大數據與商業分析 期中報告

股價漲跌關鍵字及模型預測

第17組

財金四 **黃伊華**　會計四 **張簡晴耘**　會計四 **周敏慈**　會計四 **鄭宇樺**　機械四 **戴柚誠**

EVA AIR

Outline

1. 選股
2. 資料前處理
3. 建構向量空間：N-gram / PCA
4. 分類模型：Linear SVM / Random Forest / Decision Tree / KNN / Logistic Regression / Naive Bayes
5. 資料回測
6. 補充：非監督式模型 & LLM

EVA AIR

1 選股

➢ **為台灣主要的航空公司之一，營運涵蓋多條國際航線**

產業背景清晰，具代表性與國際市場連動性

➢ **穩定的新聞討論度，文章數量適中**

有助於蒐集足夠語料，同時避免過多雜訊干擾

➢ **股價波動明顯，有助於標註文章的情緒傾向**

有利於模型從文本中學習情緒爛市場反應之關聯性

選擇 長榮航 2618 作為研究標的

EVA AIR

2 資料前處理

➢ **合併文章標題與內容，並移除標點符號/英文/數字**

減少非中文資訊對關鍵字提取與後續分析造成困擾

➢ **使用多階段關鍵字篩選，從 新聞 中挑出 6702 篇文章**

類別	關鍵字	說明
company keywords	長榮航、長榮航空、EVA Air、2618、張國煒	直接與長榮航空有關的字詞
impact keywords	股價、上漲、下跌、EPS、財報、營收、配息、除息、目標價、罷工、油價、匯率、疫情、邊境、解封、旅遊、免簽、航班、載客率、航線、戰爭、台海、空域	對股價可能造成影響的詞彙

EVA AIR

3 建構向量空間 — 標記文章

➢ **根據每則新聞發布日後 <u>最近三個交易日</u> 的股價變動進行加權評分**

計算 score 的方式：

(1) 若收盤價變動大於 ±3% 則標記為漲(1)或跌(-1)，其餘標記不出手(0)

(2) 再分別乘上權重：D+1天*0.5，D+2天*0.3，D+3天*0.2

➢ **根據評分將相關新聞自動加上漲跌標記，並分類為看漲 /看跌文章**

判斷score：

(1) > 0 標記為漲(1)

(2) < 0 標記為跌(-1)

(3) = 0 標記為不出手(0)

EVA AIR

3 建構向量空間 — 方案選擇

Options	Accuracy (based on Linear SVC)	
1. Bert + Fine-Tune	54.4%	❌
2. 合併漲/跌文章，以 tf-idf 建立向量空間	72.1%	❌
3. 將漲/跌文章分開，以多種統計 值建立向量空間	83.6%	✅

