伯', '本日', '積極', '收增', '回測試', 'cam', '炫', '宣祖', '招募', '和解', '過熱', '油電氣類', '美國大選', '切勿', '中概', '扯後腿', '巨蛋', '敘利亞', '群聯', '展現', '餐飲及', '大眾', '旅遊', 'ky', '華為', '鴻家', '太陽能', '豐金', '中多', '美大選', '軟銀', '驚魂', '交割股', '新電郵', '數量', '醣', '通常', 'nb'], dtype='<U11')
```

訓練模型

台積電

模型 Random Forest

文章集(4895)

- 漲(2694)
- 跌(2201)

StratifiedKFold(n_splits=3, random_state=0, shuffle=False)					
	precision	recall	f1-score	support	
Down	1.00	1.00	1.00	734	
Up	1.00	1.00	1.00	898	
micro avg	1.00	1.00	1.00	1632	
macro avg	1.00	1.00	1.00	1632	
weighted avg	1.00	1.00	1.00	1632	
Accuracy = 1.0					
	precision	recall	f1-score	support	
Down	1.00	1.00	1.00	734	
Up	1.00	1.00	1.00	898	
micro avg	1.00	1.00	1.00	1632	
macro avg	1.00	1.00	1.00	1632	
weighted avg	1.00	1.00	1.00	1632	
Accuracy = 1.0					
	precision	recall	f1-score	support	
Down	1.00	1.00	1.00	733	
Up	1.00	1.00	1.00	898	
micro avg	1.00	1.00	1.00	1631	
macro avg	1.00	1.00	1.00	1631	
weighted avg	1.00	1.00	1.00	1631	
Accuracy = 1.0					

訓練模型

鴻海

模型
Random Forest

文章集(3884)
• 漲(2200)
• 跌(1684)

StratifiedKFold(n_splits=3, random_state=0, shuffle=False)					
	precision	recall	f1-score	support	
Down	1.00	1.00	1.00	562	
Up	1.00	1.00	1.00	734	
micro avg	1.00	1.00	1.00	1296	
macro avg	1.00	1.00	1.00	1296	
weighted avg	1.00	1.00	1.00	1296	
0.9984567901234568					
	precision	recall	f1-score	support	
Down	1.00	1.00	1.00	561	
Up	1.00	1.00	1.00	733	
micro avg	1.00	1.00	1.00	1294	
macro avg	1.00	1.00	1.00	1294	
weighted avg	1.00	1.00	1.00	1294	
1.0					
	precision	recall	f1-score	support	
Down	1.00	1.00	1.00	561	
Up	1.00	1.00	1.00	733	
