國立政治大學風險管理與保險學系

碩士學位論文

利用集成學習預測台灣加權股價指數漲跌

Applying Ensemble Learning to Enhance TAIEX Trend Prediction

指導教授：黃泓智 博士

研究生：陳羿妘 撰

中華民國 一一〇年六月

DOI:10.6814/NCCU202100893

摘要

本文旨在利用台灣加權股價指數 TAIEX 衍生之技術指標預測未來市場漲跌趨勢，藉由集成學習方法提升整體機器學習預測效果，結合羅吉斯迴歸、隨機森林、支持向量機三個異質演算法，增加模型間之差異性，並依據個別模型的特性，採用不同變數挑選方式，以提升資料品質，最終以單一模型作為標竿模型比較預測成效。整體而言，集成學習後之預測結果較單一模型具有更高的準確度，特別針對預測漲的部分，集成學習的效果較顯著，此外在長天期的趨勢預測中，集成學習的效果也更加明顯。

關鍵字：集成學習、羅吉斯迴歸、隨機森林、支持向量機、台灣加權股價指數、股價趨勢預測

DOI:10.6814/NCCU202100893

Abstract

This study aims to enhance prediction of trends on TAIEX with ensemble learning. As the input, several technical indicators are selected to train the model. To increase diversity of ensemble model, we used three heterogeneous models (logistic regression, random forest, support vector machine) instead of homogeneous models as component learners. Besides, depends on characteristic of component learners, different methods of feature selection are applied to increase the quality of data. To evaluate performance of ensemble models, we used single classifier models as benchmark models, and we found that accuracy of ensemble models is higher than single models. Especially in long-term case, the improvement of ensemble learning is more significant.

Keywords: Ensemble Learning, Logistic Regression, Random Forest, Support Vector Machine, TAIEX, Stock Trend Prediction

DOI:10.6814/NCCU202100893

目錄

DOI:10.6814/NCCU202100893

圖表目錄

DOI:10.6814/NCCU202100893

表格目錄

DOI:10.6814/NCCU202100893

第一章　緒論

第一節 研究動機與背景

觀察國人之理財習慣，除了透過定期儲蓄、保險ㄙ商品等保守的理財方式外，多數人也選擇透過專業經理人代操基金或是進入股票市場以賺取較高的報酬，然而高報酬投資同時伴隨高風險，因此如何洞察股市整體動向，提前預測變動趨勢為所有投資人關心之議題。

隨著科技發展，電腦硬體設備的突破，數據處理能力上有了極大的改善，同時促進了機器學習（Machine Learning, ML）的發展，機器學習的概念於1980 年代形成，為一門結合資訊與統計的學科，起初機器學習的理論以統計學與機率學為主，後來應用於資訊科學上，並逐漸受到重視。過去人們解決問題時，總希望可以透過規則解釋問題，進而提出解決辦法，而機器學習則顛覆了這樣的概念，透過電腦自行尋找資料規律，再進一步找出解答，藉由電腦算力的提升，過去無法處理的大數據（Big Data），也一一的被應用在機器學習上，提升整體學習效能，現今機器學習的概念已廣泛運用在醫學診斷、影像處理、資料探勘等領域上。

其中，機器學習的技術亦大量被運用在股價預測上，例如：針對投資組合配置，利用機器學習挑選未來可能上漲之標的，形成最小化風險、利潤最大化

DOI:10.6814/NCCU202100893

之投資組合；亦可預測大盤股市的趨勢，捕捉股市漲跌時機，以建構最佳化之投資組合。

機器學習與傳統解決方法差異在於，傳統方法透過規則解決問題，而機器學習則透過演算法引導電腦解決問題的思考邏輯，一般而言依據資料型態，主要可以分為三種學習方式：監督式、非監督式與強化式學習。

（一） 監督式學習（Supervised Learning）：具有目標變數，所有的資料用於預測標記值，例如：迴歸分析（預測連續值）、羅吉斯迴歸（預測類別）。

（二） 非監督式學習（Unsupervised Learning）：無目標變數，透過機器學習尋找資料規律與特性，例如：資料分群。

（三） 強化式學習（Reinforcement Learning）：無資料，由機器學習中自行摸索，透過環境中的反應，以獲取最大利益，例如：動態控制。

一般而言，機器學習是透過是極大化或極小化目標函數的方式解決問題，而常用的目標函數為損失函數，依據不同的學習方式以及資料特性，使用不同的計算方式，例如：監督式學習中，目標變數為連續值的迴歸分析，多使用均方誤差（Mean Square Error，MSE）作為損失函數，計算預測結果與實際情形的均方誤差，藉由最小化目標函數，最佳化模型預測成效；而分類問題則常用交叉熵（Cross Entropy）作為損失函數，藉由極小化損失函數，訓練最佳模

DOI:10.6814/NCCU202100893

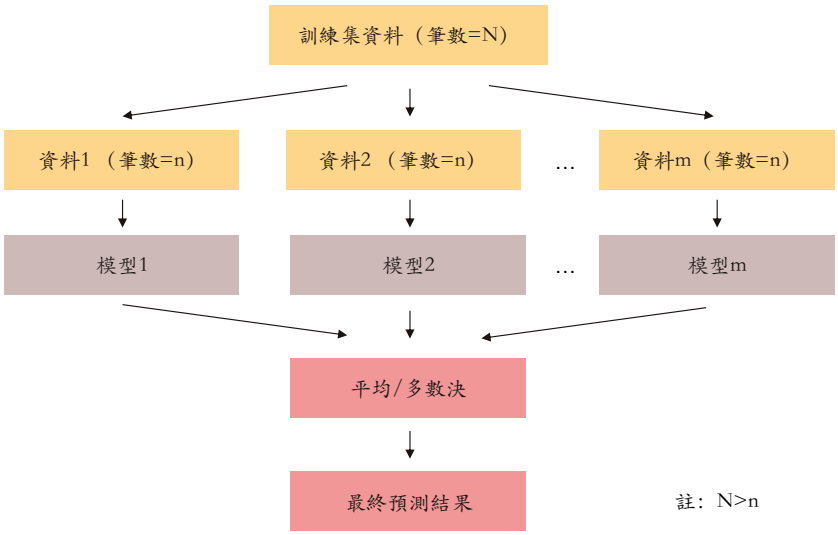
型；非監督式學習中，群集分析（K-means）目標則為最小化群內資料與群心的誤差平方和，藉此將資料分群。

然而透過單一模型訓練，常會遇到模型預測結果不穩定的現象，因此集成學習概念就此誕生，藉由整合不同模型預測結果，以及個別模型的差異，促進集成模型的多樣性，降低整體模型的變異程度（Variance）或偏差（Bias），避免單一模型之預測偏差影響最終預測結果，進而提升最終預測效果，依據組成方式不同，主要方法有三類：

（一） Bagging：

為 Bootstrap Aggregation 的縮寫，Leo Breiman 於 1994 年提出，建構多個學習器，模型間彼此獨立，並將原有資料重新以抽取後放回的方式採樣，同時訓練多個模型，再將各個模型預測結果以多數決或平均方式，整合為最終集成模型預測結果，常見的模型有：隨機森林。

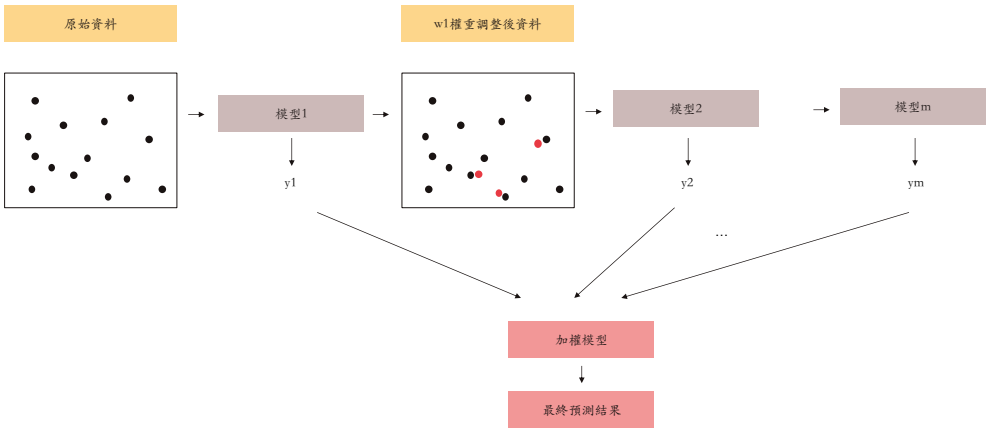
DOI:10.6814/NCCU202100893



圖表 1 Bagging 模型訓練流程

（二） Boosting：

由 Yoav Freund, Robert E. Schapire 於 1996 年提出，結合多個學習器，並以順序性方式訓練各個模型，每一次訓練中紀錄兩個值：模型權重與資料權重，依據每一筆資料的難易程度給予資料權重，以更新訓練資料，加強訓練容易錯誤之資料，並用於下一個模型訓練之中，最終利用模型權重（下圖之 y）將模型組合集成模型，預測最終結果，常見模型有：AdaBoost、Gradient Boosting、XGBoost。



