

# AI 及社群數據 V.S. 投資決策

## 第7組

資管三 杜沛慈 | 經濟三 胡南圳 | 圖資四 凌麗  
圖資四 林奕萱 | 會研所 陳詩婷 | 會研所 江泓葦



# INDEX

- ▶ 1. 選股
- ▶ 2. 資料前處理
- ▶ 3. 建構向量空間
- ▶ 4. 分類模型
- ▶ 5. 資料回測
- ▶ 6. 觀察與結論



1

# 選股



# 選股



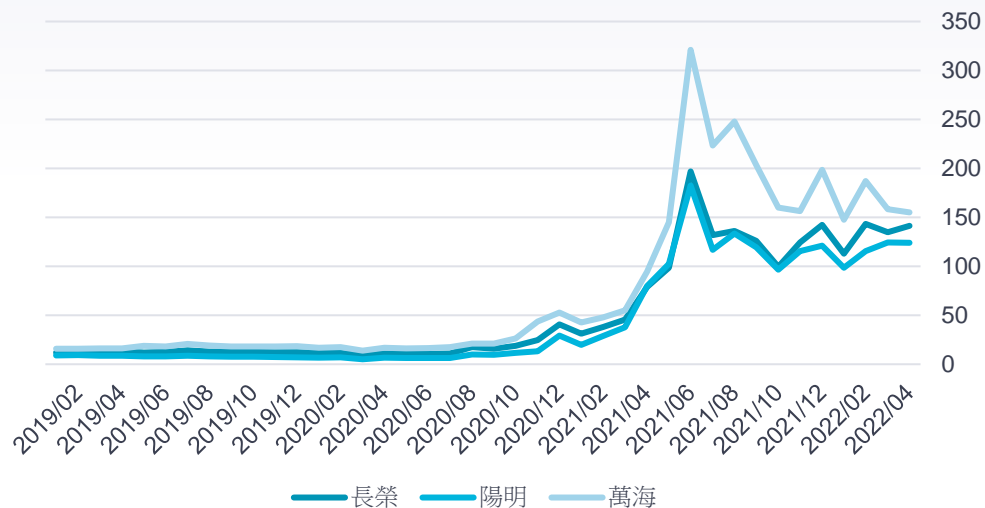
分析個股:長榮 ( 2603 ) 、陽明 ( 2609 ) 及萬海 ( 2615 ) 個別預測

資料集 : 2019 ~ 2021 年 PTT 文章資料集

# 個股分析

股票名稱	長榮 ( 2603 )	陽明 ( 2609 )	萬海 ( 2615 )
個股文章數	<b>6,948</b>	<b>4,127</b>	<b>2,569</b>
成交量 (百萬股)	<b>2,694</b>	<b>1,597</b>	<b>586</b>

2019~2021航運三雄股價變化



2

## 資料前處理



# 文章篩選 / label設定

挑出開盤日且  
PTT文章標題或內容有  
「長榮」、「陽明」、  
「萬海」的文章



合併文章標題  
和文章內容



使用  
CKIPTransformer  
斷詞

使用個股收盤價  
作為漲、跌、平  
判斷資料



調整漲跌幅  
判斷參數  
(5天 3%)



輸出檔案

# 漲跌幅判斷參數調整

間隔天數	漲跌%數	總出手率	總準確率	無資料月份數
3	3	0.9493	0.6068	7
3	5	0.9524	0.6786	13
3	7	0.9368	0.6854	15
3	10	1	0.7525	20
5	3	0.947	0.6493	1
5	5	0.9317	0.6021	8
5	7	1	0.698	14
5	10	1	0.7525	16
10	3	0.9448	0.5906	2
10	5	0.9628	0.6105	5
10	7	0.9544	0.6509	10
10	10	0.959	0.6631	13

- 市場上的消息(文章)應該會在短期內反映於股價變動上  
→  $n$  以3 / 5 / 10天做測試
- 天數間隔較短，所以將%數定為中小的股價波動率  
→  $\sigma$  以3 / 5 / 7 / 10%做測試



3

# 建構向量空間

結論2說明副標



# 資料處理

- 剔除平的資料，並重新設定文章編號

股票名稱	長榮 ( 2603 )	陽明 ( 2609 )	萬海 ( 2615 )
跌的文章數	<b>375</b>	<b>270</b>	<b>163</b>
漲的文章數	<b>525</b>	<b>412</b>	<b>242</b>

- 漲 = 1，跌 = 0

# TF-IDF + 卡方

- ▶ 將剩下資料去除非中文字詞、中文停用詞後，轉成TF-IDF，使用卡方建構2,000維的向量空間

	一下原則	一下去	一下子	一兩	一半	一半也	一度	一往	一律	一成	...	點收	點火	點燈	齊喊	齊挫	齊揚	齊殺	齊跌	龍哥	龍頭
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