➢ **再分別嘗試後，決議以Option3作為最終模型**

EVA AIR

3 建構向量空間 — 關鍵字篩選

- **對兩批文章分別使用 2~4 gram 斷詞**
- **分別計算看漲/看跌文章的統計值**

詞	TF	DF	TF-IDF	全部 TF	全部 DF	全部 TF-IDF	TF 卡方值	DF 卡方值	MI	Lift	綜合排序分數

- **移除重複出現在看漲/看跌的詞**

判斷標準: [全部DF - 自己DF > 30] 的詞

- **以「綜合排序分數」分別排序看漲 /看跌關鍵字**

綜合排序分數 = 0.4 * DF卡方值 + 0.2 * TF-IDF + 0.2 * MI + 0.2 * Lift

EVA AIR

3 建構向量空間 — 分類結果

SVC	真實為漲	真實為跌
預測為漲	184	33
預測為跌	20	86

Accuracy: 83.59%

RF	真實為漲	真實為跌
預測為漲	183	41
預測為跌	21	78

Accuracy: 80.80%

DT	真實為漲	真實為跌
預測為漲	160	48
預測為跌	44	71

Accuracy: 71.52%

KNN	真實為漲	真實為跌
預測為漲	110	9
預測為跌	94	110

Accuracy: 68.11%

LR	真實為漲	真實為跌
預測為漲	188	34
預測為跌	16	85

Accuracy: 84.52%

NB	真實為漲	真實為跌
預測為漲	141	23
預測為跌	63	96

Accuracy: 73.37%

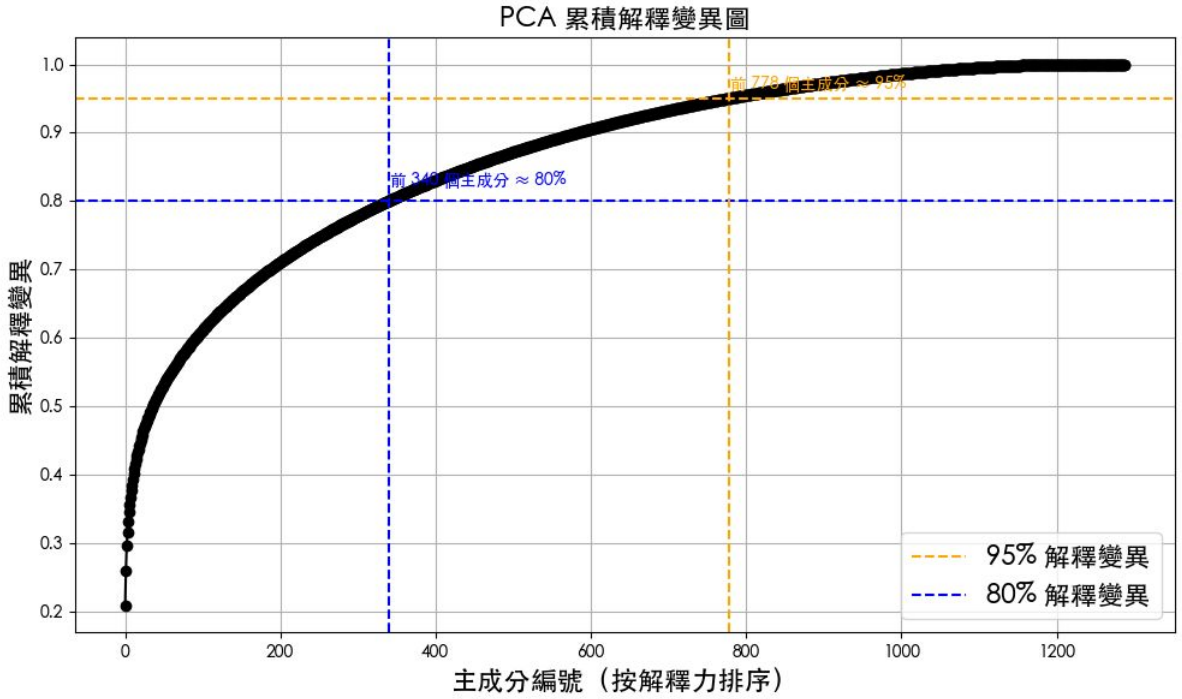
EVA AIR

3 建構向量空間 — PCA

➢ 以看漲/看跌關鍵字做主成份分析 (PCA)

