國立政治大學風險管理與保險學系研究所

碩士學位論文

以集成學習建構混合模型預測台灣加權股價指數之趨勢

Forecasting the Trend of TAIEX by Using Ensemble Learning

指導教授：黃泓智 博士

研究生：徐維延 撰

中華民國一零八年七月

DOI:10.6814/NCCU201900623

致　謝

在政治大學的日子，已到了尾聲。當初以大學生的身分來到政治大學，總是對於未來有些迷惘與徬徨，遲遲未能找尋到自己所嚮往的領域。雖然蹣跚於乾涸沙漠中，但卻發現了一片綠洲，開始灌溉與滋潤我的思想，這正是商學院給予我的新視野。驀然回首，在這近十年的青春中，政治大學總是陪伴著我的喜怒哀樂，太多的回憶都交織於此。然而，碩士是成長最多的階段，在相關課程紮實的磨練下，除了學習專業的知識，也開始懂得與同儕、老師們溝通及討論，並獲得回饋與成長。在此，非常感謝黃泓智老師在我入學時，邀請我加入老師的研究團隊，除了在論文上的循循善誘之外，老師也在我學習上需要幫助與指導的時候，全心全力地付出。另外，在老師的研究團隊中，也要特別感謝蕭鈞銓學長，在這兩年來，不論在研究的技術層面與資源上的支援都相當盡心盡力。我很高興能夠加入106級風險管理與保險學系研究所的大家庭中，因大家來自不同的背景，更能夠激盪出不同的思維與創意。尤其是精算組的好朋友們，無論是在學習上或者日常生活中，都是我最好的陪伴。

由衷感謝在旁默默付出的家人與親朋好友。在台北生活的這些年，不論何時何處，您們都是我最堅固的後盾，無時無刻給予百分之百的支持與鼓勵。最後，期許自己能盡其所學，回饋給社會，讓世界更好。

維延 寫於

2019 年 06 月

DOI:10.6814/NCCU201900623

摘　要

本研究的目標在於如何準確地預測台灣加權股價指數在數日後是否上漲至超過預設門檻，蒐集並萃取台灣加權股價指數之技術指標、其他國際重要股市指數及台灣總體經濟指標三種面向資料作為特徵值，總共有 192 個特徵。藉由集成學習的概念提出一個混合模型，並以單純的隨機森林模型作為標竿進行比較。因蒐集之資料皆具有時間性，故使用增長式視窗滾動法(Increasing Window Rolling)以驗證模型績效表現。結果顯示，單純的隨機森林模型雖在短天期的預測準確率高，但易受門檻標準訂定的影響，使得樣本呈現分類失衡的現象；反之在長天期的預測準確率較低，但對於不同門檻值也較為穩定，同時 AUC 指標也呈現較佳的表現。雖然此研究提出的混合模型並無在模型準確率上有明顯優於單純的隨機森林模型，但也觀察到混合模型的預測若能避開國際金融動盪的時期，模型表現應能不錯。

關鍵字：台股大盤、集成學習、混合模型、技術分析指標、總體經濟指標

DOI:10.6814/NCCU201900623

Abstract

The purpose of this study is emphasized how to accurately forecast the uptrend of Taiwan Capitalization Weighted Stock Index (TAIEX) in next few days, which is required to exceed different default thresholds. The data collections in three aspects comprise technical indicators of TAIEX, other influential stock markets in the world and Taiwan's macroeconomic indicators as model inputs. After extracting the crucial information behind these variables, there are 192 features in total.

By proposing a blending model based on ensemble learning, the study will present a comparison with the simple random forest model. Besides, it is worth noting that raw data is temporal ordering; therefore, "Increasing Window Rolling" will be the validation method to evaluate the performance of models. The results have shown that the simple random forest model has high predictions in short periods but prone to be affected by different default thresholds, which may make sample imbalanced. On the contrary, predictions are less accurate in long periods but more stable under different default thresholds. In addition, the AUCs are also better. Although the proposed blending model is not significantly superior to the simple random forest model, it may still provide a good performance if phase of financial crisis is disregarded.

Keywords: Taiwan Capitalization Weighted Stock Index, Ensemble Learning, Blending Model, Technical Indicators, Macroeconomic Indicators

DOI:10.6814/NCCU201900623

目　次

DOI:10.6814/NCCU201900623

表　次

DOI:10.6814/NCCU201900623

圖　次

DOI:10.6814/NCCU201900623

第一章　研究背景與動機

第一節　機器學習之發展

在科技快速發展的趨勢下，以往無法達到的數據處理能力，現今已有顯著地突破。機器學習作為人工智慧的分支，在二十世紀中已被提出，而對於機器學習的期待是在於如何在沒有明確地指引下，可使電腦進行自我的學習，進而對目標產生解答。現今，資料分析師、科學家及工程師在不同的領域上已有廣泛的發展，包含資料探勘、搜尋引擎、醫學診斷、DNA 序列檢測等。

機器學習可分析資料之規律，並對未知的資料進行預測，其中涉及的演算法大量使用了統計學原理，故機器學習與統計領域關係也相當密切。回顧歷史，早在十九世紀時，Legendre 和 Gauss 發表了線性迴歸的概念並成功地應用在天文學領域。線性迴歸可被用來預測量化的數值，例如個人薪資，亦可預測質化結果，例如股價漲跌與否。而後於 1936 年 Fisher 提出線性判別分析(Linear Discriminant Analysis)、1940 年代發展出羅吉斯迴歸分析(Logistic Regression)，後來於 1970 年代初期由 Nelder 與 Wedderburn 共同建構出廣義線性模型(Generalized Linear Models)，涵蓋了許多統計學習的方法。1970 年代後期，有越來越多的資料學習技術開始發展，但因計算能力上的不足，大多都只侷限在線性方法中。1980 年代中期，Breiman、Friedman、Olshen 與 Stone 引入了分類與迴歸樹模型，首次展示了這些模型在實務上強大的處理能力與交叉驗證的技術。Hastie 與 Tibshirani 於 1986 年將廣義線性模型更加拓展至廣義加成模型(Generalized Additive Models)使其納入非線性模型的部分。此後，機器學習開始被啟發而統計學習開啟了新的領域，逐漸著重在監督式與非監督式學習以及預測方面。

在財務領域上，如何準確預測股價的走勢一直以來是最複雜的問題之一。有太多的因子可能會影響預測，無論是環境因子、經濟狀況、政治影響、心理層面、抑或是理性或不理性的投資者行為等，這些不同面向的因素皆導致股價的波動及

