買點小幫手

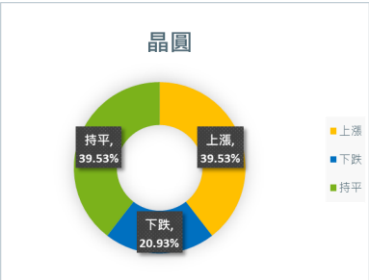
2018年2月9日 – 台積電(2330)

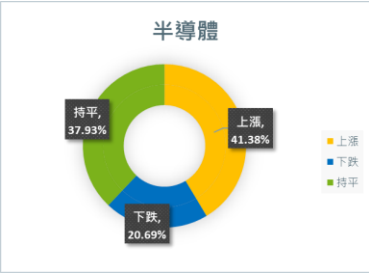
文字探勘方式進行台積電漲跌預測

產業: 半導體

漲跌關鍵字:

1. 指數和相關部分	2. 關鍵字
費城半導體指數	人工智慧
大盤	晶圓
S&P 500	7奈米
台積電ADR	半導體
APPL	蘋果





台積電(2330)：未來對AI的高速運算、物聯網、車用電子及挖礦機對繪圖晶片及特殊應用IC需求強烈，對晶圓代工的價值提升，台積電將會是對大受惠者。

- **研究目的**：將非結構化但即時的消息面(如該個股相關新聞或當日發布的重要消息)予以結構化，將能提升進退場時機的精確度，作為判斷該時點是否進場指標。

- **資料蒐集分兩大部分**：

1. 數據部分，採用 XQ 操盤高手，蒐集 2017 整年度台積電股價及美國標普 500 等的歷史股價資料。

2. 歷史新聞，以 Python 網路爬蟲擷取新聞網站資料，運用專業判斷製作詞庫，並使用開源套件 “jieba” 斷詞並算出字頻，最後選出關鍵字，統計各關鍵字出現的頻率。

- **資料萃取及清理**：結合數據及歷史新聞部分整理出一份資料集，彙整每一篇新聞出現的關鍵字及該段期間所發生的股價波動。在此，本團隊使用漲跌幅平均的上下一個標準差作區間，每篇新聞刊出後依該新聞發生的時間區間歸類為上漲、下跌及持平。特徵值及各文章字頻及目標值出來後，將數據正規化，並以 Python 的 Pandas 和 Rapidminer 做資料整理，將資料集空值填補為零。

- **數據分析的方法**：採 Python 及 Rapidminer 兩項工具進行文章漲跌預測。運用 Python 的 sickit-learn 機器學習套件以及當前 KDnuggets 上資料採礦最受歡迎的平台 Rapidminer 做各個演算法比較。
  演算法部分以 XGBoost 為主軸，並同時採用 SVM、Random Forest、GBDT 等演算法以 F1 score、Confusion Matrix 和 cross validation 做預測精確度比較。運用 XGBoost 演算法預測新聞文章漲跌的準度高達 80%以上。

買點小幫手

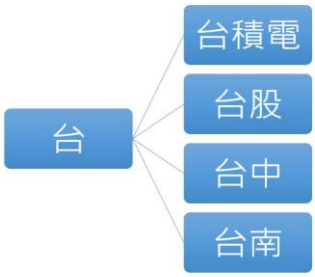
2018年2月9日 – 台積電(2330)

文字探勘方式進行台積電漲跌預測

- **XGBoost 演算法** : XGBoost 是 GBDT 的強化版，原理還是基於 GBDT。

XGBOOST	GBDT
CART、線性分類器	CART(gbtree)
一階、二階函數	一階函數
正規化項	
特徵抽樣(column subsampling)	
可以並行計算	無法並行計算
近似樹學習理論	貪心演算法

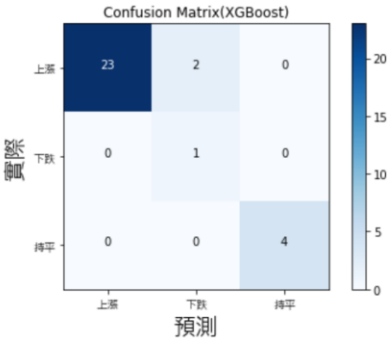
- **結巴演算法** : 基於前綴樹 (Trie Tree) 結構生成字詞



本團隊進一步從 XGBoost 裡面選取出具份量的特徵值做更深入分析，以視覺化方式呈現，並挖掘出有用資訊，可作為預測未來台積電股價漲跌的依據。

調整後的 R 平方越接近 1，台積電與該股價或指數波動程度相關性越高

美股指數部分，可以當作台積電股價的先行指標。



TSMC	大盤	S&P	費半	TSM	APPL
調整後的 R 平方	0.90	0.82	0.91	0.98	0.77

結語：以 2018 年 1 月之新聞做測試，準確度高於 93%，相信可以作為一個出色的輔助工具。本工具僅作為在投資人於基本面及技術面分析之後之輔助，不建議單純使用之便做出投資決策。

更多詳細內容請詳 Git Hub :
https://github.com/geniusbonny/TextMining-Stock

買點小幫手 Website : http://35.201.224.57/

報告內容僅供參考，不得作為任何投資引用之唯一依據，且其投資風險及決定應由投資人自行判斷並自負損益