大數據與商業分析_期中報告

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6組
B08701227 林姝廷、B08701244 蔡鉎驊、B08705017 陳煒勳
B08705036 朱修平、B08705042 董安宜、B08705046 高何銓

Agenda

Topics Covered

R1.
(1)選股
(2)資料處理與向量空間建構

R2.
(1)模型建立

R3.
(1)移動回測
(2)結果與討論

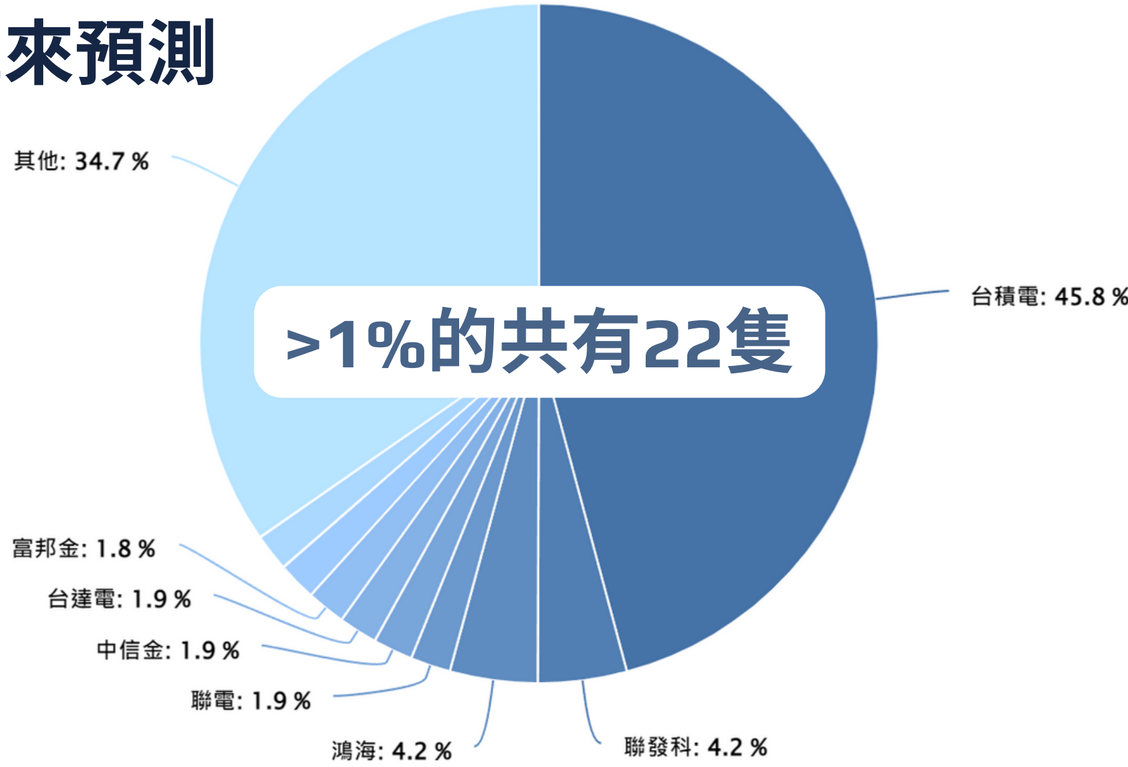
R1:各挑選出看漲及看跌的一批文章，從中取出關鍵字列表， 建構向量空間。

— 股價處理 & 文章資料處理

(1)選股

我們想用佔比較大的股票文章去預測 ETF 的漲跌

取元大台灣50 投資佔比 >1% 之股票的文章來預測



"2330 台積電", "2454 聯發科", "2317 鴻海", "2317 聯電",
"2891 中信金", "2308 台達電", "2881 富邦金", "2882 國泰金",
"1303 南亞", "1301 台塑", "2412 中華電", "2002 中鋼",
"2886 兆豐金", "2884 玉山金", "2603 長榮", "3711 日月光投控",
"5871 中租-KY", "2883 開發金", "1216 統一", "2885 元大金",
"2892 第一金", "5880 合庫金"

(2)資料處理

- 以2021年1月至2021年12月的文章資料來訓練一個分類器

- 以單日漲跌幅> 0.75 variation 標記為漲，< 0.25 variation 標記為跌

- 漲跌預測天數分別試隔天、隔三天 (test)
  並將第n天的文章與第 n+1 or n+3 的股市漲跌標籤合併

(2)資料處理

訓練集文章向量化處理 → 測試集文章向量化處理

- **文章向量化處理：**
  透過sklearn套件中TfidfVectorizer將斷詞結果去除停用詞後轉為空間向量

- **選擇對分類結果有較顯著影響的詞彙作為向量空間的維度，我們透過Chi-square計算各詞彙與漲跌標籤的獨立性作為選擇向量空間維度的依據**

選取特徵詞 -> SelectKBest(chi2, k = 調整)