DOI:10.6814/NCCU201900623

產生高度的不確定性。因此，股票市場通常被認知為是充滿動態、非參數、混亂以及充滿雜訊的本質，其短期走勢亦被認為是一種隨機過程。基於股票本身帶有不確定性的因子，投資人在股票市場往往要承受高度風險。為了降低風險，必須要先捕捉到股價在未來走勢的資訊。投資人偏好持有未來期望上漲的股票，而會選擇放空未來期望下跌的股票，所以，如何準確地預測股價的走勢，使資本利得極大化以及風險極小化，則需仰賴建構機器學習的模型以具備高度的預測能力。故本研究將以機器學習的技術，試圖準確地捕捉股價走勢的訊息。

第二節　監督式學習框架

在機器學習中，演算法通常引導著電腦運作的邏輯與思路。然而，機器學習能夠解決多少問題，則取決於演算法的發展。而在機器學習的領域中，主要可將模型學習的方式分為三類，分別為 (一) 監督式學習、(二) 非監督式學習及 (三) 強化式學習(見表 1)。

本研究的目標將以台灣加權股價指數於未來一定期間後之上漲與否作為標記值，並以潛在的特徵作為建構模型的因子，故採取「監督式學習」框架。為了達到建構模型與評估其績效，須將資料集合區分為「訓練集」與「驗證集」(見圖 1)。另外須注意的是，此兩種集合必須是互斥的。

第三節　模型偏差與變異數之抵換關係

在大多的統計學習方法，不論模型複雜度，目標通常為捕捉真實母體的樣貌，但往往不可能知道真實的母體。藉由抽樣方法，可對母體作統計推論。假使真實母體為某種函數形式：$f = f(x)$，經抽樣得到 m 筆訓練資料 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$，並使用統計方法建立模型$\hat{f}(x)$以估計母體，此時滿足$y = \hat{f}(x) + \varepsilon$，同時$E(\varepsilon) = 0$。

DOI:10.6814/NCCU201900623

表 1　主要三種學習方法之框架

學習方法	使用範圍
(1) 監督式學習	資料已有標記，運用已標記資料訓練。例如：分類、迴歸分析。
(2) 非監督式學習	資料沒有標記，從中找出擁有相同特徵的資料群。例如：分群。
(3) 強化式學習	可能沒有任何資料，直接使模型執行，將執行結果反饋回模型再訓練。例如：動態系統、機器人控制。

然而，對於建構出的預測模型是否能準確地捕捉訓練樣本通常並不是研究上感興趣的部分，而應是關心對於尚未使用在建立模型的測試樣本，是否可藉由此模型被準確地預測。故假設有 n 筆測試資料 $T = \{(x_1', y_1'), (x_2', y_2'), \ldots, (x_n', y_n')\}$，則測試之均方誤差(test mean-squared error; test MSE)為：

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i' - \hat{f}(x_i'))^2 \quad (1.1)$$

再將取其期望值可評估模型的績效。經過分解後，可知$E\{[y_i' - \hat{f}(x_i')]^2\}$是由三個成分組成，如(1.2)式，詳細推導請見附錄 1：

$$E\{\text{MSE}\} = E\left\{\frac{1}{n}\sum_{i=1}^{n}\left(y_i' - \hat{f}(x_i')\right)^2\right\} = \frac{1}{n}\sum_{i=1}^{n}E\{[y_i' - \hat{f}(x_i')]^2\}$$

$$= \text{Average of}\{\,\text{VAR}(\varepsilon) + [\text{Bias}(y_i')]^2 + \text{VAR}(y_i')\} \quad (1.2)$$

為了要使得測試之均方誤差(test MSE)最小化，已知殘差項的變異數$\text{VAR}(\varepsilon)$此項是無法消除的，故如何將模型的偏差(Bias)和變異數(VAR)降低則是使模型提升預測能力的方法，但實際上並不容易達到。假使有一模型不理會任何樣本的資訊，對於預測採取固定值，如此一來可使得模型的變異數為零，但同時卻會使得模型的偏差變大；反之，當模型可百分之百完美地捕捉樣本的偏差程度，但會使得每次預測的結果會因為樣本的樣貌會有明顯的起伏，使得模型的變異數變大。

DOI:10.6814/NCCU201900623

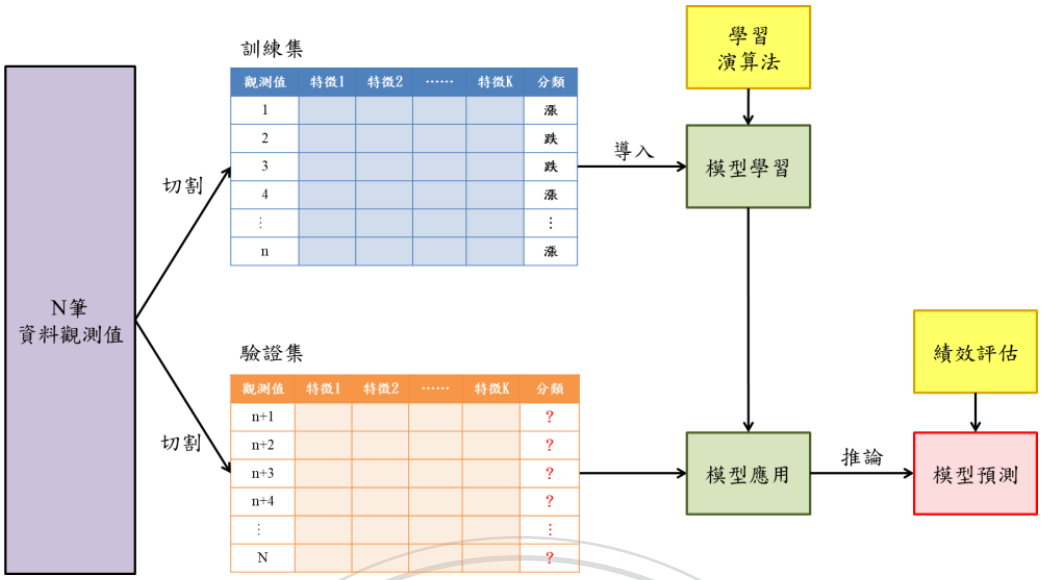

圖 1　監督式機器學習流程圖

一般而言，訓練誤差可以透過越複雜的模型降低，但測試誤差通常並非單調遞減的函數，而可能呈現 U 字型的趨勢，反而會在過於複雜的模型上預測表現更差(見圖 2)，故如何使得測試誤差達到最小可能的值，正是找到最佳模型的方法。若模型使用過於複雜而使得測試誤差上升，則呈現「過度配適」的問題，使得模型對於預測的偏差會降低，但變異數提升；反之，若模型使用過於簡單而使得測試誤差上升，則呈現「低度配適」的問題，使得模型對於預測的變異數會降低，但偏差則會提升。這兩者互相消長，即為模型偏差與變異數之抵換關係。