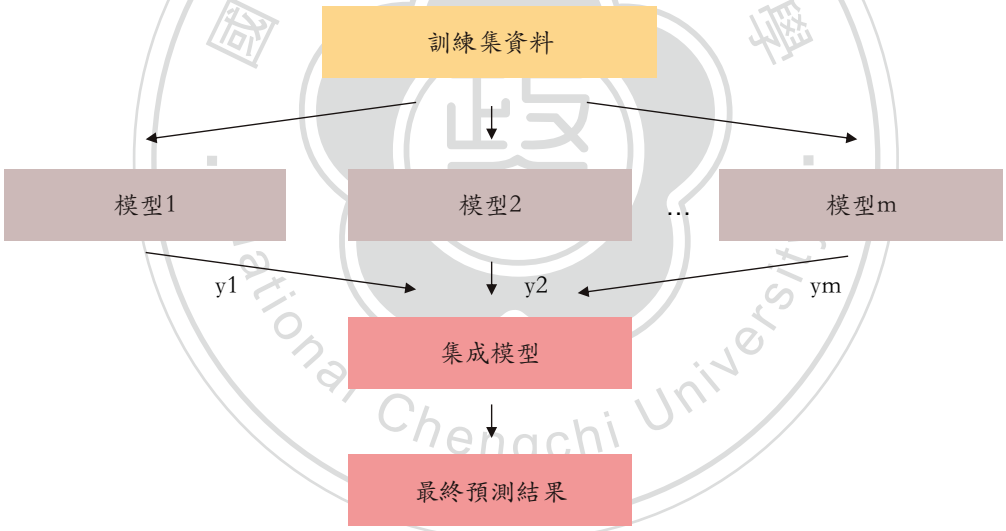
圖表 2 Boosting 模型訓練流程

DOI:10.6814/NCCU202100893

（三） Stacking：

相較於前述兩者，最終使用平均、投票或加權方式整合最終預測結果，Stacking 將集成學習分為兩步驟：Stacking 及 Blending。

1. Stacking：訓練多個模型，並將個別模型預測結果以平均或投票方式形成 Meta-Data。
2. Blending：建立整合模型（Meta Model, Blender, Ensemble Model），以 Meta-Data 作為特徵值訓練模型，並預測最終結果。



圖表 3 Stacking 模型訓練流程

DOI:10.6814/NCCU202100893

第二節 研究目的

國人對於股票投資之需求量高，然而投資市場上資訊龐雜，難以單一因素判斷未來漲跌趨勢，另一方面，隨科技發展，機器學習技術已廣泛運用在各個領域，包含影像處理、語言辨識、投資市場等，因此本研究旨在利用機器學習之技術，回應投資人對於股票投資之需求，利用台灣股價加權指數相關指標，建立集成學習模型，預測未來一段期間內之股市漲跌趨勢，以尋找股市買進時機點。

本研究欲達成以下之研究目的:

(一) 結合機器學習之技術，建立羅吉斯迴歸、隨機森林、支持向量機模型，用於預測整體股市漲跌趨勢，以滿足投資人對於股票投資之需求。

(二) 探討集成學習之成效，藉由集成個別模型之預測結果，提升單一模型之預測準確率。

第三節 研究流程

依據上述之研究動機與目的，本文使用台灣加權股價指標之漲跌趨勢作為預測目標，研究集成學習對於股價趨勢預測之成效，並分為以下五個章節，個別敘述如下:

一、緒論

DOI:10.6814/NCCU202100893

檢視國人投資理財之需求以及機器學習發展背景，並闡述本文之研究動機及目的。

二、文獻探討

回顧本研究相關文獻資料，包含資料前處理、特徵值挑選與機器學習三大部分。

三、研究方法

本章節中分為五個小節，說明本研究使用之方法，包含整體研究架構、資料處理方式、特徵值挑選、單一模型與整體集成模型建構流程。

四、實證結果

以單一模型作為標竿模型，比較集成學習模型與個別單一模型之預測準確率差異，分析集成學習之效果。

五、結論與建議

就集成學習於股價預測上之應用，包含模型架構及交易策略建構兩面向，提出共五點建議，供未來相關研究參考。

圖表 4 研究流程圖

DOI:10.6814/NCCU202100893

第二章 文獻探討

本文使用技術指標預測大盤股市漲跌趨勢，將資料預先處理過後，挑選適當特徵值，以避免資料品質過低導致機器學習無效，接著使用集成學習概念，進行多個模型漲跌預測訓練，因此本章節中會分為資料預處理、特徵值挑選與機器學習模型三部分進行文獻回顧:

第一節 資料預處理

股票是變動程度最高、反應最迅速且風險最高的投資工具之一，短期內股價的噪點較多，容易影響資料品質，造成機器學習效率低落，因此如何去除資料雜訊，取得適當的歷史股價趨勢，是奠定整體模型成敗的基石。

Basak, Kar, Saha, Khaidem and Dey (2019)將大盤股價指數資料，透過指數平滑方式，去除短期資訊噪點後用於預測未來股價漲跌趨勢。而 Jiang, Liu, J., Zhang and Liu, C. (2019)亦將 S&P 500、道瓊工業指數、那斯達克指數資料進行指數平滑轉換，以避免短期股價波動影響模型對於長期趨勢的預測，並將平滑後的結果用於預測 20 天後的大盤指數漲跌。

除了透過平滑轉換修正股價短期趨勢外，過往有少部分學者使用時間 序列資料斜率作為特徵值，近期使用過類似方法的論文包含：Mierswa and Morik (2005)使用具時序性的音訊資料分類音樂風格、Moews, Herrmann and Ibikunle

DOI:10.6814/NCCU202100893

(2019)利用股價資料預測未來趨勢，皆是使用線性迴歸計算時間區間內之斜率，以擷取期間內之資料趨勢作為特徵值。

第二節 特徵值挑選

股市漲跌變動成因複雜，難以單一變數捕捉，去除短期波動後，市場中仍然隱含大量雜訊，因此特徵值的選擇為預測漲跌變動之關鍵，過往不乏學者將價與量進行計算分析，藉由長時間的股價觀察，透過動量、趨勢、資金等面向預測未來趨勢，2010 年 Larsen 僅使用技術分析與 k 線便擊敗大盤股價，然而隨著技術演變，獲取即時市場資訊不再是難事，意味著市場資訊愈加龐雜，然而多維度的資料，大幅降低模型訓練效率，同時伴隨過擬合現象的發生，因此維度下降（Dimension Reduction）與特徵值挑選（Feature Selection）逐漸受到重視，主成分分析(Principle Component Analysis)、Boruta 演算法也開始受到廣泛應用，其中 Boruta 演算法於 2010 年被 Kursa and Rudnicki 提出並以 R 語言實踐，此後，Naik and Mohan 於 2019 年將 Boruta 演算法，應用於個股股價趨勢預測，藉由篩選出具重要性之變數，提升模型預測效果。

第三節 機器學習方法

單單使用技術分析，無法充分反映龐雜的市場資訊，配合市場對於大數據資料運算的需求，因此投資人開始嘗試將機器學習引入股價趨勢預測中，其中常見的演算法包含以下:

DOI:10.6814/NCCU202100893

（一） 羅吉斯迴歸:

Dutta, Bandopadhyay and Sengupta（2012）挑選印度股市中交易頻率最高的 30 間公司，以基本面資料作為特徵值，使用羅吉斯迴歸模型判斷公司未來年度股價表現是否高於市場報酬，並達到 7 成以上之準確度。

Li, H., Yang, Z. and Li, T.(2014)使用 2000 年至 2014 年 NASDAQ 與 NYSE 之股價日資料，以計算五日後之漲跌趨勢，並利用股價外資訊，包含：能源價格、黃金價格等，訓練羅吉斯迴歸模型進行二元分類，預測五日後股價漲跌趨勢，並達到 55.65%之準確率。

（二） 隨機森林:

Patel, Shah, Thakkar and Kotecha 於 2015 年利用不同的變數型態作為特徵值，將 10 個技術指標分為連續型與類別型兩套，應用於四類演算法：神經網路、支持向量機、隨機森林、樸素貝葉斯（Naive-Bayes）建立模型預測隔日股價的漲跌趨勢，以兩個股價指數 NIFTY 50 與 BSE-Sensex，及兩個公司 Infosys 與 Reliance 的股價作為預測目標，並使用 2003 年到 2012 年共 10 年之日資料建立模型與回測績效，連續型資料以隨機森林模型之表現最優，平均達到八成以上之準確率；類別型資料則以樸素貝葉斯模型為最優，平均準確率為九成。

DOI:10.6814/NCCU202100893

（三） 支持向量機:

2014 年 Di 收集 84 個技術指標後，利用極限隨機樹演算法挑選 3 成特徵值作為支持向量機模型變數，以預測 Apple、Amazon、Microsoft 三家公司的股價漲跌趨勢，並使用 2010 年 1 月至 2014 年 10 月近五年之資料，達到 7 成之準確率。

Żbikowski 於 2015 年使用交易量加權支持向量機分類器，以 Fisher 方法挑選技術指標作為特徵值訓練，建構股價交易策略，平均報酬率為 168.71%

（四） 集成學習（Ensemble Learning）：

Ballings, Van den Poel, Hespeels and Gryp（2015）使用歐洲個股資料進行實證分析，比較單一模型與集成學習模型的成效差異，並發現相較於單一模型如羅吉斯迴歸、支持向量機、集群分析以及神經網路，集成學習的隨機森林在 ROC (Receiver Operating Characteristic Curve)與 AUC(Area Under Curve)的表現上為所有模型中最佳。

DOI:10.6814/NCCU202100893

第三章 研究方法

第一節 研究架構

本研究使用 1990 年至 2020 年間台灣股價加權指數，共 30 年之日資料，並將收盤、開盤、最高、最低價格與單日成交量，使用指數平滑方式轉換原始資料，將平滑後之成交量與股價，計算技術指標及相關訊號指標，共 75 個變數作為機器學習模型之特徵值，接著藉由資料歸一化與斜率擷取的過程，轉換資料以降低特徵值單位不一致之影響並提升資料品質。