(1) 找到解釋變異 95% 的主成份

(2) 從 90134 維降到 778 維



EVA AIR

3 建構向量空間 — PCA分類結果

SVC	真實為漲	真實為跌
預測為漲	179	37
預測為跌	25	82

Accuracy: 80.80%

RF	真實為漲	真實為跌
預測為漲	181	54
預測為跌	23	65

Accuracy: 76.16%

DT	真實為漲	真實為跌
預測為漲	142	53
預測為跌	62	66

Accuracy: 64.40%

KNN	真實為漲	真實為跌
預測為漲	193	94
預測為跌	11	25

Accuracy: 67.49%

LR	真實為漲	真實為跌
預測為漲	191	41
預測為跌	13	78

Accuracy: 83.28%

NB	真實為漲	真實為跌
預測為漲	-	-
預測為跌	-	-

PCA後有負值, 無法計算

EVA AIR

4 資料回測

➢ 逐日移動式訓練

使用LR模型，取第 D日前60天的資料預測第D +5日的漲跌

LR	真實為漲	真實為跌
預測為漲	528	127
預測為跌	178	208

➢ **Accuracy：70.70%**

EVA AIR

4 資料回測

➢ **加上出手篩選:若篇數差距小於當天文章 25% 則不出手**

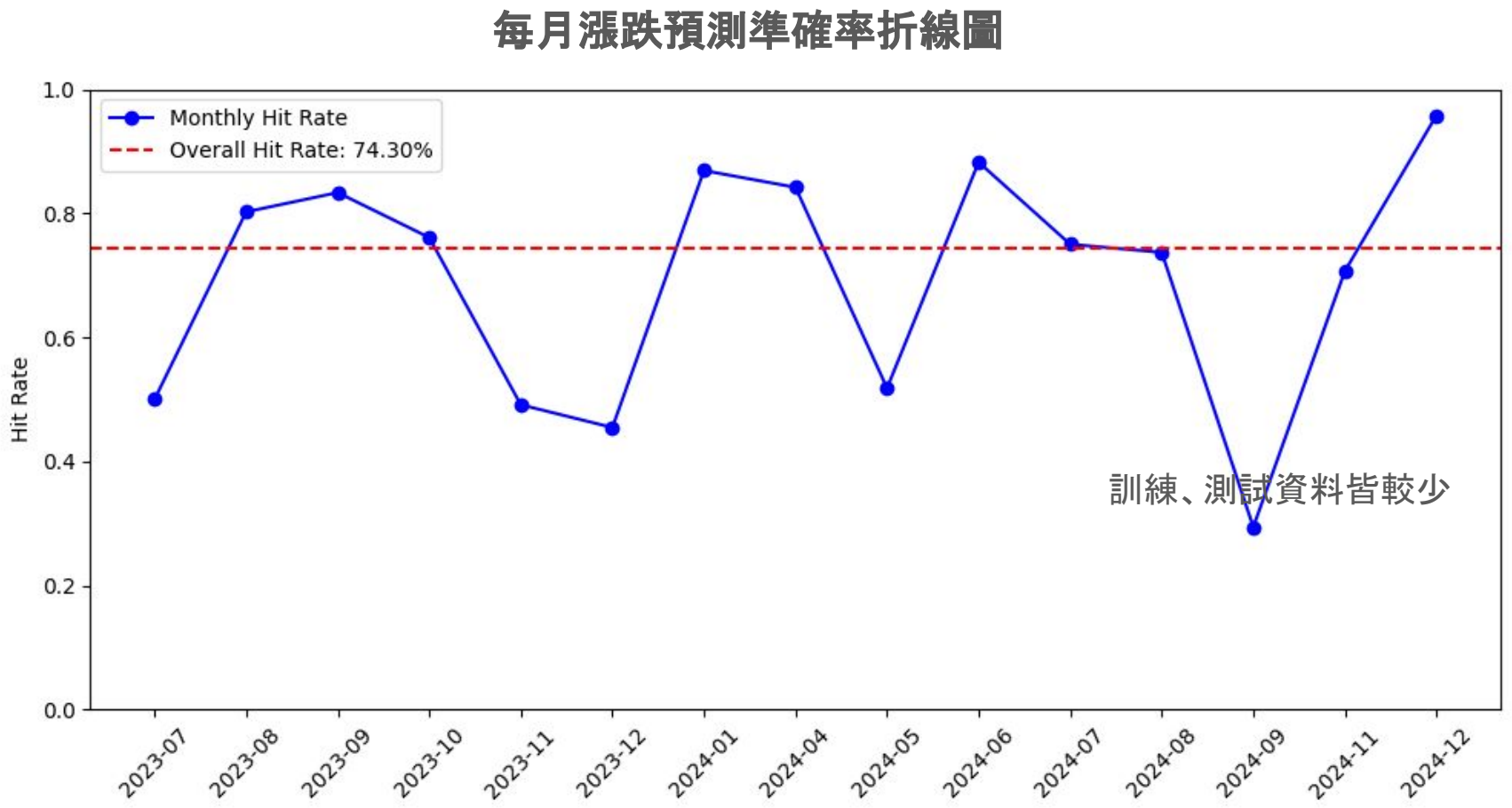
LR	真實為漲	真實為跌
預測為漲	458	107
預測為跌	112	175

➢ **出手率:81.84%**

➢ **Accuracy(出手時):74.30%**

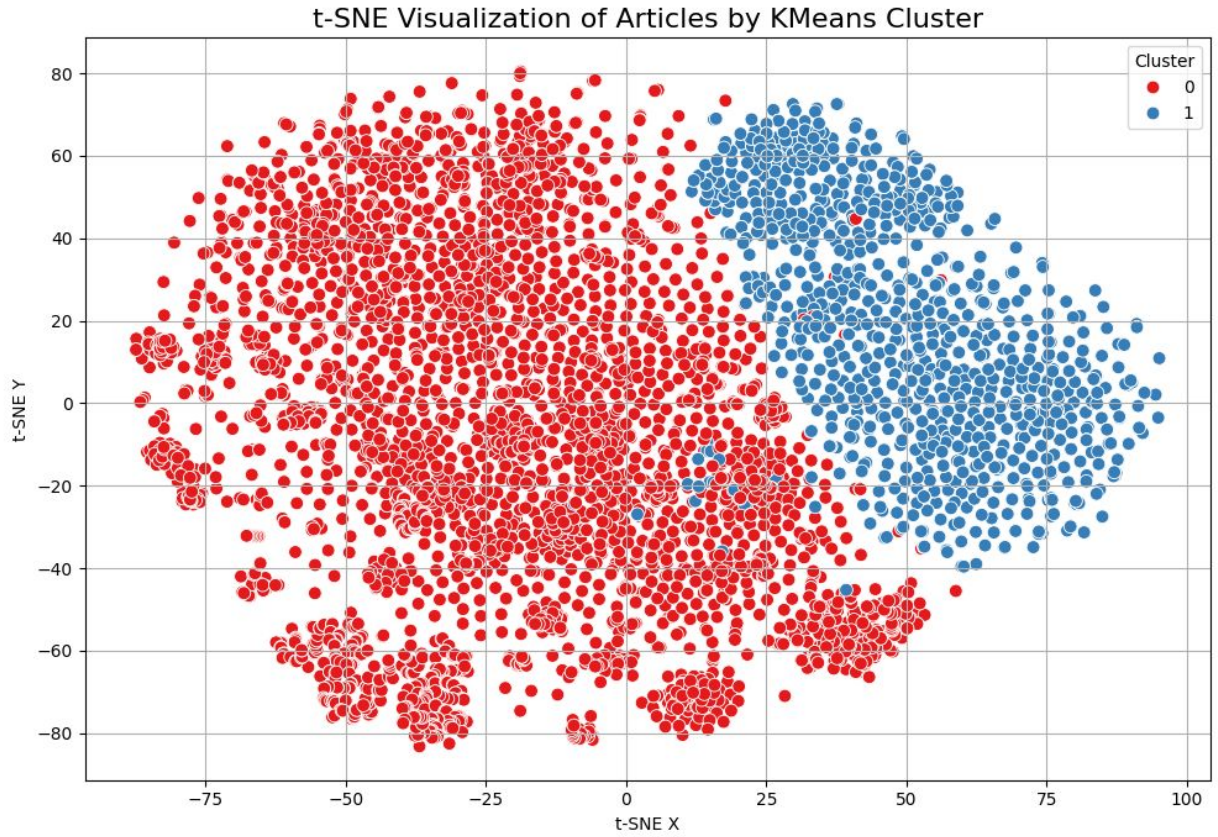
EVA AIR

4 資料回測


每月漲跌預測準確率折線圖
Monthly Hit Rate
Overall Hit Rate: 74.30%
Hit Rate
1.0
0.8
0.6
0.4
0.2
0.0
2023-07
2023-08
2023-09
2023-10
2023-11
2023-12
2024-01
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09
2024-11
2024-12
訓練、測試資料皆較少