第四節　集成學習方法

集成學習，亦稱為多分類器系統(Multiple classifier systems)，是由多個基礎學習器(Base learners)所組成，其精神在於集結「群體的智慧(Wisdom of crowds)」。一般而言，集成學習的泛化能力(Generalization ability)[1]通常比基礎的學習器更強。實際上，集成學習會如此強大是在於其能夠增強弱學習器。雖然可將基礎學習器稱作弱學習器，但實務上可在基礎學習器選擇「不那麼弱」的學習器，在效果上也會表現得相當不錯。

[1] 泛化能力是指機器學習模型對於新樣本的適應能力。

DOI:10.6814/NCCU201900623

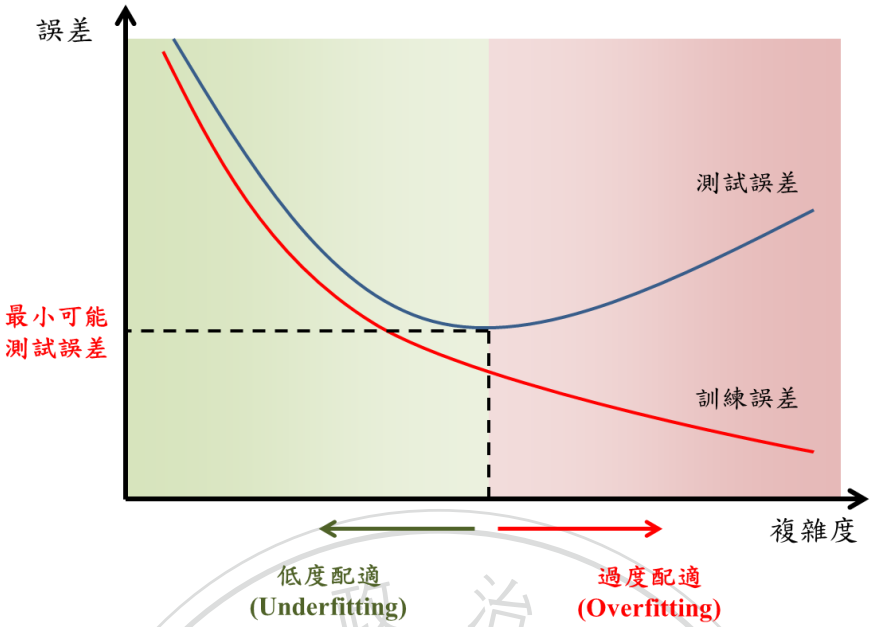

圖 2　模型複雜度與誤差之關係示意圖

基礎學習器常常可以藉由基礎學習演算法(Base learning algorithm)代入訓練樣本而產生，而這樣的基礎演算法包含了決策樹、神經網路等。雖然大多的集成學習方法使用了單一一種基礎學習演算法而產生同質性的基礎學習器，但仍有其他的方法可以使用不同的學習演算法以產生異質性的學習器，同時因為沒有單一的基礎學習演算法的緣故，又可將基礎學習器稱作成分學習器(Component learners)或者個別學習器(Individual learners)。

原則上，集成學習分為兩個步驟。首先，以平行或有順序[2]的模式生成數個基礎學習器。而後，所有的基礎學習器一併被使用，常見的合併方法包含了多數投票的概念(分類問題)以及加權平均的概念(迴歸問題)。一般而言，要得到一個好的集成學習模型，基礎學習器應該盡可能地準確，同時也要盡可能地具有多樣性。檢測一個學習器的準確度可以使用交叉驗證或是 Hold-out 測試，但多樣性卻沒有一個很嚴謹的指標來衡量。然而，陸陸續續有許多衡量多樣性的指標被提出，但 Kuncheva and Whitaker (2003) 仍對於這些多樣性的指標存有懷疑。目前在實務上，基礎學習器的多樣性可以藉由不同的方式引入，例如對於訓練樣本隨

[2] 一個基礎學習器的產生會影響到下一個基礎學習器。

DOI:10.6814/NCCU201900623

機抽樣、屬性處理、輸出處理、加入隨機性到學習演算法中、或者同時有多種的機制等。只要是經過不同基礎學習器的生成過程或者不同的合併方式，都會導致不一樣的集成學習方法。集成學習的方法有很多，主要有三種主流的方法：

(一) Bagging

又稱作 Bootstrap Aggregation 或 Bootstrap，由 Leo Breiman (1994) 提出，為一種簡單且具有力量的集成學習，同時考量許多同質(Homogeneous)的弱學習器，而這些弱學習器皆是獨立且平行建構的，再將其各自的結果經由平均或投票的過程以決定最終的結果。

(二) Boosting

首度由 Yoav Freund 和 Robert E. Schapire (1996) 提出。Boosting 亦是考量到許多同質的弱學習器，與 Bagging 不同的是，這些基礎模型是以順序性的方式適應與學習，再將結果以某種決定性的策略結合起來。

(三) Stacking

相較於 Boosting 與 Stacking 使用同質性的弱學習器，Stacking 則是使用異質性(Heterogeneous)的弱學習器，以平行的方式建構各自的模型，並將各自不同的弱學習器的預測結果結合起來，以訓練一個 Meta 模型，並得到輸出。

後續內容安排如下：第二章回顧各領域如何預測股價及機器學習在預測股價之相關文獻貢獻；第三章呈現整個研究的架構並確立問題與設定解決的目標，以不同的觀點蒐集多面向的數據資料並說明其來源。然而，資料須做進一步的萃取與精煉的緣故，於第四章將展示資料如何前置處理，並生成可能潛在的特徵；第五章則利用訓練集資料以建構混合模型，以不同的績效指標，並以隨機森林模型作為標竿，將驗證集資料作為評估模型的依據，以分析模型之配適程度；最後，第六章以結論說明本研究的貢獻以及未來的發展。

DOI:10.6814/NCCU201900623

第二章　文獻回顧

早期，有許多學者認為股價的改變是不能被預測的，最有名的學說為 Fama (1970) 提出的隨機漫步假說(Random Walk Hypothesis)以及 Jensen (1978) 提出的市場效率假說(Efficient Market Hypothesis)，其論述如果不能從市場中獲取經濟收益，即當前市場是對所有資訊有效率的。不過，有學者持有相反論點。透過個人有限的資訊集合起來，如果以適當的方式將眾多信息萃取，藉由眾人的智慧以及協同過濾(Collaborative Filtering)之效果，則可以提供準確的評估。Avery, Chevalier, Zeckhauser (2016) 則以自然人與法人立場應證了協同過濾效果，打敗市場並獲取利潤。