此外，為了提高模型間的差異性，以增加集成學習效率，就個別模型使用不同特徵值，本文一共使用三個模型與兩個變數挑選方式：羅吉斯迴歸使用邊際迴歸顯著程度挑選重要變數；隨機森林使用 Boruta 演算法挑選重要變數；支持向量機則保留原始的 75 個變數，接著透過集成學習賦予各模型相同權重，以多數決或三分類方式集成最終模型預測結果，並使用單一模型作為學習標竿，將前 25 年資料作為訓練集、最後 5 年資料作為測試集，比較單一模型與集成模型之訓練績效。

第二節 資料預處理

本文之預測目標為台灣加權股價指數，並以其股價與成交量衍生之技術指標作為模型訓練之特徵值，採用之資料型態如下：

DOI:10.6814/NCCU202100893

（一） 資料期間：1990 年 7 月 13 日至 2020 年 7 月 30 日，共計 30 年

（二） 資料來源：TEJ 台灣經濟新報資料庫

（三） 資料類型：台灣加權股價指數

資料品質高低為模型訓練成效的關鍵，最初收集到的原始資料，最可能遇到的問題為資料缺漏、噪點、名稱不一致等，因此在生成變數前，本研究先透過刪除資料點的方式將這些狀況排除，此外為了增進模型之訓練成效，同時透過變數生成及資料變換等方式，形成模型使用之特徵值，以下分為五部分，分別介紹本研究資料預處理之方法。

（一） 指數平滑：

本研究目的為預測大盤股價指數K日內之漲跌趨勢，以助於投資買進時機點之判斷，因此利用指數平滑 3 日之轉換，去除短期投資市場噪音，以取得長期股市價與量之趨勢，轉換公式如下：

$$S'_t = a \times S_t + (1-a) \times S'_{t-1}$$

$$a = \frac{2}{3+1}$$

其中，S_t為原始價量，S'_t為指數平滑後之價量。

（二） 生成技術指標：

本研究採用之變數包含：價量資訊、技術指標以及由技術指標衍生之訊號指標，共計 75 個變數，作為模型訓練之特徵值。利用過去市場價與量之走勢，計算出歷史股價趨勢水準以及股價變動速度，並藉由當前股價概

DOI:10.6814/NCCU202100893

況與歷史資料之比較，提前預測未來可能之趨勢變動，近一步捕捉股價交易策略進出場時機點，全文採用之變數可分為以下兩類別：

1. 台灣加權股價指數：開盤價、最高價、收盤價、最低價、當日成交量
2. 技術指標與訊號指標：共採用 15 種技術指標，簡單移動平均（Simple n-days Moving Average）、加權移動平均（Weighted moving average）、動量(Momentum)、平滑異同移動平均線（Moving Average Convergence & Divergence，MACD）、平均趨向指標（Average Directional Indicator，ADX）、順勢指標（Commodity Channel Index）、隨機指標（Stochastic Oscillator，KD）、相對強弱指數（Relative Strength Index, RSI）、布林通道指標（Bollinger Bands）、蔡金波動指標（Chaikin Volatility, CV）、阿隆指標（Aroon Indicator）、簡易波動指標（Ease of Movement Value, EMV）、資金流向指標(Money Flow Index , MFI)、成交量簡單移動平均（Volume Simple n-days Moving Average）、威廉指標（William's %R）。此外，配合不同技術指標特性，設定相應門檻值，以計算買進賣出的訊號指標，捕捉股價趨勢轉折之處，各指標詳細計算方式如表格 1 說明：

DOI:10.6814/NCCU202100893

表格 1 變數說明

指標	名字	描述	公式
SMA	簡單移動平均 (Simple n-days Moving Average)	計算特定期間內股價與成交量之平均值。	$\frac{C_t + C_{t-1} + \cdots C_{t-n+1}}{n}$ 收盤價C_t：n=2,3,4,5,6,10,20,60
VSMA			$\frac{V_t + V_{t-1} + \cdots V_{t-n+1}}{n}$ 成交量V_t：n=2,5,20,60
WMA	加權移動平均 (Weighted moving average)	計算特定期間內股價平均值，越近期的股價加權比重越高。	$\frac{C_t \times W_1 + \cdots C_{t-n+1} \times W_n}{n}$
MOM	動量 (Momentum)	計算特定期間之股價差異，MOM 由下往上突破 0 時為買進訊號，MOM 由上往下突破 0 時為賣出訊號。	$C_t - C_{t-n}$ n=2,3,5,10,20
MOM_Signal_u			$\begin{cases} 0, & otherwise \\ 1, & MOM_{t-1} < 0 \,\&\, MOM_t > 0 \end{cases}$
MOM_Signal_d			$\begin{cases} 0, & otherwise \\ 1, & MOM_{t-1} > 0 \,\&\, MOM_t < 0 \end{cases}$
fastK	隨機指標 (Stochastic Oscillator，KD)	計算目前股價與特定期間內高低價的相對位置，以預測股價逆	$\frac{C_t - LL_{t-(n-1),n}}{HH_{t-(n-1),n} - LL_{t-(n-1),n}} \times 100$ $LL_{t,n}$：過去 n 天最低價 $HH_{t,n}$：過去 n 天最高價

DOI:10.6814/NCCU202100893

fastD		轉時機，當 fastK 大於 fastD 的時候，表示處於上漲趨勢。	$\frac{fastK_t + fastK_{t-1} + fastK_{t-2}}{3}$
KD_Signal			$\begin{cases} 0, & otherwise \\ 1, & fastK_t > fastD_t \end{cases}$
ArUp	阿隆指標 (Aroon Indicator)	計算近期股價達到最高與最低點所經過之時間，以及兩者的期間差距，設定 ArUp 高於 ArDn 且 ArUp 大於 70 時為買進訊號。	$\frac{n - n_{tHH}}{n} \times 100$ n_{tHH}:過去 n 日內最高價距離 t 的天數
ArDn			$\frac{n - n_{tLL}}{n} \times 100$ n_{tLL}:過去 n 日內最低價距離 t 的天數 n=25
Oscillator			ArUp- ArDn
Aroon_Signal			$\begin{cases} 0, otherwise \\ 1, \text{ArUp} > \text{ArDn} \,\&\, \text{ArUp} > 70 \end{cases}$
Dn	布林通道指標 (Bollinger Bands)	計算 20 日簡單移動平均之標準差，加（減）兩倍標準差後為股價之上下限（壓力線與支撐線），並將當日股價轉換為百	$\text{SD}_t = \sqrt{\frac{\sum_0^{n-1}(C_{t-i} - MA_t)^2}{n}}$ $\text{Dn}_t = MA_t - 2 \times SD_t$ n=20
Up			$MA_t + 2 \times SD_t$
%B			$\frac{C_t - \text{Dn}_t}{Up_t - \text{Dn}_t}$

DOI:10.6814/NCCU202100893

Bbands_Signal		分比指標，當百分比指標大於 1 時，為買進訊號。	$\begin{cases}0, otherwise \\ 1, PctB_t > 1\end{cases}$
DIF	指數平滑異同移動平均線（Moving Average Convergence & Divergence）	計算長短期（26 日、12 日）股價之指數平均差異為 DIF，將差異值取 10 日指數平均為 MACD，兩者相減為柱線（OSC），OSC 由負轉正為買進訊號，由正轉負則為賣出訊號。	$DIF_t = EMA(C_t, 12)_t - EMA(C_t, 26)_t$
MACD			$MACD_t = \mathrm{EMA}(\mathrm{DIF}_t, 10)$
Macd_Signal1			$\mathrm{OSC_t} = \mathrm{DIF_t} - \mathrm{MACD_t}$ $\begin{cases}0, & otherwise \\ 1, OSC_t > 0 \ \& \ OSC_{t-1} < 0\end{cases}$
Macd_Signal2			$\begin{cases}0, & otherwise \\ 1, OSC_t < 0 \ \& \ OSC_{t-1} > 0\end{cases}$
DIp（positive）	平均趨勢指標（Average Directional Indicator, ADX）	依據當日與前一交易日之最高與最低價狀況，紀錄±DM(Directional Movement，動向變化值)，依前 14 天內之	L_t：第 t 日最低價 H_t：第 t 日最高價 +DM：$H_t > H_{t-1} \& L_t > L_{t-1}$ −DM：$H_t < H_{t-1} \& L_t < L_{t-1}$ $TR_t = max(H_t - L_t, H_t - L_{t-1}, C_{t-1} - L_t)$ +DM(14)：14 天內+DM 平均