4

# 分類模型

結論3說明副標



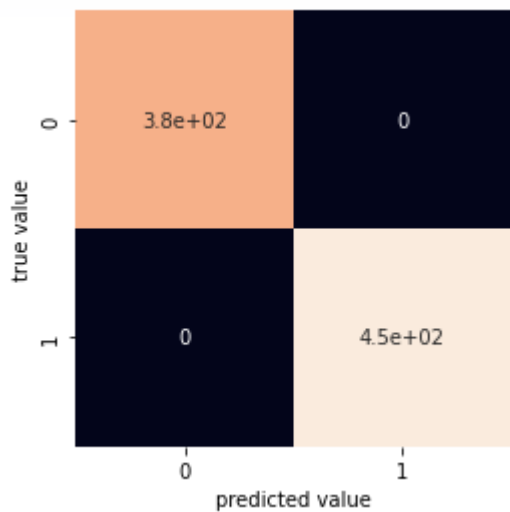
# 以SVM模型為例

- ▶ 將資料集隨機分為：80%訓練資料、20%測試資料
- ▶ 使用SVM (rbf) 模型分類  
( 另有嘗試隨機森林、XGBoost等模型，但結果相似 )
- ▶ 3支個股預測準確率皆為1.0

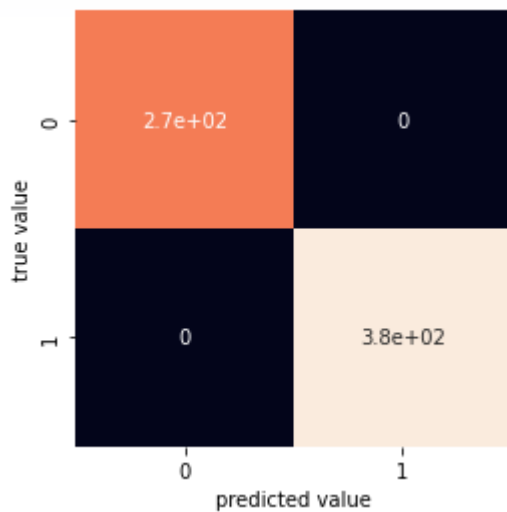
	precision	recall	f1-score	support
下跌	1.00	1.00	1.00	378
上漲	1.00	1.00	1.00	447
accuracy			1.00	825
macro avg	1.00	1.00	1.00	825
weighted avg	1.00	1.00	1.00	825

# 模型結果

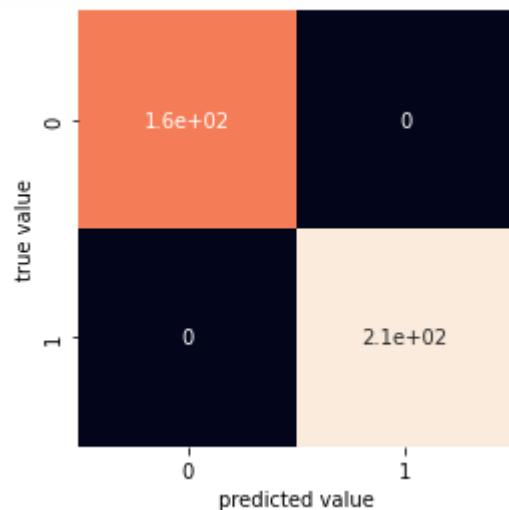
長榮



陽明

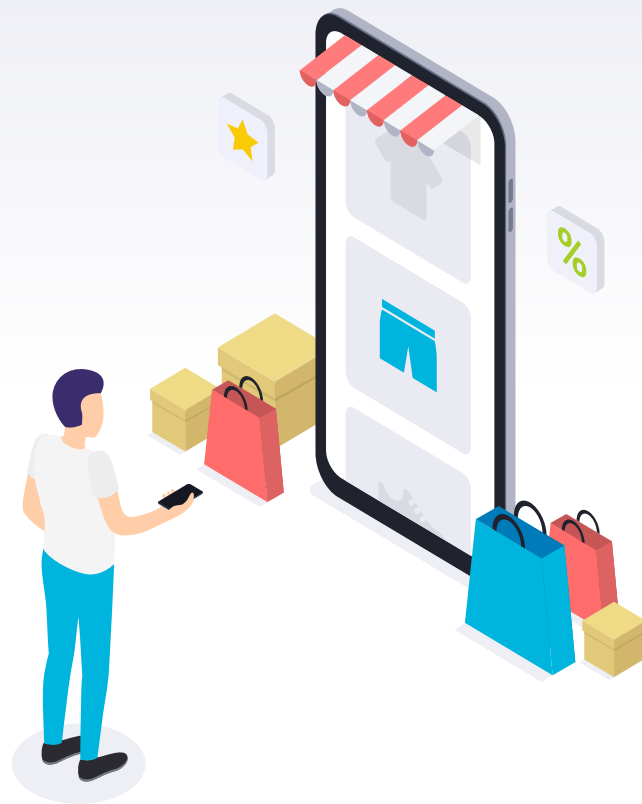


萬海



5

## 資料回測



# 資料回測



- ▶ 每6個月預測下一個月，再持續往後移動1個月，重複進訓練、預測
- ▶ 使用上階段的資料（剔除平） → 可能造成無訓練/測試資料
- ▶ 每6個月預測下一個月，再持續往後移動1個月，重複進訓練、預測
- ▶ 使用SVM模型 (另有使用XGBoost等模型，但出手率及準確率之綜效較差)
- ▶ 出手率判斷：