在過往研究股價的行為有從不同的面向探討。包含 (一) 技術指標分析，Khaidem et al, (2016) 以 RSI、隨機指標等技術指標作為模型之輸入，其對於股票價格的處理先進行了平滑化處理，並運用了隨機森林模型分析以不同時間長短對於股價的預測能力，準確度可達到九成。Jan Ivar Larsen (2010) 則利用技術指標及 K 線理論建構雙層模型，以挪威 Oslo 股票市場實證，打敗大盤 250 個百分點。(二) 時間序列預測與多變量分析，Berninger, Jordan (2018) 以 Apple 公司的股價作為預測目標，以 ARIMA 及 ARIMA+GARCH 進行分析，顯示 ARIMA 較合適短期的預測，而 ARIMA+GARCH 的預測較適合長期，最後提出在長期預測下，多變量模型在長期的表現皆比單變量模型更好。(三) 機器學習與資料採礦，Suryoday Basak et al, (2019) 利用隨機森林及 XGBoost 模型對於各個公司之股價進行不同時間長度的預測，並認為中長期的預測較為準確。(四) 利用微分方程以建構模型與預測股票波動，Saha, Routh and Goswami (2014) 使用了擴散模型 (Diffusion Model)以解釋選擇權價值。

近期在股價市場行為的領域主要是以機器學習為主，這些方法脫離了傳統的預測方法以及隨機理論。早期研究的方法大多是以時間序列模型以及多變量分析為主，如 Gencay (1999), Timmermann and Granger (2004), Bao and Yang (2008)，

DOI:10.6814/NCCU201900623

是將股價的移動當成是一個時間的函數，並以迴歸的問題解決。然而，要準確地估計股價確實並非易事,但倘若將問題轉為分類問題,預測的表現就會相對較好。因此，許多模型藉由市場資料結合了學習技巧，包含了使用最多的支持向量機(Support Vector Machine ; SVM)、神經網絡、Naïve Bayesian 分類器等，目標著重在如何準確預測股價的走向。在 Li, Li and Yang (2014) 使用對於股價較為敏銳的外部數據，包含黃金價格、原油價格、天然氣等，與美國 NYSE 及 NASDAQ 自 2000 年 01 月 01 日至 2014 年 11 月 10 日的日資料，以羅吉斯迴歸分析的結果顯示，模型之勝率表現可達 55.65%。Dai and Zhang (2013) 將 2008 年 01 月 09 日至 2013 年 11 月 08 日的 3M 公司股價，總樣本數為 1,471 筆，使用多個預測系統進行模型配適，包含羅吉斯迴歸、二次判別分析(Quadratic Discriminant Analysis)及 SVM，預測隔日以及多日後的股價，而結果顯示，隔日的準確率落在 44.52%至 58.20%。其同時應證美國股票市場是半強勢效率市場，說明不論是基本面資訊或技術分析都無法獲取明顯的利得。然而，長期預測的模型在天期為 44 日的時候表現最好，表現最好的 SVM 有著 79.30%的高準確率。Xinjie (2014) 則使用了三檔股票，分別是 Apple、Amazon 及 Microsoft，期間為 2010 年 01 月 04 日至 2014 年 12 月 10 日，結合技術指標生成 84 個特徵，使用極限隨機樹(Extremely Randomized Tree)作為特徵選取的方法，最後使用 RBF Kernelized SVM 作為學習模型，並使用貪婪搜尋法(Grid Search)以調整參數。最終模型的準確率依照不同的預測期間，落在 60%至 80%。Devi, Bhaskaran and Kumar (2015) 建議一個支持向量機配合著混合布穀鳥搜尋法(Hybrid Cuckoo Search)，布穀鳥搜尋法是一種支持向量機最佳化參數的方法，使用了技術指標作為特徵，並以實際市場數據分析自 2013 年 01 月至 2014 年 07 月的資料。Giacomel, Galante and Pareira (2015) 則使用了神經網絡集成方法以預測北美股票市場及巴西股票市場的漲跌。以上談及的文獻能幫助瞭解在集成學習演算法的領域上仍有許多尚未發掘的部分。本研究將以集成學習作為研究股價趨勢的架構,試圖集合不同模型的組合使得模型預測效果更佳。

DOI:10.6814/NCCU201900623

第三章　研究方法

第一節　研究架構

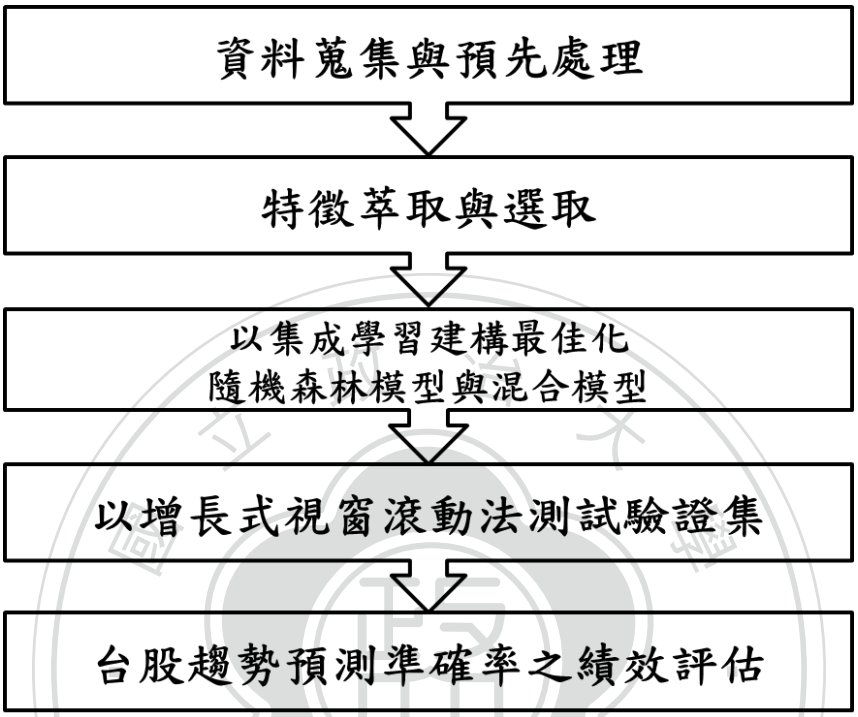

圖 3　研究流程圖

本研究將蒐集台灣加權股價指數從 1999 年至 2018 年，共二十年之日資料，嘗試以不同領域的特徵資料，包含技術指標、其他重要國家股市指數及總體經濟指標。藉由集成學習的精神，集結不同模型的結果，試圖極大化未來 N 日後股價為上漲的預測準確率。然而，為了要測試模型的績效，會將資料集合依照時間先後做切割，以避免使用未來資料影響模型的建構，故使用了增長式視窗滾動法(Increasing Window Rolling)測試驗證集。從驗證集可調整模型的參數以達到最佳化的模型，並以多個模型的績效指標判斷模型對於台灣加權股價指數上漲預測之學習程度。同時，以不同未來天期的預測時間，分析模型預測準確率是否受到預測時間長短的影響。本研究之流程請見圖 3。