DOI:10.6814/NCCU202100893

		±DM 個數與股價真實波幅（TR），計算趨向指數，當 DIp 向上穿越 DIn 且高於 ADX 時，為買進訊號，當 DIp 向下穿越 DIn 且低於 ADX 時，為買進訊號。	TR(14)：14 天內 TR 平均 $DIp_t = \frac{+\mathrm{DM}(14)}{TR(14)} \times 100$
DIn(negative)			−DM(14)：14 天內−DM 平均 $DIn_t = \frac{-\mathrm{DM}(14)}{TR(14)} \times 100$
DX			$\frac{\lvert DIp_t - DIn_t \rvert}{\lvert DIp_t + DIn_t \rvert} \times 100$
ADX			$\frac{\mathrm{DX_t} + \mathrm{DX_{t-1}} + \ldots \mathrm{DX_{t-13}}}{14}$
ADX_Signal1			$\begin{cases} 0, \text{otherwise} \\ 1, \mathrm{DIp} > \mathrm{DIn} \ \& \ \mathrm{DIp} > \mathrm{ADX} \ \& \\ \quad \mathrm{DIn} > \mathrm{ADX} \end{cases}$
ADX_Signal2			$\begin{cases} 0, \text{otherwise} \\ 1, \mathrm{DIp} < \mathrm{DIn} \ \& \ \mathrm{DIp} < \mathrm{ADX} \ \& \\ \quad \mathrm{DIn} < \mathrm{ADX} \end{cases}$
CCI	順勢指標（Commodity Channel Index）	計算最高、最低及收盤價之平均值為典型價格（Typical Price, TP），並計算 TP 之移動平均值為 MATP，將 MATP 與 TP 之差距絕對值加總平均為 MD，利用 TP、MATP 與 MD 計算 CCI 指標以觀察股	$\mathrm{TP_t} = \frac{\mathrm{H_t} + \mathrm{L_t} + \mathrm{C_t}}{3}$ $\mathrm{MATP_t} = \frac{\mathrm{TP_t} + \cdots \mathrm{TP_{t-n+1}}}{\mathrm{n}}$ $\mathrm{MD_t} = \frac{1}{n} \times (\lvert \mathrm{MA_t} - \mathrm{TP_t} \rvert + \cdots \lvert \mathrm{MA_{t-n+1}} - \mathrm{TP_{t-n+1}} \rvert)$ $\mathrm{CCI_t} = \frac{\mathrm{TP_t} - \mathrm{MATP_t}}{0.015 \times \mathrm{MD_t}}$ n=20
CCI_Signal1			$\begin{cases} 0, otherwise \\ 1, CCI_t < -100 \end{cases}$

DOI:10.6814/NCCU202100893

		市異常走勢，當 CCI 小於-100 為超賣指標。	
EMV	簡易波動指標（Ease of Movement Value, EMV)	計算最高、最低價之平均值為 MM，前一日與當日MM差距為移動終點 MID（Distance moved），並計算單位成交量 VPU（Box Ratio），兩者相除為每單位成交量之股價變動數 EMV，取移動平均後為 MAEMV，當 EMV 大於 0 時，為買進訊號，反之則為賣出訊號。	$MM_t = \frac{H_t + L_t}{2}$ $MID_t = MM_t - MM_{t-1}$ $VPU_t = \frac{V_t}{H_t - L_t}$ $EMV_t = \frac{MID_t}{VPU_t}$
MAEMV			$\frac{EMV_t + \cdots EMV_{t-n+1}}{n}$ n=9
EMV_Signal			$\begin{cases} 0, EMV_t < 0 \\ 1, EMV_t > 0 \end{cases}$
MFI			$MF_t = TP_t \times V_t$ ＋MF(14)：

DOI:10.6814/NCCU202100893

	資金流向指標 (Money Flow Index , MFI)	典型價格與成交量香城後為資金流量 (Money Flow, MF)，若當日典型價格較前一交易日高，將資金流量計為正資金流量，反之則計為負資金流量，計算 14 日內之正負資金流量比率 (Money Flow Ratio, MFR)，並求出 MFI，當 MFI 大於 80 為超買指標，MFI 小於 20 為超賣指標。	14 日內正資金流量總值 －MF(14): 14 日內負資金流量總值 $MFR_t = \frac{+MF(14)}{-MF(14)}$ $MFI_t = 100 - \frac{100}{1 + MFR_t}$
MFI_Signal1			$\begin{cases} 0, otherwise \\ 1, MFI_t > 80 \end{cases}$
MFI_Signal2			$\begin{cases} 0, otherwise \\ 1, MFI_t < 20 \end{cases}$
CV	蔡金波動指標 (Chaikin Volatility, CV)	計算高低價差之指數移動平均值，以衡量股價波動性 CV，當 CV 小於 0 時為買進訊號。	$REMA_t = EMA(H_t - L_t, 10)$ $CV_t = \frac{REMA_t - REMA_{t-10}}{REMA_{t-10}} \times 100$
CV_Signal			$\begin{cases} 0, otherwise \\ 1, \quad CV_t > 0 \end{cases}$

DOI:10.6814/NCCU202100893

RSI	相對強弱指數（Relative Strength Index, RSI）	計算前一交易日與今日收盤價上漲、下跌幅度，分別記為 U 與 D，並將 RS 壓縮至 0 與 100 之間，當 RSI 大於 80 為超買訊號，RSI 小於 20 則為超賣訊號。	$U_t = max(C_t - C_{t-1}, 0)$ $D_t = max(C_{t-1} - C_t, 0)$ $RS_t = \frac{EMA(U_t, 10)}{EMA(D_t, 10)}$ $RSI_t = 100 - \frac{100}{1 + RS_t}$
RSI_Signal1			$\begin{cases} 0, & otherwise \\ 1, & RSI_t > 80 \end{cases}$
RSI_Signal2			$\begin{cases} 0, & otherwise \\ 1, & RSI_t < 20 \end{cases}$
WPR	威廉指標（William's %R）	計算過去 10 天之最高價以比較當日收盤價與歷史高低價之差異。	$WPR_t = \frac{C_t - HH_{t,10}}{HH_{t,10} - LL_{t,10}} \times 100$
WAD	威廉多空力度線（Williams Accumulation / Distribution）	計算每日最高價與最低價的變化並忽略短期股價變化，以捕捉價格趨勢，當股價下跌且 AD 上漲時為買進訊號，當股價上漲	$TRL_t = min(C_{t-1}, L_t)$ $TRH_t = max(C_{t-1}, H_t)$ If $C_t > C_{t-1}$: $AD_t = C_t - TRL_t$ If $C_t < C_{t-1}$: $AD_t = C_t - TRH_t$ If $C_t = C_{t-1}$: $AD_t = 0$ $WAD_t = WAD_{t-1} + AD_t$
AD_Signal1			$\begin{cases} 0, & otherwise \\ 1, & C_t < C_{t-1} \& WAD_t > WAD_{t-1} \end{cases}$

DOI:10.6814/NCCU202100893

AD_Signal2		且 AD 下跌時為賣出訊號。	$\begin{cases} 0, & otherwise \\ 1, C_t > C_{t-1} \& WAD_t < WAD_{t-1} \end{cases}$

（三） 擷取斜率:

本研究的目標為漲跌趨勢預測，為了使特徵值與預測目標一致，因此使用特徵工程中擷取趨勢的方式轉換資料，過去有些研究中採用迴歸係數以獲取資料趨勢，本研究中為了簡化作法，直接以資料斜率取代，計算當期與前一期資料的變動數，以捕捉前後期股價變動趨勢，並去除股價高低對於預測漲跌趨勢之影響，計算公式如下:

$$x_t' = x_t - x_{t-1}$$

其中，x_t為該變數第 t 點資料之原始數值

（四） 變數歸一化:

機器學習時，不同單位的特徵值以及資料數值大小對模型建構都會造成影響，尤其是離群值的存在會大幅影響模型訓練結果，因此為了解決特徵值性質不一的問題，並增加特徵值間的可比性，變數歸一化是常用的資料轉換方式，透過訓練集中資料的最大及最小值將資料區間壓至-1 與 1 之間。

歸一化公式如下:

$$SD_t'' = \frac{x_t' - \min(x')}{\max(x') - \min(x')}$$

$$x_t'' = SD_t'' \times (1 - (-1)) + (-1)$$

DOI:10.6814/NCCU202100893

其中，x_t'為該變數第 t 點資料之跨期變動數，$\min(x')$為該變數之最小值，$\max(x)$為該變數之最大值。

（五） 目標變數訂定

本研究探討之目標為「對於第 t 筆觀測值，K 個交易日後的平滑後收盤價與今日平滑後收盤價相比是否上漲」，以公式表達如下：

$$Signal_t = \begin{cases} 0, & if \ \ EMA(C_t, 3) \geq EMA(C_{t+K}, 3) \\ 1, & if \ \ EMA(C_t, 3) < EMA(C_{t+K}, 3) \end{cases}$$

第三節 特徵值挑選

透過機器學習進行訓練時，除了學習器的挑選外，資料品質對於模型之成效亦具有極大影響力，當資料品質低落的時候，將造成「Garbage-in, garbage-out」這種無效學習的現象，因此特徵值的挑選格外重要。

本研究中，針對羅吉斯迴歸與隨機森林模型，使用羅吉斯迴歸邊際檢定與 Boruta 演算法作為特徵值挑選的方式，初步排除不重要的變數，以提升整體資料品質，並藉由使用不同特徵值挑選方法，增進集成學習模型多樣性，提升整體學習效果。

（一） 羅吉斯迴歸

建構羅吉斯迴歸，變數的挑選將影響整體模型的解釋力，因此進行模型訓練前，先針對每一個原始變數進行邊際檢定，並依據個別變數的顯著程度，保留 p-value 小於 0.2 的變數作為羅吉斯迴歸模型的特徵值。

DOI:10.6814/NCCU202100893

（二） 隨機森林

Boruta 演算法在 2010 年被 Kursa and Rudnicki 提出，將特徵值與目標變數間之關聯性打亂，產生與目標變數無關之特徵值資料，並重複訓練模型，以紀錄各個變數對於模型訓練的重要度，並依據隨機特徵值的重要度，將原始特徵值分為三類：重要、不確定、不重要。為了提升模型訓練效率與模型間之差異性，此處保留判定為重要及不確定之變數作為隨機森林模型之特徵值。

第四節 個別模型架構

本研究採用近似於 Bagging 之集成學習概念，同時使用三個演算法建立模型，為了提高模型間之差異性，分別使用三套不同變數訓練三個模型，避免集成學習效率不彰。羅吉斯迴歸與隨機森林模型使用的變數，分別透過變數顯著程度與 Boruta 演算法進行挑選，以下將分為兩個小節討論，第一部分介紹個別模型架構，第二部分介紹集成學習方法與建模流程。

