以機器學習分析券商分公司交易資訊預測個股走勢

學生姓名：曾柏諺，指導教授：王義明老師
Department of Electrical Engineering, National Chi Nan University

摘要

籌碼面分析被普遍認為是市場數據分析方式中，有較強參考性的分析方法[1]，其中分公司交易資料(又名分點籌碼)隱含較多的市場資訊，但也是最難進行人工分析的一項籌碼數據。本研究使用機器學習分析3支個股由2017至今的券商分公司交易資料，預測10、20個交易日後的股價漲跌情況，以作為投資交易參考，並學習在跨領域研究中使用機器學習對連續性分布的數據進行分析與預測的方法。本模型的預測準確度最高為35%，並未超過基於三大法人交易資訊進行的機器學習分析之63%準確度[2]，推測為機器學習之訓練資料庫建立方式有誤。

一、研究動機

籌碼面資訊包含外資、投信、自營商的交易紀錄，其中自營商包含券商、大散戶、小散戶與公司內部人。而Ippolito (1989)[3]研究發現，以及SEC(美國證管會)表示，知悉關鍵市場資訊的交易人，能夠獲得異常的市場報酬，該類交易人多為特定本土散戶(又名關鍵分點、地緣分點)。投資人若能跟隨該類交易人進行布局，便能降低投資風險並提升獲利機率。但由於他們幾乎每天都會有交易，可能會連續多日少量的買入，並混雜部分的賣出，或使用股票匯撥後賣出以製造仍然持有該股的假象，或頻繁進行鉅額隔日沖銷(無法預測出波段漲跌)，且每支個股的關鍵/地緣分點均不同。若能將機器學習套用在分點交易資料，製作成一個客觀的方法對其進行定量分析，應能成為相關投行的程式交易參考工具。

二、系統架構

相似於一般程式交易系統，包含資料取得、分析、交易、修正。

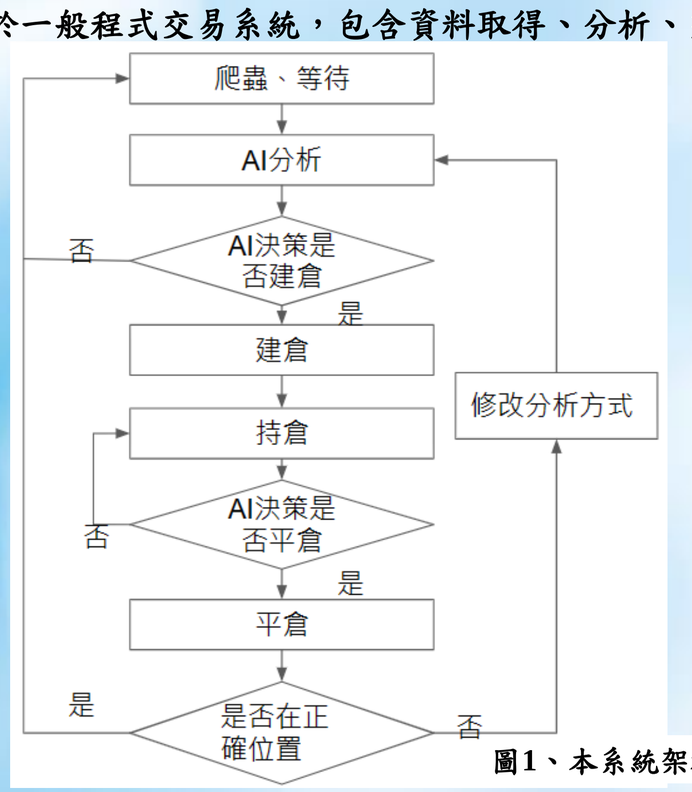

圖1、本系統架構流程圖

三、研究方法

本專題使用python爬取可成、台達電、宏遠，共3支個股，源於證交所公開的券商分公司交易資料，資料時間由2017/7/6至2023/4/19，並以最近60筆資料做為為準確度檢驗資料庫。由於資料較少，故聯合並編號多間券商分公司交易資料共同組成約1萬5千筆的訓練資料庫。

一段趨勢通常在10~20個交易日達到峰/谷值[4]，為提升交易準確度，將10天後的價格訂為標竿，若未來10天會上漲超過10%則訂為買入(答案值=10)；下跌超過10%定為賣出(答案值=-10)；並避免盤整過程對預測造成不利影響[5]，漲跌10%以內則不交易(答案值=0)。目標是要知道今天的「答案值」應是多少，預測未來漲跌。訓練資料之答案值與股價之關係如圖2所示。