DOI:10.6814/NCCU201900623

第二節　資料來源與預先處理

(一) 資料概述

1. 資料時間：1999 年 01 月 01 日至 2018 年 12 月 31 日，共 20 年期間。

2. 資料來源：TEJ 台灣經濟新報資料庫。

3. 資料類型與頻率：

(1) 台灣加權股價指數，日頻資料。

(2) 其他國際重要股市指數，日頻資料。

(3) 台灣總體經濟變數，月頻資料。

(二) 資料預先處理

資料的乾淨程度與品質對於模型學習是非常重要的。對於機器學習而言，「Garbage in, garbage out.」，如果資料沒有預先處理，建模的效果將會有所影響。故在資料蒐集的過程中，必須考量到以下細節：(1) 完整性、(2) 噪音、(3) 一致性、(4) 遺漏值、(5) 離群值。然而，以 TEJ 資料庫所維護的狀況，應都能符合以上的要求，唯有在蒐集總經變數，發現有些資料紀錄時點不足，會缺少多筆早期資料，故在挑選變數時，將存有遺漏值之變數移除，不納入變數考量。

第三節　變數介紹

一、台灣加權股價指數(日頻資料)

台灣加權股價指數(TAIEX，以下簡稱為「台股」)為台灣證券交易所所編制的股價指數，以 1966 年為基期，設定為 100。台股計算的方式是採用 Paasche 公式，與美國 S&P 相同，是採用整體市場股票價值變動的指標。以上市的股票市值作為權重計算股價指數，採用的樣本為所有掛牌交易中的普通股，故股本較大的公司對於台股的影響力較大，例如台積電、鴻海、國泰金、大立光等。

DOI:10.6814/NCCU201900623

本研究從 TEJ 台灣經濟新報資料庫，蒐集台股從 1999 年之第一個交易日至 2018 年之最後一個交易日，選取各交易日之開盤價、最高價、最低價、收盤價、交易量(Trading Volume)與交易值(Trading Value)，作為原始資料變數。圖 4 為台股近二十年走勢，長期觀察有逐漸上升之趨勢，但也於數個時期面臨嚴重的下跌：(1) 21 世紀初面臨網路泡沫破滅、經濟負成長等因素，重創台股；(2) 2007 年至 2008 年爆發次級房屋信貸危機，引發全球金融危機；(3) 2015 年 08 月面臨全球股災，台股創下有史以來最大跌幅紀錄。

從圖 5 可觀察近二十年台股之日報酬率表現，分別於 21 世紀初、金融海嘯時期及 2015 年的日報酬率之中位數略低於其他年份。另外，波動最大的時期則是落在 2000 年至 2003 年及最劇烈的 2008 年。然而近年來，台股在負報酬率的離群值較正報酬率的離群值明顯。總言之，台股市場波動大，如何找出較高報酬的訊號，是值得研究與探討的。

二、其他國際重要股市指數(日頻資料)

隨著全球貿易頻繁，貿易體系逐漸融為一體。回顧歷史，股票市場曾多次經歷泡沫化，牽動著全球股市，如 1989 年至 1992 年的日本房市泡沫、1997 年至 1999 年的亞洲金融風暴、2000 年至 2001 年的美國網路泡沫以及 2007 年至 2009 年的金融海嘯。本研究蒐集了全球主要經濟體各交易日之股市收盤價資料，共有 12 個股市之日資料(見表 2)。根據此十二國股市與台股進行相關係數分析，發現大多股市皆與台股呈現高度正相關[3](見表 3)。

三、台灣總體經濟指標(月頻資料)

根據許多文獻針對不同國家之總體經濟與該國股市間關係的研究，皆顯示二者之間長期而言有著密不可分的關係。Joseph Tagne Talla (2013) 蒐集瑞典股票市

[3] 相關係數為介於-1 至 1 區間的數值。其絕對值小於 0.1 為無相關，介於 0.1 至 0.39 為低度相關，介於 0.40 至 0.69 為中度相關，介於 0.7 至 0.99 為高度相關，等於 1 為完全相關。

DOI:10.6814/NCCU201900623

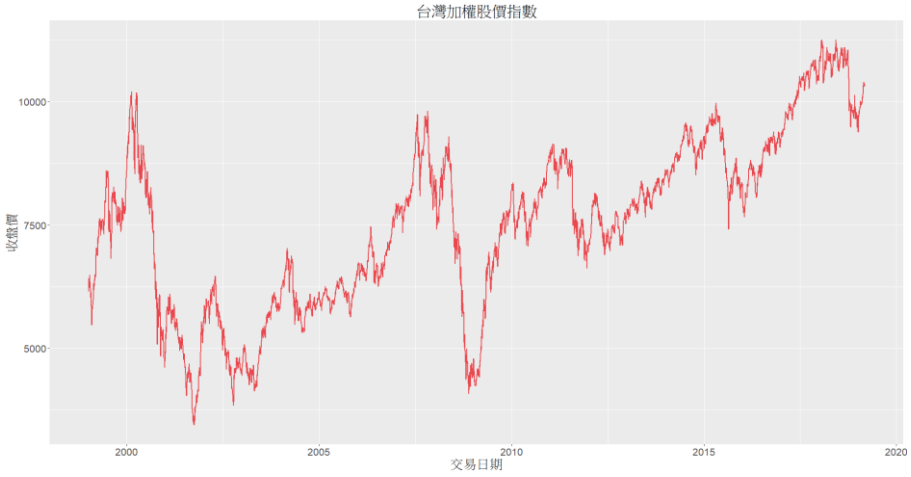

圖 4　近二十年台灣加權股價指數走勢圖

圖 5　近二十年台灣加權股價指數之日報酬率箱型圖

場(Stockholm Stock Exchange)自 1993 年至 2012 年的資料，分析總體經濟與股市間的關係，發現通貨膨脹率及貨幣貶值與股價有著顯著負相關，反倒是利率沒有統計上顯著的關係。Ahmed Imran Hunjra et al. (2014) 以及 Jawad Khan et al. (2018) 皆以巴基斯坦股市作為研究。其發現短期而言，股市雖與總經變數較無明顯關係，但長期而言是有統計上的顯著關係。Amith Vikram Megaravalli et al. (2018) 以亞洲三大國家中國、印度及日本以統計檢定分析匯率、消費者物價指數(為「通膨率」之替代變數)與股市之間的關係，發現長期匯率與股市之間的關係是顯著的。