（一） 邏輯斯迴歸（Logistic Regression）

迴歸分析模型主要是用來建立目標變數與解釋變數間的關係，其中邏輯斯迴歸專門處理目標變數為類別型資料的二元分類問題，不過由於目標變數為離散型資料，因此需要透過乙狀函數（Sigmoid Function）轉換目標變數至 0 到 1 之間的連續型態，並將特徵值與目標變數間以條件機率

DOI:10.6814/NCCU202100893

（Conditional Probability）與乙狀函數（Sigmoid Function）建立關係，公式如下：

$$p(X) = Pr(Y = 1|X = x) = \frac{e^{(\beta_0+\beta_1 X)}}{1 + e^{(\beta_0+\beta_1 X)}}$$

$Pr(Y = 1|X = x)$為目標變數 Y=1 的發生機率，可以用特徵值 x 的非線性函數解釋，經由簡化後，公式又可以轉換為特徵值 x 對數勝率（Odds）的線性函數：

$$ln(\frac{Pr(Y = 1|X = x)}{1 - Pr(Y = 1|X = x)}) = \beta_0 + \beta_1 X$$

$ln(\frac{Pr(Y=1|X=x)}{1-Pr(Y=1|X=x)})$形式又可稱為 Logit 或 Log-odds，其中$\frac{Pr(Y=1|X=x)}{1-Pr(Y=1|X=x)}$為事件 Y=1 發生之勝率(Odds)，即為事件 Y=1 發生機率除以事件 Y=1 不發生機率的比率。

最終使用最大概似估計法（Maximum Likelihood Estimation）估計模型係數β，即得邏輯斯迴歸模型。

（二） 隨機森林（Random Forest）

隨機森林是結合了 Breiman 的袋裝算法（Bootstrap aggregating，Bagging）與何天琴的隨機子空間理論（Random Subspace Method）所發展而成的集成學習模型，隨機森林內包含多個決策樹模型，依據目標變數為

DOI:10.6814/NCCU202100893

連續資料或類別資料，以多數決或加總平均將所有決策樹的預測結果，形成最終隨機森林之決策。

由於隨機森林內皆為同質性之決策樹模型，為了增加整體模型多樣性，隨機森林中包含兩項隨機因子：隨機抽取訓練樣本、隨機選擇變數，以避免樹與樹間具有過強的相關性。

多模型集成之隨機森林相較於單一決策樹的優點在於：避免過擬合現象發生以及降低模型偏差與變異數。單一決策樹若不經過修剪，容易使用過多變數產生過量節點，因而面臨過擬合（Overfitting）的問題，然而集成大量決策樹，可以幫助隨機森林透過大數法則避免此一現象的發生。此外，多模型的集成，提升了整體模型複雜度，降低了預測偏差，而平均或多數決的集成方式，則減少了模型的變異程度。

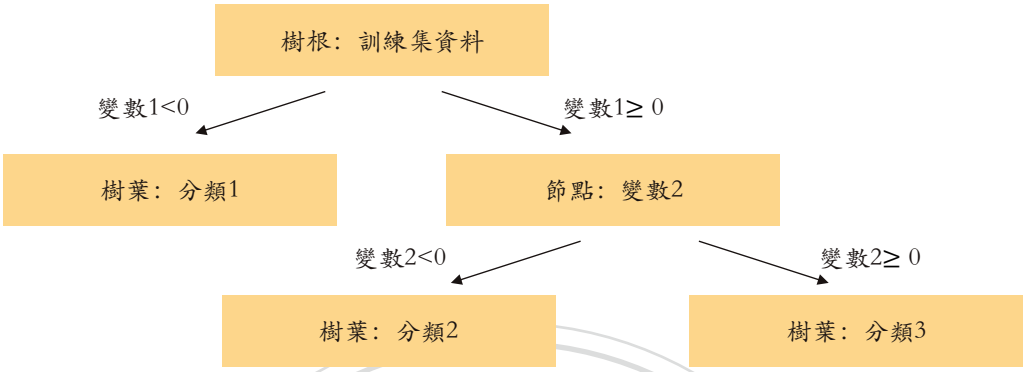
以下分為決策樹與隨機森林兩部分說明模型建構流程，並介紹相關模型調校可以使用之參數：

1. 決策樹建構流程：

決策樹模型是採用遞迴分割（Recursive Partitioning）的方式，歸納訓練集資料所形成之分類規則，建立類似流程圖之樹狀圖，每一棵決策樹中皆包含三個要素：樹根（Root）、節點（Node）、樹葉（Leaf），樹根包

DOI:10.6814/NCCU202100893

含所有訓練資料，在每一個節點上，資料依據不同變數向下分裂為節點或樹葉，樹葉則為最終決策結果。



圖表 5 決策樹分類過程

(1) 將資料分為訓練樣本與測試樣本。

(2) 將訓練樣本放入決策樹的樹根，建立決策樹。

(3) 依據資訊理論（Information Theory）或不純度指標（Impurity），選擇含有最大資訊量的變數作為分割條件。

(4) 重複步驟 2 與步驟 3，直到所有變數都作為分類規則，或是全部資料都被分類完畢。

(5) 使用測試樣本修剪決策樹以避免過擬合現象。

2. 隨機森林建構流程：

(1) 決定隨機森林內的決策樹數量，以及每棵決策樹內使用的變數個數，其中，每棵決策樹內使用之變數個數較原始變數數量小。

DOI:10.6814/NCCU202100893

(2) 從資料中隨機抽取子樣本，作為決策樹的訓練集，並於抽後放回。

(3) 隨機抽取變數作為決策樹每次分裂的準則，同一決策樹中，分裂過後的變數不再重複使用。

(4) 重複步驟 2 到步驟 3，直到建立至預定決策樹數量。

(5) 收集所有決策樹對於新資料的預測後，以多數決（離散資料的分類問題）或是加總平均（連續資料的迴歸問題），做出最終決策。

3. 決策樹模型建構可調整之參數：

(1) 樹的最大深度：在變數量大資料較少時，可以設定樹的分枝數，當樹分裂超過此深度時，自動修剪超出之分枝，避免過擬合現象產生。

(2) 分裂前最低樣本數量：此參數限定了每個節點具有之樣本數量，若低於此數量則不再繼續分裂。

(3) 分裂後最低樣本數量：此參數限定了每次樹進行分裂時，兩類中分別需要包含之樣本數量，否則分裂不會發生，一般搭配最大深度使用，可以幫助模型變得平滑，若要求之觀測值過高，則模型學習無效率；過低則容易產生過擬合。

DOI:10.6814/NCCU202100893

(4) 分裂方法：資訊理論（Information Theory）或不純度（Impurity）指標是決策樹中用於挑選分裂變數的兩大準則，決策樹訓練時會在每一個節點挑選可以使資訊或純度提升最多的變數進行分裂，其中常見的指標包含：熵（Entropy）、Gini 指標、分類錯誤率指標（Classification error）、資訊增益（Information Gain）。

(5) 最低改善率：限制每個節點分裂時之最低改善率，若新的分裂沒有低於最低改善率，則不繼續向下分裂。

4. 隨機森林建模型構可調整之參數：

(1) 決策樹數量：一般而言，樹的數量越多越不容易產生過擬合現象，然而需要較多時間運算。

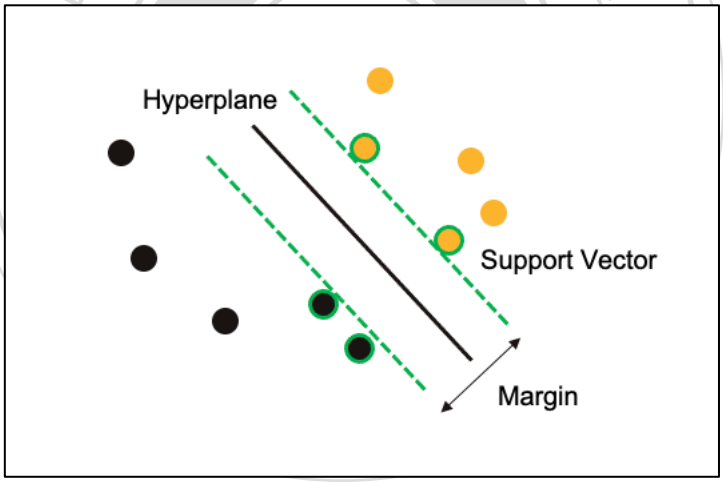
(2) 觀測值比例：建立決策樹時，不會將所有資料納入訓練集中，以製造決策數間之差異性，因此需要決定每棵樹採用之觀測值比例，如果採用的比例越低，需要建立的決策樹數量就越多，全部的資料才能被整體模型使用到。

(3) 變數個數：為了提高決策樹間差異性，因此在建立決策樹時需要決定每棵樹使用的變數數量，若變數數量越高，則越容易產生相似的決策樹。

DOI:10.6814/NCCU202100893

（三） 支持向量機（Support Vector Machine, SVM）

支持向量機模型是在 1995 年由 Vapnik et al.以統計學習理論為基礎所發展出來的，其概念是將原先無法以線性函數區分之資料，利用非線性之透過核心函數（Kernal Function），將原始資料投影至高維度的特徵空間（Feature Space），並在特徵空間中找出對兩類別資料具有最大邊界距離（Margin），且可以將資料區分為兩類之超平面（Hyperplane），圖表 2 中超平面將黃色與黑色兩類別資料區分，其中邊界距離為$\frac{2}{\|w\|}$，w 為超平面上的法向量（Normal Vector）。