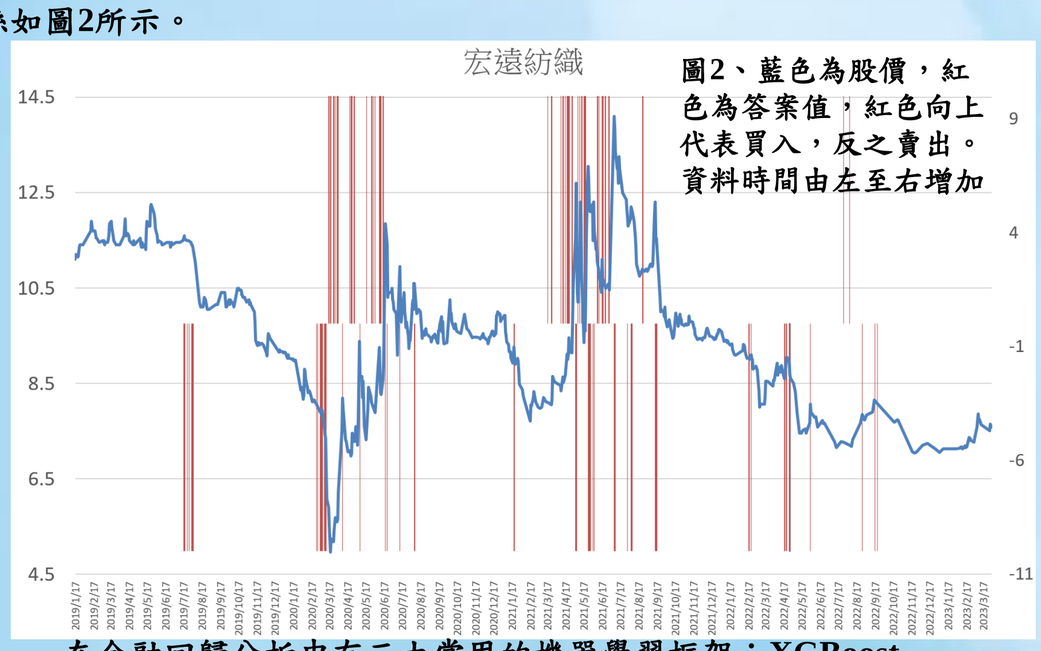

圖2、藍色為股價，紅色為答案值，紅色向上代表買入，反之賣出。資料時間由左至右增加

在金融回歸分析中有三大常用的機器學習框架：XGBoost, lightGBM, CatBoost，其中lightGBM由於有著最好的準確度與速度[6]，故本專題使用之。經反覆測試，其神經生長參數除了特徵使用率設為0.2、學習速率設為0.1，有最好的預測成果，其餘均使用預設值。 再使用近30個交易日的預測結果與其「答案值」進行貼合度驗證，由於要將準確程度與其他研究進行相比較，驗證函數使用具備正規化效果的R-square。

R-square公式說明：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

yi = 真實值(本研究中的±10)
ŷ = 預測值
ȳ = 平均值

$R^2 < 0$ 代表誤差劇烈
$R^2 = 0$ 代表零相關
$R^2 = 1$ 代表完全正確

四、實作成果

表1、特殊分點檢測

檢驗地緣分點	R2	檢驗關鍵大戶分點	R2
可成+5個台南分點	-0.04	宏遠+5個大戶	0.02
可成+5個台中分點	-0.11	宏遠+15個大/散戶	0.03
可成+5個台北分點	-0.14	宏遠+40個大/散戶	0.02
可成+15個分點	-0.01	台達電+5個大戶	-0.04
		台達電+15個大/散戶	0.01

表2、簡易濾網效果

濾網測試者	濾網前R2	濾網後R2
宏遠+5個大戶	0.02	-0.03
宏遠+15個大/散戶	0.03	0.17
宏遠+40個大/散戶	0.02	0.35

*此處大戶定義為日成交金額逾2億之分公司，而散戶(/小戶)為2億以下

實測結果如表1，可以發現預測準確度R^2不甚理想，因為良好的訓練模型R^2分數應大於 +0.5。但是仍可發現，資料量越多，預測結果相對越準確一些(R^2分數會高一些)。並且以宏遠這個股本最小的公司的分點交易資料具有較佳的預測準確度，故對其加入簡易濾網，移除結果中的眾數及將預測結果值拉開差距(將眾數設為閾值，其餘開根號*5)之後，得到表2，並將預測結果之累計損益繪製如圖3。