DOI:10.6814/NCCU201900623

表 2　其他國際重要股市指數列表

其他國際重要股市指數名稱
(1)　美國—紐約道瓊工業平均指數
(2)　日本—東京日經 225 指數
(3)　香港—恆生指數
(4)　新加坡—富時海峽時報指數
(5)　德國—DAX 指數
(6)　泰國—曼谷 SET 股價指數
(7)　馬來西亞—吉隆坡綜合股價指數
(8)　菲律賓—馬尼拉綜合股價指數
(9)　韓國—綜合股價指數
(10) 加拿大—多倫多綜合股價指數
(11) 法國—巴黎 CAA ALL-TRADABLE 股價指數
(12) 英國—倫敦金融時報一百種股價指數

故本研究推測，台灣總體經濟指標對於股市走勢的分析應有預測能力，因此將其納入特徵。詳細的總經變數列表請參附錄 2，主要分成人口就業薪資統計、工業產銷用電量、批發零售商業營業額、物價統計數據、景氣指標以及證券統計數據六項指標，總共 55 個變數。

DOI:10.6814/NCCU201900623

表 3　台灣與其他國際重要股市指數之相關係數矩陣

	台灣	美國	日本	香港	新加坡	德國	泰國	馬來西亞	菲律賓	韓國	加拿大	法國	英國
台灣	1												
美國	0.82	1											
日本	0.72	0.77	1										
香港	0.89	0.80	0.55	1									
新加坡	0.85	0.70	0.47	0.95	1								
德國	0.87	0.94	0.78	0.85	0.77	1							
泰國	0.79	0.89	0.51	0.84	0.82	0.86	1						
馬來西亞	0.81	0.79	0.41	0.89	0.90	0.83	0.95	1					
菲律賓	0.81	0.90	0.56	0.84	0.81	0.91	0.97	0.94	1				
韓國	0.83	0.77	0.41	0.93	0.93	0.81	0.91	0.96	0.89	1			
加拿大	0.85	0.81	0.56	0.94	0.93	0.86	0.86	0.90	0.85	0.93	1		
法國	0.69	0.67	0.89	0.61	0.56	0.75	0.42	0.40	0.47	0.42	0.64	1	
英國	0.85	0.81	0.83	0.76	0.73	0.89	0.68	0.67	0.75	0.66	0.75	0.85	1

DOI:10.6814/NCCU201900623

第四章　資料前置處理與特徵生成

第一節　上漲標準訂定

本篇研究探討的目標變數為「對於第 i 筆觀測值而言，N 個交易日後之收盤價與今日之收盤價的比值是否超過預設的門檻」。以數學符號表示如公式(4.1)。

$$\text{Signal}_i = \begin{cases} 1, & \text{if } \frac{\text{close}_{i+N}}{\text{close}_i} > \text{Default threshold} \\ 0, & \text{otherwise} \end{cases} \quad (4.1)$$

訂定上漲標準的主要目的是考慮到股價波動的特性，有時稍微向上攀升，並不代表有明確的上漲趨勢。故本研究將設定當股價爬升到某水準之上，才定義為上漲，藉此減輕波動的影響。此外，若水準訂太高，可能會使得樣本內與樣本外面臨資料不平衡(Imbalanced)的狀況，只有極少數樣本在上漲水準之上，進而影響模型學習的效果。

第二節　月頻資料處理

因選取的台灣總體經濟資料皆為每月公布，為了要與台股交易日配合，故將月頻資料以複製的方式至該月各個日期，如圖 6 所示。此作法的精神在於避免總經變數因缺乏高頻的資料點，對於日資料之觀測值而言必須移除存有缺漏的觀測值。若不處理資料稀疏的問題，會大幅影響樣本數。

第三節　技術指標生成

技術指標是指研究過去金融市場的資料，以預測價格的趨勢，並作為決策投資的決策標準之一。技術指標的精神在於認為「歷史會不斷重演」，故嘗試以大量的統計資料以預測走勢。然而，技術指標目前廣泛被交易者與金融專家使用。

技術指標在本研究分成四項類別(見表 4)，分別為擺動類、趨勢類、動量以

DOI:10.6814/NCCU201900623

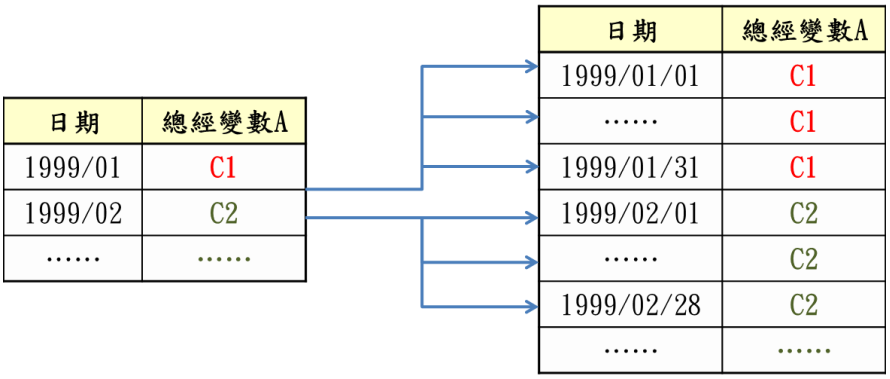

圖 6　月頻資料轉日頻資料示意圖

及量能指標，以下將逐一介紹。

(一) 擺動類指標

1. Aroon 指標

Aroon 指標是由 Tushar Chande 於 1995 年發明的，透過計算自價格達到近期之最高值與最低值以來所經過的時間，而 Aroon 指標可以幫助投資者辨認資產的趨勢轉變以及趨勢的力道。其背後主要的原理是，當有強烈上漲的趨勢，往往會看到創新高的價格；反之，當有強烈下跌的趨勢，往往會看到創新低的價格。Aroon 指標包含了 Aroon up 和 Aroon down 兩種線以衡量上升及下降的強度。Aroon 指標公式如(4.2)與(4.3)：

$$\text{Aroon Up} = \frac{\text{N}-\text{自 N 期中創造最高價格至今的時間}}{\text{N}} \times 100 \quad (4.2)$$

$$\text{Aroon Down} = \frac{\text{N}-\text{自 N 期中創造最低價格至今的時間}}{\text{N}} \times 100 \quad (4.3)$$