圖表 6 超平面示意圖

支持向量機模型建構可調整參數：

1. 懲罰值（Cost）：支持向量機之目的為找到一個可以將資料分為兩類之超平面，然而訓練模型之最終目的為一般化的預測能力，而非追求訓

DOI:10.6814/NCCU202100893

練集的完美分類，因此為了避免過擬合（Overfitting）的現象發生，透過懲罰值的設定，給予分錯的資料容錯空間，以控制最靠近超平面邊界之資料（Support Vector），對於整體模型訓練的影響力，當 C 越大，表示容錯越小，越容易過擬合。

2. Gamma：支持向量機利用核心函數將原始資料投影至高維度空間時，核心函數會影響到資料在特徵空間中的分佈，因此透過此一參數調整核心函數，當 Gamma 越大時，越靠近超平面邊界的資料對於建構超平面的影響力越大，因此較容易發生過擬合現象；反之 Gamma 越小，則可以描繪出較平滑的超平面，較不易發生過擬合現象。

第五節 集成學習方法與建模流程

集成學習方法可以依據個別學習器間的特性，分為「同質集成」與「異質集成」，而大多集成學習方式，又多以建立同質性學習器為基礎學習器（Base Learner），藉由集成方式形成強學習器，以降低變異程度（Variance）或偏差（Bias），主要概念是透過個別學習器的差異性，形成不同預測結果，以提高整體模型預測準確率，一般而言，可以透過抽取不同資料樣本，或是在個別模型中加入不同變數，以建立具有差異性之基礎模型。

本研究中為了大幅提升學習器間的差異性，採用了異質性的模型作為個體學習器（Component Learner），同時訓練羅吉斯迴歸、隨機森林及支持向量機

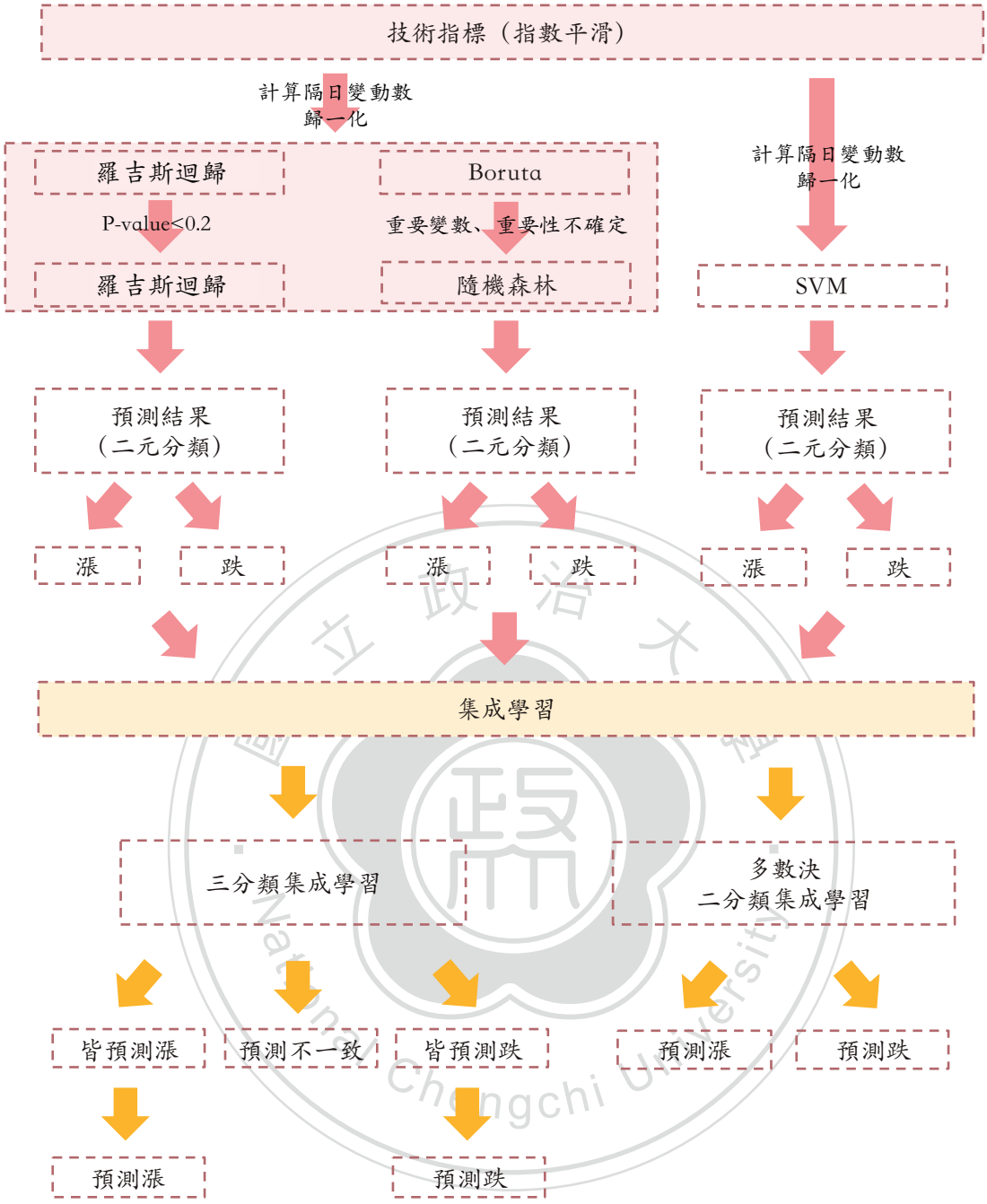
DOI:10.6814/NCCU202100893

模型，以形成最終集成模型，並針對個別模型，利用不同變數篩選的方式提升資料品質，增加個體學習器的準確率，同時提升模型間之多樣性，最終透過集成個別學習器之預測結果，建構完成模型，整體建構過程可以分為兩步驟：

（一） 特徵值挑選：將整體資料依照日期前後分為訓練與測試集，並將訓練集的資料加入羅吉斯迴歸與 Boruta 演算法中，分別依據變數重要性進行篩選，羅吉斯迴歸採用邊際檢定 p-value=0.2 作為篩選標準，去除 p-value 大於 0.2 之特徵值，保留 p-value 小於 0.2 的變數作為第二步驟羅吉斯迴歸模型的特徵值；Boruta 演算法部分，則透過重複模擬訓練，挑選重要性為「重要」與「不確定」之變數，作為第二步驟隨機森林之特徵值；為了保留所有原始變數對模型建構的效果，支持向量機模型之特徵值不進行篩選，使用所有原始變數。

（二） 集成學習：針對個別模型挑選特徵值後，同時訓練三個模型，以預測二元之漲跌分類，並採用不同模型個數，建構集成學習模型。本研究中共採用兩種集成學習方式：多數決二分類集成學習與三分類集成學習，一般而言集成學習大多採用多數決方式形成最終預測，此方式下最終決策結果仍然為二分類；然而在二模型集成學習中，會遇到模型間無共識之情況，因此將基礎模型之二元預測類別分化為三類：全預測漲、全預測跌、預測結果不一致，並著重探討全預測漲與全預測跌兩個類別之預測效果。

DOI:10.6814/NCCU202100893



圖表 7 集成學習訓練流程

DOI:10.6814/NCCU202100893

第四章 實證結果

本研究目的為利用集成模型提升台灣股價加權指數趨勢預測之準確率，以捕捉適當買進時機，資料期間共 30 年，由於股價具時間序列特性，因此將前 25 年作為訓練集、最後 5 年作為測試集，此外本模型欲透過集成學習提升模型成效，故以個別模型作為標竿模型，比較集成學習效果，分別測試兩個模型與三個模型集成學習之效果，並依據不同天期之股價趨勢進行分析。

本研究中一共使用三種異質演算法，組合為五種不同的集成模型，表格 2 為本研究所採用之模型列表，其中使用之集成學習方式又可以分為多數決二分類與三分類。

多數決二分類與一般的集成學習方式無差異，透過模型間投票，以多數之預測結果作為集成模型最終決策，本研究將此方式運用在三模型中，當有兩個以上的模型預測結果為漲時，最終模型預測分類為漲，反之則預測為跌。

股市漲跌趨勢問題原先為二元分類問題，一般而言集成學習採用的方式為多數決，然而本研究採用之二模型集成學習，會遇到模型間無法取得共識之情形，因此本研究將最終集成學習之結果分成三種預測類別：「全預測漲」、「全預測跌」、「預測不一致」，若所有模型預測為漲，則最終模型預測分類為漲；若所有模型預測為跌，則最終模型預測為跌；若模型間預測結果不一致，則另外分為一類，最終側重於探討「預測漲」與「預測跌」兩類別之準確率。

DOI:10.6814/NCCU202100893

表格 2 集成學習模型預測方式

<table>
<tr><td></td><td colspan="3">二模型集成學習</td><td colspan="2">三模型集成學習</td></tr>
<tr><td>模型</td><td>羅吉斯迴歸
隨機森林</td><td>羅吉斯迴歸
支持向量機</td><td>隨機森林
支持向量機</td><td colspan="2">羅吉斯迴歸
隨機森林
支持向量機</td></tr>
<tr><td>集成學習
預測方式</td><td colspan="3">全預測漲：預測漲
全預測跌：預測跌
模型預測不一致：排除</td><td>多數決：
兩模型以上預測漲：預測漲
兩模型以上預測跌：預測跌</td></tr>
</table>