micro avg	1.00	1.00	1.00	1294	
macro avg	1.00	1.00	1.00	1294	
weighted avg	1.00	1.00	1.00	1294	
0.999227202472952					

訓練模型

大立光

模型 Random Forest

文章集(3653)

- 漲(1946)
- 跌(1707)

StratifiedKFold(n_splits=3, random_state=0, shuffle=False)					
	precision	recall	f1-score	support	
Down	1.00	1.00	1.00	569	
Up	1.00	1.00	1.00	649	
micro avg	1.00	1.00	1.00	1218	
macro avg	1.00	1.00	1.00	1218	
weighted avg	1.00	1.00	1.00	1218	
Accuracy 1.0					
	precision	recall	f1-score	support	
Down	1.00	1.00	1.00	569	
Up	1.00	1.00	1.00	649	
micro avg	1.00	1.00	1.00	1218	
macro avg	1.00	1.00	1.00	1218	
weighted avg	1.00	1.00	1.00	1218	
Accuracy 1.0					
	precision	recall	f1-score	support	
Down	1.00	1.00	1.00	569	
Up	1.00	1.00	1.00	648	
micro avg	1.00	1.00	1.00	1217	
macro avg	1.00	1.00	1.00	1217	
weighted avg	1.00	1.00	1.00	1217	
Accuracy 1.0					

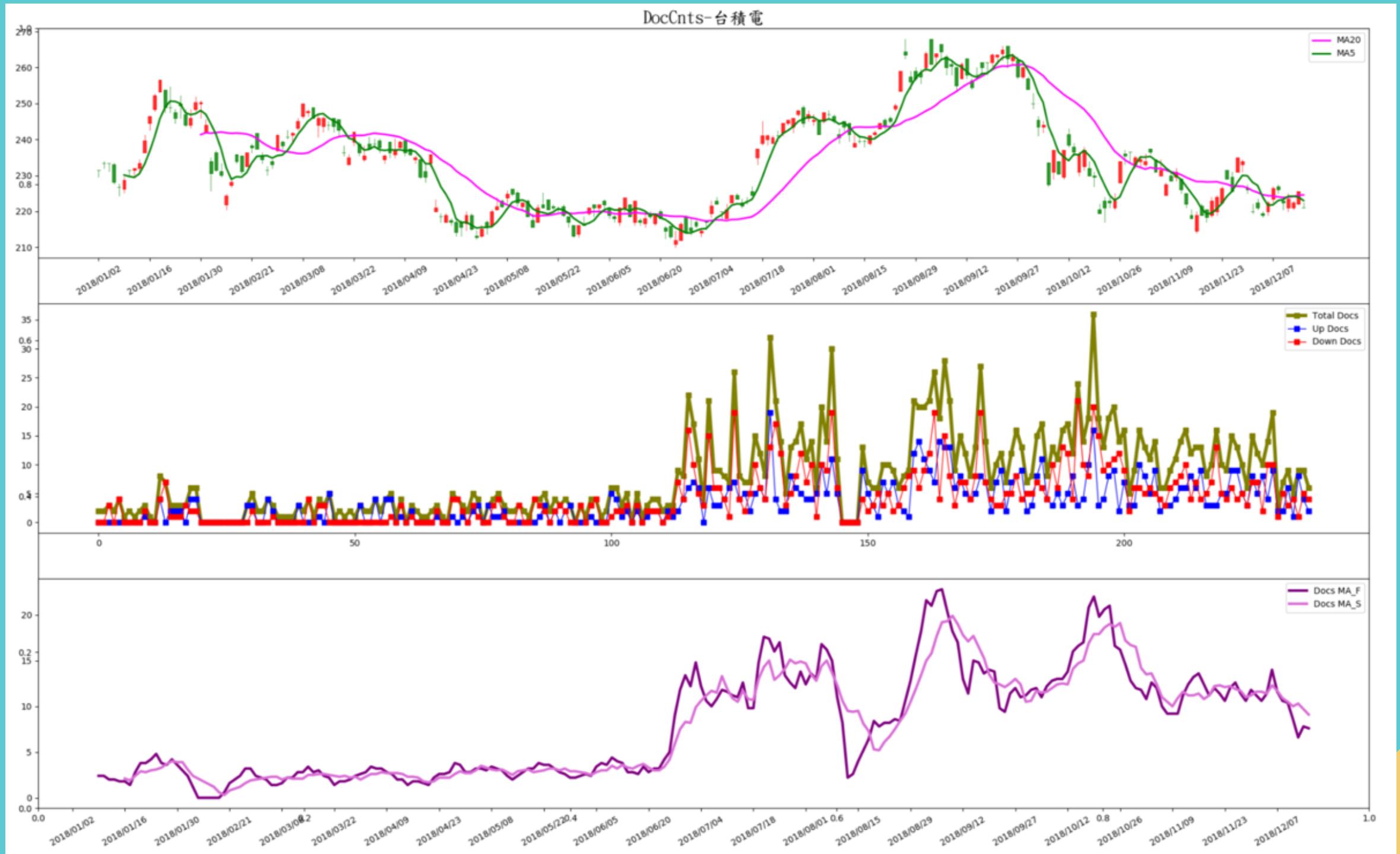
事件偵測

- 當日文章數超過5篇
- 快速平均由下穿過慢速平均
- 快速平均超過慢速平均(0.4)篇
- 個股變數 預測天數 n
篇數平均速度當天

出手頻率 台積電

有偵測到事件便是出手時機

```
Test Date= [Timestamp('2018-01-21 00:00:00')], DocCnt=8(U 5 D 3 P 0.40), Prediction=1, Acutual=1.0
Test Date= [Timestamp('2018-02-01 00:00:00')], DocCnt=6(U 4 D 2 P 0.50), Prediction=1, Acutual=1.0
Test Date= [Timestamp('2018-06-08 00:00:00')], DocCnt=6(U 4 D 2 P 0.50), Prediction=1, Acutual=0.0
Test Date= [Timestamp('2018-07-15 00:00:00')], DocCnt=7(U 4 D 3 P 0.25), Prediction=1, Acutual=1.0
Test Date= [Timestamp('2018-07-22 00:00:00')], DocCnt=32(U 16 D 16 P 0.00), Prediction=0, Acutual=1.0
Test Date= [Timestamp('2018-08-05 00:00:00')], DocCnt=20(U 6 D 14 P 0.57), Prediction=0, Acutual=0.0