圖3中，資料時間由左至右，藍色為宏遠紡織股價，紅色是依據「答案值」進行交易的累計損益，綠色是依據預測結果進行交易的累計損益。

由圖3中也可以發現，預測損益(綠)低於答案值之理想損益(紅)，但在資金最大回撤(由高點回落)程度，預測損益有較好的表現。

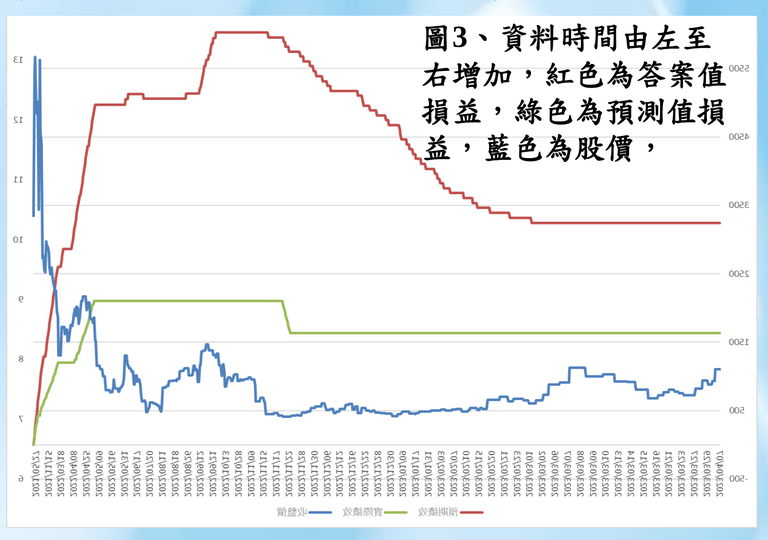

圖3、資料時間由左至右增加，紅色為答案值損益，綠色為預測值損益，藍色為股價，

五、結論與改進空間

使用lightGBM機器學習框架分析券商分公司於可成、台達電、宏遠之個股交易資訊，在預測10天後漲跌情況方面，加入簡易濾網以後，相對「理想答案」的準確度最高為宏遠的35%，準確度低於基於三大法人進行的機器學習分析之63%[2]。此外，預測之累計損益亦低於理想累計損益，但就損益回徹程度而言，優於「理想答案」。

儘管準確度較低，仍可致力於少數的獲利交易中取得高報酬以抵銷其餘小額虧損交易，故若將此分析套用於現貨，再依據預測結果交易具虧損抑制能力的可轉債、可轉債選擇權或具不對稱損益槓桿特性的權證便能將此特性凸顯，應能呈現更好的績效回報。

研究成果的R^2分數甚低，且未能呈現關鍵分點、地緣分點的成效，推測使用的機器學習資料分析方法在以下方面應有錯誤，因製作時間有限，未能及時解決改正：

1. 儘管有對每個分公司做編號，仍應將同日期的訓練資料合成1筆、多個欄位，因其中摻雜許多應被忽略的無用資料，若拆成多筆進行訓練，該類無用資料的影響力會擴大，對訓練結果產生負面作用，但這麼做會讓資料量大幅縮減至2千筆以下，因此要在避免過擬合方面多加著墨。 換句話說，每個分公司呈現的績效不同，績效好的散戶可以被列為正指標、績效差的散戶可被列為反指標，而無明顯績效的應被去除，現行做法可能會使績效差、績效好、無績效的分公司交易資料排列至相同水準進行機器學習，造成彼此干擾。但就算能夠人工挑選績效好、績效差的進行訓練，或者將多筆績效好壞的資料都合成1筆，讓模型自行判斷，資料量都會大幅縮水，不利於機器學習。故應設計不易過擬合的模型進行訓練。
2. 「答案值」的制定方法應做修正，因圖3中，答案值損益曲線回撤程度過於激烈，並非理想答案會有的現象。(監督式學習的標籤方法不正確)
3. 預測結果的濾網應使用更具參考性的制定。
4. 資料庫內所有的資料都應正規化後再行訓練，且對於無交易之日也應納入訓練資料。

參考資料

[1]張維碩、張智淵、張書豪 (2018)，「以向量自我迴歸模式探討台灣50成分股報酬率與技術面及籌碼面之關聯性」，全球商業經營管理學報，第10期，177-187。

[2]吳彥璋（2019）。機器學習與程式交易在建構交易系統之研究-以台灣指數期貨為例。〔碩士論文。輔仁大學〕臺灣博碩士論文知識加值系統。

[3] Ippolito, R. A., 1989, Efficiency with Costly Information: A Study of Mutual Fund Performance, 1965-84, Quarterly Journal of Economics, 104, 1-23.

[4]林哲緯（2022）。台灣資本市場證券公司分點之主力大戶的預測能力研究。〔碩士論文。國立中正大學〕臺灣博碩士論文知識加值系統。

[5]曾振楠（2010）。多頭、空頭與盤整市場台股指數價格調整係數之比較。〔碩士論文。國立臺灣大學〕臺灣博碩士論文知識加值系統。

[6] Al Daoud E (2019). Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset. International Journal of Computer and Information Engineering.