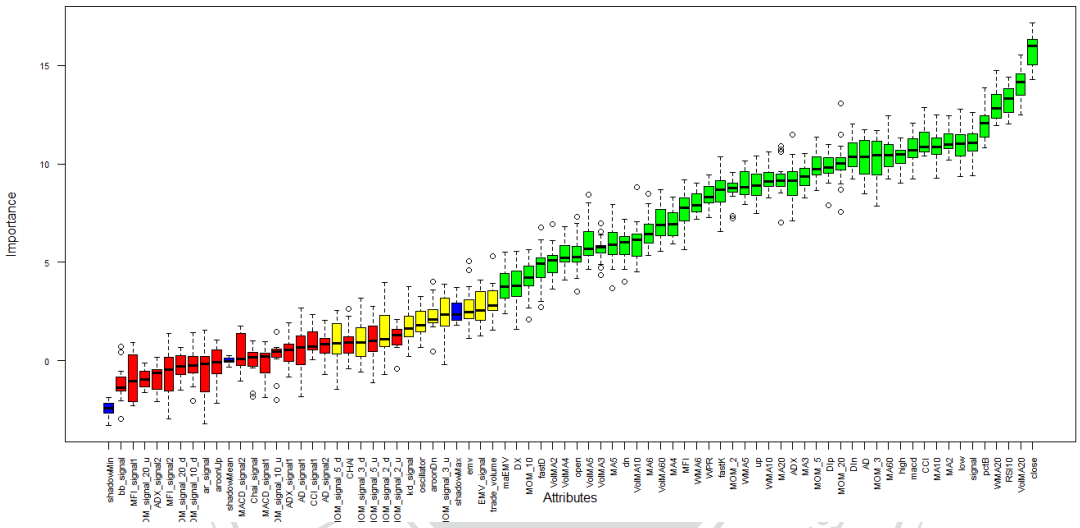
第一階段特徵值篩選中，為了增加資料品質及提升模型間之多樣性，針對個別模型，採用了不同的特徵值，由於使用 Brouta 演算法及邏輯斯模型的邊際迴歸時皆會涉及目標變數，因此表格 3 中不同演算法與不同預測天期，採用的特徵值數量皆不相同。

表格 3 個別模型使用特徵值個數

預測天期/模型	羅吉斯迴歸	隨機森林	支持向量機
5 天	6 個特徵值	55 個特徵值	75 個特徵值
10 天	6 個特徵值	54 個特徵值	75 個特徵值
15 天	4 個特徵值	51 個特徵值	75 個特徵值

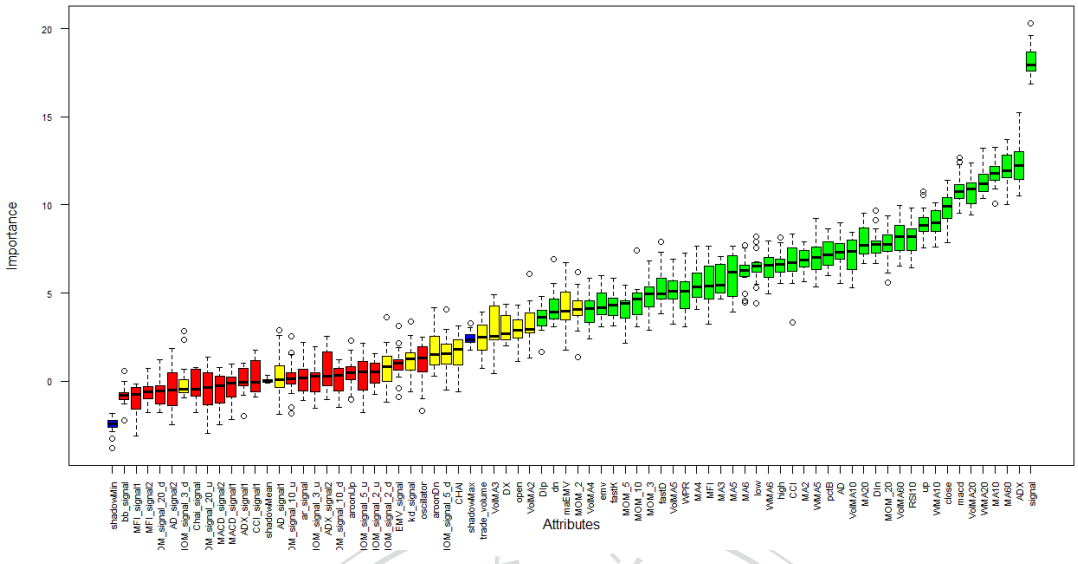
DOI:10.6814/NCCU202100893

圖表 8、圖表 9、圖表 10 中，為各個預測天期以 Boruta 演算評斷之變數重要性，綠色為重要之變數，黃色為重要性不確定之變數，紅色則為不重要之變數。透過下圖可以發現，預測天期越長重要之變數越少，不過本研究中，僅排除紅色部分之不重要變數，並保留重要與重要性不確定之變數，用於隨機森林模型之建構，因此不同天期所保留之變數數量沒有太大的差異。

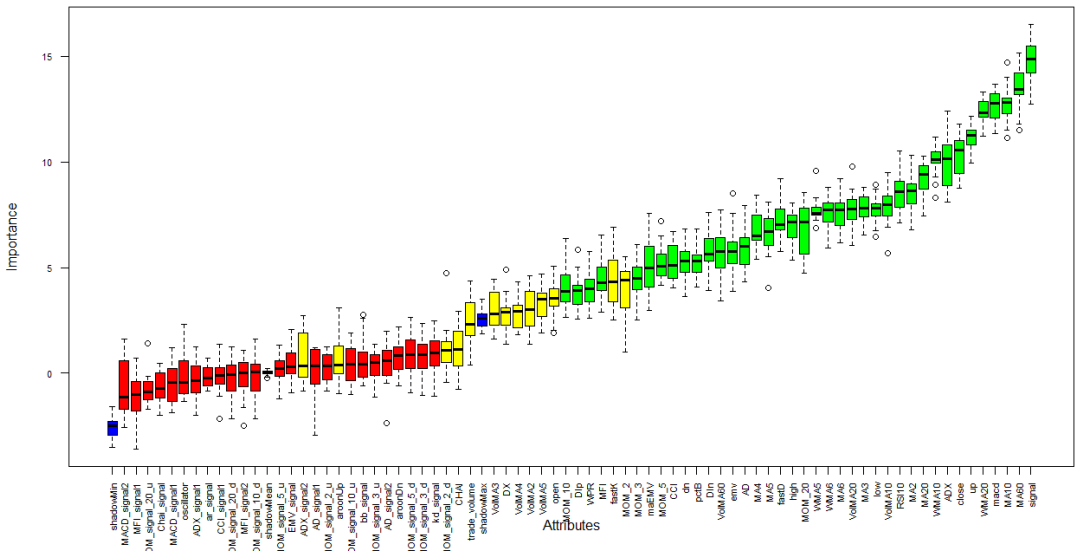


圖表 8 隨機森林預測天期 5 日之變數挑選結果

DOI:10.6814/NCCU202100893



圖表 9 隨機森林預測天期 10 日之變數挑選結果



圖表 10 隨機森林預測天期 15 日之變數挑選結果

下圖表格 4 至表格 9 為不同預測天期下，單一模型之預測準確率與集成學習模型之預測結果。預測天期五日的部分，單一模型與集成模型中預測漲之準確率相較於預測跌之準確率高許多，此外所有集成學習的總準確率都有 1%~2%之提升，其中又以預測跌的準確率上升幅度較預測漲大一些。

DOI:10.6814/NCCU202100893

表格 4 預測天期 5 日單一模型之預測準確率

績效/模型	羅吉斯迴歸	隨機森林	支持向量機
預測漲準確率	67%	66%	67%
預測跌準確率	51%	47%	49%
總準確率	62%	60%	62%

表格 5 預測天期 5 日集成模型之預測準確率

	二模型集成學習			三模型集成學習	
模型	羅吉斯迴歸 隨機森林	羅吉斯迴歸 支持向量機	隨機森林 支持向量機	羅吉斯迴歸 隨機森林 支持向量機	
集成學習方式	全預測漲/全預測跌/預測不一致				多數決
預測漲準確率	69%	68%	68%	69%	68%
預測跌準確率	52%	51%	51%	51%	51%
總準確率	64%	63%	64%	64%	63%

DOI:10.6814/NCCU202100893

預測天期十日的部分，單一模型中預測漲之準確率相較於預測跌之準確率高許多，預測跌的表現相對預測天期五日較差，不同單一模型間的準確率差異性較大，不過集成學習後大部分模型準確率皆有提升。

表格 6 預測天期 10 日單一模型之預測準確率

績效/模型	羅吉斯迴歸	隨機森林	支持向量機
預測漲準確率	69%	69%	68%
預測跌準確率	42%	38%	40%
總準確率	63%	58%	61%

表格 7 預測天期 10 日集成模型之預測準確率

	二模型集成學習			三模型集成學習	
模型	羅吉斯迴歸 隨機森林	羅吉斯迴歸 支持向量機	隨機森林 支持向量機	羅吉斯迴歸 隨機森林 支持向量機	
集成學習方式	全預測漲/全預測跌/預測不一致				多數決
預測漲準確率	69%	69%	70%	70%	68%
預測跌準確率	45%	40%	40%	42%	40%
總準確率	64%	63%	63%	65%	62%

DOI:10.6814/NCCU202100893

預測天期十五日的部分，單一模型中預測漲之準確率相較於預測跌之準確率高許多，預測跌的表現相對短天期的較差，因此總準確率隨之下降，不過集成學習後的總準確率大致上都有提升，若分為預測漲與預測跌來觀察，可以明顯觀察到，集成學習在長天期預測跌的準確率並無提升，這部份是因為長期而言股價趨勢為漲的資料比例較高，導致單一模型預測為跌之次數較少，造成不同模型間預測跌之準確率波動幅度大，如表格 8 所示，因而導致集成學習後預測跌之準確率不增反減。