Test Date= [Timestamp('2018-08-23 00:00:00')], DocCnt=10(U 7 D 3 P 0.57), Prediction=1, Acutual=1.0
Test Date= [Timestamp('2018-09-22 00:00:00')], DocCnt=10(U 6 D 4 P 0.33), Prediction=1, Acutual=1.0
Test Date= [Timestamp('2018-10-04 00:00:00')], DocCnt=8(U 2 D 6 P 0.67), Prediction=0, Acutual=0.0
Test Date= [Timestamp('2018-10-08 00:00:00')], DocCnt=13(U 4 D 9 P 0.56), Prediction=0, Acutual=0.0
Test Date= [Timestamp('2018-11-17 00:00:00')], DocCnt=12(U 9 D 3 P 0.67), Prediction=1, Acutual=0.0
Test Date= [Timestamp('2018-11-30 00:00:00')], DocCnt=13(U 9 D 4 P 0.56), Prediction=1, Acutual=1.0
Test Date= [Timestamp('2018-12-03 00:00:00')], DocCnt=15(U 8 D 7 P 0.12), Prediction=1, Acutual=1.0
```



MA FAST LEN=5
MA SLOW LEN=10

預測結果

台積電



N=3

MA FAST LEN=5

MA SLOW LEN=10

出手次數=13

		precision	recall	f1-score	support
DOWN		0.75	0.60	0.67	5
	UP	0.78	0.88	0.82	8
micro avg		0.77	0.77	0.77	13
macro avg		0.76	0.74	0.75	13
weighted avg		0.77	0.77	0.76	13
台積電 Tested accuracy = 0.7692307692307693					

預測結果

鴻海



N = 4

MA FAST LEN = 5

MA SLOW LEN = 15

出手次數 = 8

		precision	recall	f1-score	support
DOWN		1.00	0.80	0.89	5
	UP	0.75	1.00	0.86	3
micro avg		0.88	0.88	0.88	8
	macro avg	0.88	0.90	0.87	8
	weighted avg	0.91	0.88	0.88	8
鴻海 Tested accuracy = 0.875					

預測結果

大立光



N=3

MA FAST LEN=2

MA SLOW LEN=5

出手次數=8

		precision	recall	f1-score	support
DOWN		0.60	1.00	0.75	3
	UP	1.00	0.60	0.75	5
micro avg		0.75	0.75	0.75	8
macro avg		0.80	0.80	0.75	8
weighted avg		0.85	0.75	0.75	8
大立光 Tested accuracy = 0.75					



END

BDA GROUP 12