R2:將兩批文章作為訓練資料及測試資料，使用監督式學習之分類演算法，評估分類模型之準確率。

— model 建立

建立預測模型

- 以2021年1月至2021年12月的文章資料來訓練一個分類器
  以 8 : 2 拆分訓練集及測試集

80%

訓練集

20%

測試集

- 文章只取第一部份被標記為漲和跌的文章

前25%

後25%

建立預測模型

TRY ：

- 漲跌預測天數分別是**隔天**、**隔三天**
- selectKBest分為**k=5000** 、**k=10000**
- 套入五種分類模型

建立預測模型

結果 (Accuracy) :

	隔天		隔三天
	k = 5000	k = 10000	k = 10000
GBC	0.677	0.645	0.637
RF	0.648	0.697	0.698
DT	0.644	0.619	0.633
SVM	0.632	0.670	0.690
KNN	0.569	0.593	0.6263

Gradient Boosting Classifier

Random Forest

Decision Tree

建立預測模型

(最後決定選擇隔三天, 10000詞彙)

	隔天		隔三天
	k = 5000	k = 10000	k = 10000
GBC	0.677	0.645	0.637
RF	0.648	0.697	0.698
DT	0.644	0.619	0.633
SVM	0.632	0.670	0.690
KNN	0.569	0.593	0.6263

Gradient Boosting Classifier

Random Forest

Decision Tree

預測天數隔三天
term10000

GBC	真實為漲	真實為跌
預測為漲	753	709
預測為跌	399	1195

RF	真實為漲	真實為跌
預測為漲	891	571
預測為跌	352	1242

DT	真實為漲	真實為跌
預測為漲	851	611
預測為跌	511	1083

SVM	真實為漲	真實為跌
預測為漲	881	581
預測為跌	367	1227

KNN	真實為漲	真實為跌
預測為漲	741	721
預測為跌	421	1173

(預測天數隔天term5000)

GBC	真實為漲	真實為跌
預測為漲	914	744
預測為跌	414	1515

RF	真實為漲	真實為跌
預測為漲	781	877
預測為跌	385	1544

DT	真實為漲	真實為跌
預測為漲	635	1023
預測為跌	253	1676

SVM	真實為漲	真實為跌
預測為漲	853	805
預測為跌	512	1417

KNN	真實為漲	真實為跌
預測為漲	972	686
預測為跌	860	1069

(預測天數隔天term10000)

GBC	真實為漲	真實為跌
預測為漲	855	803
預測為跌	469	1460

RF	真實為漲	真實為跌
預測為漲	938	720
預測為跌	367	1562

DT	真實為漲	真實為跌
預測為漲	880	778
預測為跌	590	1339

SVM	真實為漲	真實為跌
預測為漲	755	903
預測為跌	279	1650

KNN	真實為漲	真實為跌
預測為漲	738	920
預測為跌	540	1389

R3:判斷 n 日後指數或股價歸類為看漲或看跌，進行移動回測。

— 移動回測 & 預測結果

移動回測

前述選定的MODEL:
隔三天, 10000詞彙, Random Forest

- **2021年一整年，一次取三個月的資料作訓練集，第四個月為測試集**
- **一次往後移動 1 個月，重複進行訓練、預測**

結果

label vs predict label

round0	真實為漲	真實為跌
預測為漲	214	162
預測為跌	178	120

準確率：49.6%
出手率：83.3%

round1	真實為漲	真實為跌
預測為漲	476	780
預測為跌	462	738

準確率：49.2%
出手率：93.8%

round2	真實為漲	真實為跌
預測為漲	163	143
預測為跌	339	156

準確率：39.8%
出手率：60.0%

round3	真實為漲	真實為跌
預測為漲	253	463
預測為跌	306	433

準確率：47.3%
出手率：87.5%

round4	真實為漲	真實為跌
預測為漲	277	369
預測為跌	411	601

準確率：52.9%
出手率：87.5%

round5	真實為漲	真實為跌
預測為漲	13	99
預測為跌	81	446

準確率：71.8%
出手率：85.7%

round6	真實為漲	真實為跌
預測為漲	15	366
預測為跌	14	338

準確率：48.2%
出手率：88.9%

round7	真實為漲	真實為跌
預測為漲	23	470
預測為跌	26	356

準確率：43.3%
出手率：90.0%

round8	真實為漲	真實為跌
預測為漲	36	158
預測為跌	9	58

準確率：36.0%
出手率：80.0%

討論

移動回測：

- 訓練準確率都接近完美，但測試的結果很差
  --> overfit
- 應嘗試增加訓練資料以解決此問題
- 我們亦嘗試使用其他配置，像是更少的關鍵字、其他的訓練模型、以更多個月當作訓練資料，但結果還是不甚理想

討論

模型預測力沒有非常好的原因：

- **沒有讓不同公司用不同權重進行投票**
- **佔比近一半的台積電本身可能受台灣新聞的影響較小**
- **資料量太少**

Thank you!

影片連結:
https://youtu.be/dkcHvKhQ1zw