其中，N 的數值一般為 25。從以上兩個公式可推得 Aroon Up 與 Aroon Down 的數值皆在 0 至 100 間波動，當值越接近 100，表示有強烈的趨勢；反之，接近 0 的時候，則表示趨勢較薄弱。故假設 Aroon Up 所計算出來的值較低，可以推測上升的趨勢較為薄弱、下降的趨勢較為強烈。當期數 N 等於 25，計算出 Aroon Up 若數值大於 50，表示在近 12.5 期數出現過最高價格；若 Aroon Up 非常接近 100，表示在非常近的期數中出現過最高價格。同樣的解釋邏輯也可套用至 Aroon

DOI:10.6814/NCCU201900623

表 4　四大類技術指標

分類	指標名稱
(一) 擺動類指標	1. Aroon Indicator 2. Bollinger Bands (BBands) 3. KD Indicator
(二) 趨勢類指標	1. Simple Moving Average (SMA) 2. Volume Weighted Moving Average (VWMA) 3. Exponential Moving Average (EMA) 4. Moving Average Convergence/Divergence (MACD) 5. Average Directional Movement Indicator (ADX)
(三) 動量指標	1. Relative Strength Index (RSI) 2. Commodity Channel Index (CCI) 3. Ease of Movement Value (EMV) 4. Money Flow Index (MFI)
(四) 量能指標	1. On Balance Volume (OBV) 2. Chaikin Volatility

Down 上。此外，當兩條線交叉的時候，可視為進場或出場點。當 Aroon Up 向上穿越 Aroon Down，此為買進訊號；當 Aroon Up 向下穿過 Aroon Down，此則為賣出訊號。值得注意的是，當兩條線同時低於 50 的時候表示價格正在盤整 (Consolidating)，代表不會產生著新高價或新低價。

2. Bollinger Bands (簡稱 BBands)

Bollinger Bands 是由 John Bollinger 創立的均線軌道，用以修正不同種類的資產所呈現的波動度不同，以消除以往當使用移動平均軌道的時候會有固定軌道的缺陷。其原理為利用價格與均線的標準差作為軌道的寬度，此時軌道便可以適用於不同類型的資產或股票，便能指出極端值並得到超買或超賣的資訊。

DOI:10.6814/NCCU201900623

Bollinger Bands 的計算方法如(4.4)至(4.7)：

第一步、先求出均線作為基準(N 通常為 20)。

$$MA_t = \frac{P_t + P_{t-1} + \cdots + P_{t-N+1}}{N} \quad (4.4)$$

第二步、求出每天價格對MA_t的標準差。

$$SD_t = \sqrt{\frac{\sum_{i=0}^{N-1}(P_{t-i} - MA_t)^2}{N}} \quad (4.5)$$

第三步，每日之上下界 Upper Bound (UB)和 Lower Bound(LB)則可求出。

$$UB_t = MA_t + 2 \times SD_t \quad (4.6)$$

$$LB_t = MA_t - 2 \times SD_t \quad (4.7)$$

UB 為 MA 加上兩倍標準差所建構的，又名「壓力線」；而 LB 為 MA 減去兩倍標準差所建構的，又名「支撐線」。一般而言，價格皆會落在區間內上下波動，故當價格突破上、下線的時候，顯示此時的價格波動與過去相比較為異常，亦可能表示標的物產生基本性的改變，可視為向上突破或向下跌破的訊號。

此外，由 Bollinger Bands 衍生出%b 指標，可輔助 Bollinger Bands 的判讀與運用。%b 指標以數字的形式呈現收盤價於 Bollinger Bands 的位置，以作為交易決策的關鍵指標。計算公式如(4.8)：

$$\%b_t = \frac{\text{收盤價}_t - LB_t}{UB_t - LB_t} \quad (4.8)$$

由於收盤價會在上下界中震盪遊走，幅度或有可能超過軌道的範圍，因此%b 的數值並無上下限。當走勢向上突破，收盤價落於 UB 的上方，此時%b>1；而走勢向下突破的時候，收盤價落於 LB 的下方時，%b<0。藉由觀察%b 指標可提供投資參考，亦可依照指標的強弱做為決策。

DOI:10.6814/NCCU201900623

3. KD 指標

KD 指標，又稱 Stochastic Oscillator(隨機指標)。為美國企業家 George C. Lane 於1950年代開始廣泛使用，其為一種動量分析的方法並採用超買與超賣的概念，透過比較收盤價格與價格之波動範圍，試圖預測價格趨勢逆轉的時間點。KD 指標的計算公式如(4.9)至(4.12)：

$$\text{Fast \%K} = \frac{\text{第 N 日收盤價} - \text{最近 N 日內最低價}}{\text{最近 N 日內最高價} - \text{最近 N 日內最低價}} \times 100 \quad (4.9)$$

$$\text{Fast \%D} = \text{Fast \%K 的三日簡單移動平均} \quad (4.10)$$

$$\text{Slow \%K} = \text{Fast \%K 與三日簡單移動平均取平滑} \quad (4.11)$$

$$\text{Slow \%D} = \text{Slow \%K 的三日簡單移動平均} \quad (4.12)$$

一般而言，N 的值預設為 14。K 值為「快速平均值」，反應較為敏感；D 值為「慢速平均值」。當 K 值大於 D 值，表示處在漲勢，反之則處在跌勢。K 值與 D 值的數值範圍皆介於 0 至 1 之間，落在 0.5 時為多空平衡位置，落在 0.8 以上為超買區，為多頭強勢，落在 20%以下為超賣區，為空頭強勢。

(二) 趨勢類指標

1. 簡單移動平均(Simple Moving Average; SMA)

移動平均(Moving Average; MA)又稱「均線」，其可藉由平均的方式弭平短期的波動，反應較為穩定長期的趨勢或週期。移動平均可依照不同的加權方式得到不同類型的移動平均線，包含簡單移動平均(Simple Moving Average; SMA)、加權移動平均(Weighted Moving Average; WMA)、指數移動平均(Exponential Moving Average; EMA)。

簡單移動平均是加權移動平均的特例，即當每日權重皆相同的情況。公式如(4.13)：

$$\text{SMA}_N = \frac{P_1 + P_2 + \cdots + P_N}{N} \quad (4.13)$$

DOI:10.6814/NCCU201900623

在技術分析中，會依照不同的市場而有不同的天數 N 使用。普遍而言，N 通常為 10、20、60、120、240，視乎分析時間的長短而定。如果當短期 MA 向上穿越長期 MA，可視為買進訊號；反之，當短期 MA 向下穿越長期 MA，則可視為賣出訊號。