表格 8 預測天期 15 日單一模型之預測準確率

績效/模型	羅吉斯迴歸	隨機森林	支持向量機
預測漲準確率	65%	69%	70%
預測跌準確率	26%	34%	38%
總準確率	54%	57%	60%

表格 9 預測天期 15 日集成模型之預測準確率

	二模型集成學習			三模型集成學習	
模型	羅吉斯迴歸 隨機森林	羅吉斯迴歸 支持向量機	隨機森林 支持向量機	羅吉斯迴歸 隨機森林 支持向量機	
集成學習方式	全預測漲/全預測跌/預測不一致				多數決
預測漲準確率	68%	70%	71%	71%	68%
預測跌準確率	19%	22%	35%	17%	33%
總準確率	59%	61%	62%	62%	59%

DOI:10.6814/NCCU202100893

整合不同天期之預測結果，可以發現集成學習後預測漲的準確率都維持在七成左右的水準，然而在預測跌的部分，則隨著天期增加，有逐漸下降的趨勢，尤其是最長天期的十五日趨勢預測，預測跌的準確率經過集成學習後，出現不增反減的效果。

就整體準確度而言，不同天期下總準確率都在六成左右，其中在長天期的趨勢預測上，集成學習的效果最好，總準確率提升的程度最高，整體而言，本研究發現集成學習確實可以增加機器學習的預測效果，尤其是隨機森林與支持向量機二模型的組合以及三分類的三模型組合，在不同天期下準確率都最佳。

DOI:10.6814/NCCU202100893

第五章 結論與建議

本研究以台股加權指數之漲跌趨勢作為目標變數，透過異質性模型間之集成學習，以及個別模型之變數篩選方式，提升整體模型預測成效，最終與單一模型之預測準確率比較分析，發現集成學習確實可以帶來 1%至 2%之準確率提升，尤其是長天期的趨勢預測，集成學習的效果更加明顯，此外，觀察漲與跌兩類別之預測成效，所有的模型在預測漲的準確率提升幅度，都較預測跌之準確率上升幅度顯著。

此外，本研究也進一步使用模型預測結果，形成交易策略，以計算報酬率，雖然與準確率之表現並未完全一致，但在不同天期的模型下，都可以觀察到集成學習模型間的報酬率穩定性較高。

借鑑本研究之經驗，以下五點可供後續研究參考:

（一） 提升模型個數：一般而言，平行訓練組合之集成模型，採用之模型數量較高，以提升整體模型穩定性與降低預測結果偏差之效果，然而本研究中，由於採用異質性模型進行訓練，因此模型數量較低，單一模型對於整體集成學習之影響力大，未來可以考慮提升異質性模型數量，或是採用異質同質混合之方式，針對異質演算法，個別建立多個同質性模型，例如：三模型集成學習下，建立三個隨機森林模型、三個支持向量機模型、三個羅吉斯迴歸模型，最終使用九個模型之集成結果。此外，提升模型個數亦

DOI:10.6814/NCCU202100893

可解決目前模型數較少、模型間無法取得共識之問題，提升整體模型預測漲跌趨勢之比例。

（二） 抽取不同訓練樣本以提升模型間差異性：集成學習的 Bagging 方法中，整合許多同質性演算法，並透過訓練樣本與變數的抽取，創造模型之差異性，本研究已透過異質性演算法建立模型，並利用變數重要性挑選特徵值，提高整合模型之多元性，因此未抽取樣本，直接將全段訓練集作為個別模型訓練集，未來可以考慮針對不同模型隨機抽取不同的觀測值，更加提升模型間之差異程度，優化集成學習效果。

（三） 支持向量機模型之變數挑選：本研究中為了提高個別模型差異性，因此分別針對羅吉斯迴歸模型與隨機森林模型，選用對應之變數挑選方式：羅吉斯迴歸採用羅吉斯迴歸邊際檢定結果挑選；隨機森林採用以隨機森林模型為基礎之 Boruta 演算法；為了保留所有變數之效果，因此在支持向量機模型中保留了所有原始變數。未來可以考慮透過遞迴特徵消除（Recursive feature elimination，RFE)或傳遞 Anova 支持向量機（Pipeline Anova SVM）等方法，利用特徵值與目標間的關係或迭代排序特徵影響力挑選變數，以提升模型訓練速度與成效。

（四） 增加預測分類：本研究目的除了透過集成學習方式提升台股大盤指數漲跌趨勢預測外，更希望可以透過模型建立自動交易策略，取得穩定之獲

DOI:10.6814/NCCU202100893

利，回應投資者之期待，因此建議將二分類之漲跌趨勢預測擴展為多類別之區間報酬率預測，以提高模型準確率與交易策略報酬率之一致性。

（五） 建立交易策略：雖然本研究以台灣加權股價指數之漲跌趨勢作為目標變數，但由於此一指數為整體市場之股價加權指標，並非可以直接交易之標的，因此建議可以透過其他標的，例如：台灣 50 指數、指數股票型基金等等，以實現自動化之交易策略。

DOI:10.6814/NCCU202100893

參考文獻

1. Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. Expert systems with Applications, 42(20), 7046-7056.
2. Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. The North American Journal of Economics and Finance, 47, 552-567.
3. Di, X. (2014). Stock trend prediction with technical indicators using SVM. Independent Work Report, Stanford Univ.
4. Dutta, A., Bandopadhyay, G., & Sengupta, S. (2012). Prediction of stock performance in the Indian stock market using logistic regression. International Journal of Business and Information, 7(1), 105.
5. Jiang, M., Liu, J., Zhang, L., & Liu, C. (2020). An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. Physica A: Statistical Mechanics and its Applications, 541, 122272.
6. Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. J Stat Softw, 36(11), 1-13.

DOI:10.6814/NCCU202100893

7. Li, H., Yang, Z., & Li, T. (2014). Algorithmic trading strategy based on massive data mining. Stanford University Stanford.

8. Larsen, J. I. (2010). Predicting stock prices using technical analysis and machine learning (Master's thesis, Institutt for datateknikk og informasjonsvitenskap).

9. Moews, B., Herrmann, J. M., & Ibikunle, G. (2019). Lagged correlation-based deep learning for directional trend change prediction in financial time series. Expert Systems with Applications, 120, 197-206.

10. Mierswa, I., & Morik, K. (2005). Automatic feature extraction for classifying audio data. Machine learning, 58(2), 127-149.

11. Naik, N., & Mohan, B. R. (2019, May). Stock price movements classification using machine and deep learning techniques-the case study of indian stock market. In International Conference on Engineering Applications of Neural Networks (pp. 445-452). Springer, Cham.

12. Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert systems with applications, 42(1), 259-268.

13. Vapnik, V. N. (1995). The nature of statistical learning. Theory.

DOI:10.6814/NCCU202100893

14. Żbikowski, K. (2015). Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. Expert Systems with Applications, 42(4), 1797-1805.

DOI:10.6814/NCCU202100893

附錄

表格 10 預測天期 10 日個別模型採用之變數

羅吉斯迴歸	隨機森林			
6 個特徵值	55 個特徵值			
low	close	VolMA3	MOM_20	CCI
VolMA20	open	VolMA4	fastK	emv
WMA20	high	VolMA5	fastD	maEMV
fastK	low	VolMA10	aroonDn	MFI
aroonDn	trade_volume	VolMA20	oscillator	RSI10
ADX_signal1	MA2	VolMA60	dn	WPR
	MA3	WMA5	up	AD
	MA4	WMA6	pctB	kd_signal
	MA5	WMA10	macd	EMV_signal
	MA6	WMA20	signal	MOM_signal_2_d
	MA10	MOM_2	DIp	MOM_signal_3_u
	MA20	MOM_3	DIn	MOM_signal_3_d
	MA60	MOM_5	DX	MOM_signal_5_d
	VolMA2	MOM_10	ADX	

DOI:10.6814/NCCU202100893

表格 11 預測天期 10 日個別模型採用之變數

羅吉斯迴歸	隨機森林			
6 個特徵值	54 個特徵值			
WMA20	close	VolMA3	MOM_20	emv
aroonUp	open	VolMA4	fastK	maEMV
aroonDn	high	VolMA5	fastD	MFI
DIn	low	VolMA10	aroonDn	CHAI
DX	trade_volume	VolMA20	dn	RSI10
MFI	MA2	VolMA60	up	WPR
	MA3	WMA5	pctB	AD
	MA4	WMA6	macd	kd_signal
	MA5	WMA10	signal	MOM_signal_2_d
	MA6	WMA20	DIp	MOM_signal_3_d
	MA10	MOM_2	DIn	MOM_signal_5_d
	MA20	MOM_3	DX	AD_signal1
	MA60	MOM_5	ADX	
	VolMA2	MOM_10	CCI	

DOI:10.6814/NCCU202100893

表格 12 預測天期 15 日個別模型採用之變數

羅吉斯迴歸	隨機森林			
4 個特徵值	51 個特徵值			
VolMA10	close	VolMA3	MOM_20	emv
VolMA20	open	VolMA4	fastK	maEMV
DX	high	VolMA5	fastD	MFI
bb_signal	low	VolMA10	aroonUp	CHAI
	trade_volume	VolMA20	dn	RSI10
	MA2	VolMA60	up	WPR
	MA3	WMA5	pctB	AD
	MA4	WMA6	macd	ADX_signal2
	MA5	WMA10	signal	MOM_signal_2_d
	MA6	WMA20	DIp	
	MA10	MOM_2	DIn	
	MA20	MOM_3	DX	
	MA60	MOM_5	ADX	
	VolMA2	MOM_10	CCI	

DOI:10.6814/NCCU202100893