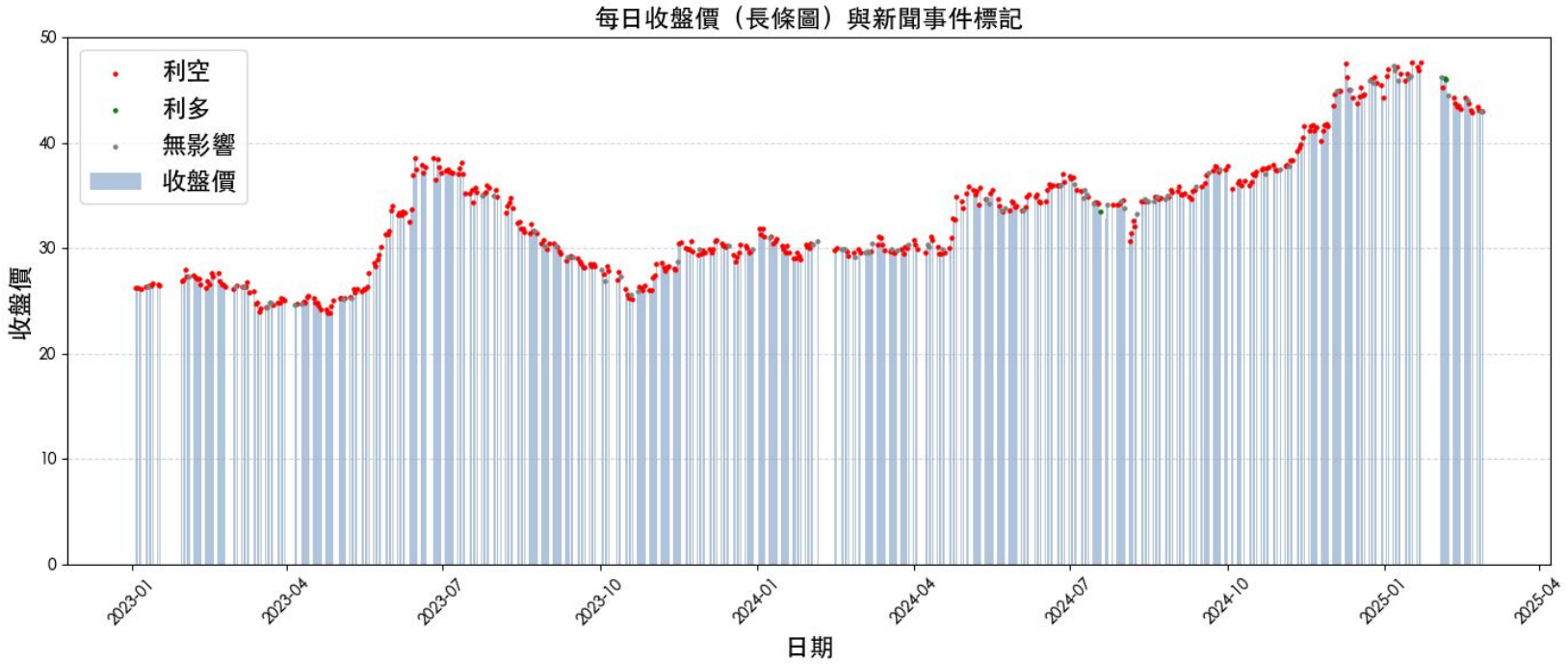
6 補充 — 非監督式學習

- **證明本組利用股價標記漲/跌文章的方式與非監督式學習下的結果相近**



6 補充 — LLM

- **LLM 依照「標題＋內文」來判斷文章漲跌**
- **結合收盤價、文章漲跌，畫出長條圖**



影片連結：https://youtu.be/AfNjAeUlfm4