(依第D 日歸類為看漲或看跌的篇數，預測第D+n日為看漲或看跌，若篇數過於接近則不出手)

A. 出手：漲的文章數>60% 或 跌的文章數>60%

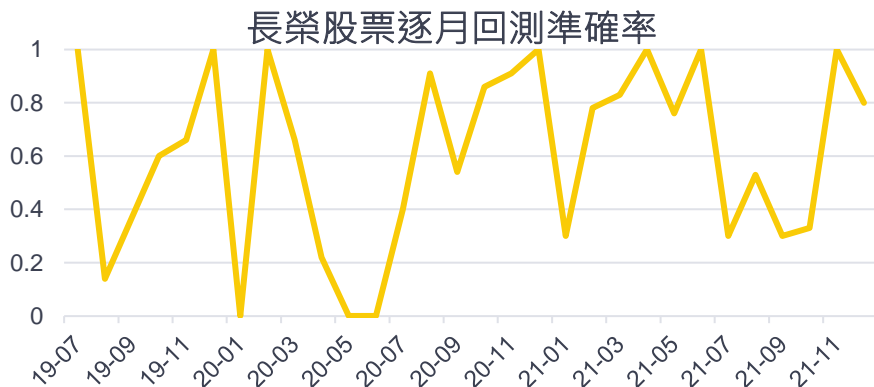
B. 其他→ 不出手



# 結果分析- 長榮



- ▶ 平均出手率大約為0.95，資料回測準確率約為0.65
- ▶ 準確率每月差異大



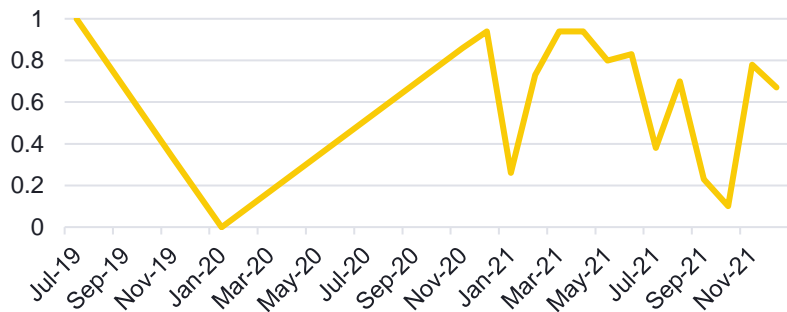
```
confusion_matrix :  
[[ 11  79]  
 [ 15 163]]
```

# 結果分析- 陽明



- ▶ 平均出手率大約為0.92，資料回測準確率約為0.65
- ▶ 文章數較少，因此較多月份無資料
- ▶ 準確率每月差異大

陽明股票逐月回測準確率



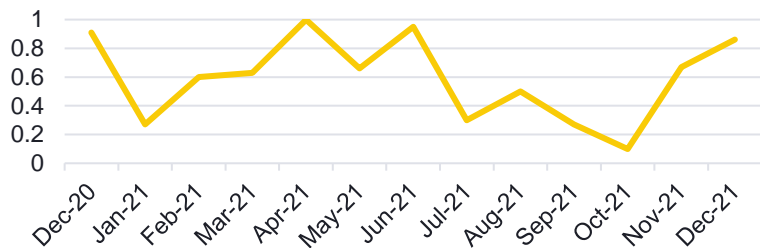
```
confusion_matrix :  
[[ 2 68]  
 [ 1 128]]
```

# 結果分析- 萬海



- ▶ 平均出手率大約為0.92，資料回測準確率約為0.59
- ▶ 文章數較少，因此較多月份無資料
- ▶ 準確率每月差異大


萬海股票逐月回測準確率



```
confusion_matrix :  
[[ 2 60]  
 [ 4 90]]
```

# 觀察與結論

使用3年的資料進行訓練，分類模型準確率可達1.0  
僅使用6個月的訓練資料做逐月回測，準確率僅0.65

- 
1. 訓練資料多寡大幅影響準確率
  2. 航運股遇到難得一見之股價飆漲，可能較不適合用歷史資料推估未來 (Requirement 2 是隨機抽取測試資料)

2021年後逐月回測準確率較佳



2021年航運股股價大幅攀升，連帶討論聲量、文章數大幅增加，使預測準確率較高

# 觀察與結論

漲的逐月回測準確率較佳

漲的文章較多，幾乎是跌的兩倍

股票名稱	長榮 ( 2603 )	陽明 ( 2609 )	萬海 ( 2615 )
跌的文章數	375	270	163
漲的文章數	525	412	242

# ▶ 說明影片

- ▶ 影片連結：<https://reurl.cc/ErKmqK>

# THANKS!