2. 成交量加權移動平均(Volume Weighted Moving Average; VWMA)

成交量加權移動平均為一關注交易量的移動平均，根據給定期間內的交易活動量以衡量價格。由於 VWMA 的加權比重是使用交易量，故得到的值應能更反應市場的真實價格。其公式如(4.14)：

$$\mathrm{VWMA_N} = \frac{P_{t0}\times V_{t0}+P_{t1}\times V_{t1}+\cdots+P_{tN}\times V_{tN}}{V_{t0}+V_{t1}+\cdots+V_{tN}} \quad (4.14)$$

判斷訊號點的原理跟 SMA 相同，當短期 VWMA 向上穿過長期 VWMA 時，可視為買進訊號；當短期 VWMA 向下穿過長期 VWMA 時，則可視為賣出訊號。

3. 指數移動平均(Exponential Moving Average; EMA)

指數移動平均是以指數的方式處理移動平均。各數值的加權影響力取決於時間的遠近，對於近期的數據給予較大的權重，較早的數據給予較小的權重。權重的大小一般而言是由一個參數α決定。其公式如(4.15)：

$$\mathrm{EMA_t} = \begin{cases} P_1, & t = 1 \\ \alpha \times P_t + (1-\alpha) \times \mathrm{EMA_{t-1}}, & t > 1 \end{cases} \quad (4.15)$$

其中α的數值介於 0 到 1 之間，是決定權重的參數。依照權重的不同，可以觀察 SMA 與 EMA 對於不同遠近的價格有不同的反應，因 EMA 在近期給予較大的權重，故其數值反應較為敏感。判斷訊號點的原理跟 SMA 相同，當短期 EMA 向上穿過長期 EMA 時，可視為買進訊號；當短期 EMA 向下穿過長期 EMA 時，則可視為賣出訊號。

DOI:10.6814/NCCU201900623

4. 指數平滑聚散移動平均(Moving Average Convergence/ Divergence; MACD)

由 Gerald Appel 於 1970 年代提出的 MACD，是透過收盤價的短期與長期的 EMA 之間的差距所計算出來的。一般而言，短期 EMA 為 12 日，長期 EMA 為 26 日。計算方式如(4.16)及(4.17)：

第一步、離差值(DIF)計算。

$$\mathrm{DIF} = \mathrm{EMA}_{(\mathrm{Close},12)} - \mathrm{EMA}_{(\mathrm{Close},26)} \quad (4.16)$$

2 第二步、訊號線 MACD 計算，通常是 DIF 的九日指數移動平均值。

$$\mathrm{MACD} = \mathrm{EMA}_{(\mathrm{DIF},9)} \quad (4.17)$$

當股市有強烈的震盪或是波動過大，此時就可能會給予錯誤的訊號。MACD 是 DIF 平滑後的結果，為一種中長線的研判指標，故反應較為遲鈍。當差離值向上穿過訊號線，可視為買進訊號；反之，當差離值向下穿過訊號線，則可視為賣出訊號。

5. 平均趨勢指標(Average Directional Movement Indicator; ADX)

平均趨勢指標是一種常用的趨勢衡量指標，與趨勢指數(DMI)同為 Welles Wilder 所提出的，用於表是市場趨勢的強弱程度，但不能指示趨勢的方向，需要另外使用+DI(Positive Directional Movement Value)及-DI[4](Negative Directional Movement Value)指標線才能確認趨勢的方向。故 ADX 指標包含了三條指標線，分別為 ADX、+DI 以及-DI。然而在動向指標的計算過程較為複雜，必須先決定股價的趨勢，以決定真實波動，才能進行指標的計算。計算公式如下：

第一步、決定股價趨勢(DM)為上漲或下跌。

若今日股價波動幅度大於昨日股價波動部分的最大值，可能是創高價或低價的部分；若今日股價波動幅度小於昨日股價波動，則 DM=0。故紀錄的原則為：

[4] -DM 前的負號“-”並不代表負數，而是表示反向趨勢。

DOI:10.6814/NCCU201900623

(1) 股價高點持續走高，為上漲趨勢，記為+DM。

(2) 若為下跌趨勢，記為-DM。

(3) 其他狀況皆為 DM=0。

第二步、尋找股價的真實波動(True Range; TR)

真實波動是以最高價、最低價以及前一日收盤價三個價格做比較，找出當日股價波動的最大幅度。

第三步、趨勢方向須經由一段時間觀察

計算出+DM、-DM、TR 的算術平均數而得到+DM14、-DM14、以及 TR14 三組指標。利用這三個數值，可以計算「方向指標(Directional Indicator; DI)」，公式如(4.18)及(4.19)：

$$+\mathrm{DI14} = \frac{+\mathrm{DM14}}{\mathrm{TR14}} \times 100 \quad (4.18)$$

$$-\mathrm{DI14} = \frac{-\mathrm{DM14}}{\mathrm{TR14}} \times 100 \quad (4.19)$$

因真實波動為股價波動的最大幅度，趨向指標(DM)不會超過真實波動的值，故方向指標(DI)的數值應該介於 0 至 100 之間。舉例而言，當今日+DI=45 且-DI=25，表示過去 14 日當中向上的趨勢佔了 45%，而向下的趨勢佔了 25%。

第四步、計算「趨向指數(Directional Movement Index; DX)」及「平均趨勢指標(Average Directional Movement Index; ADX)」

$$\mathrm{DX} = \frac{|(+\mathrm{DI14}) - (-\mathrm{DI14})|}{|(+\mathrm{DI14}) - (-\mathrm{DI14})|} \times 100 \quad (4.20)$$

$$\mathrm{ADX} = \mathrm{MA}(|\mathrm{DX}|) \quad (4.21)$$

ADX 可作為趨勢行情的判斷依據。當+DI 向上穿過-DI 且兩條線同時大於

DOI:10.6814/NCCU201900623

ADX 時，可視為買進訊號；反之，當+DI 向下穿過-DI 且兩條線同時小於 ADX，則是為賣出訊號。但如果+DI 及-DI 兩條曲線如果糾結時，代表上漲與下跌力道相當，多空勢均力敵。

(三) 動量指標

1. 相對強弱指數(Relative Strength Index; RSI)

RSI 由美國機械工程師 Welles Wilder JR.提出，為一種藉由比較價格升降的運動以表達價格的強度的技術分析工具。對於每個交易日期，可定義上漲的幅度為 U 或是下跌的幅度為 D，可以下列數學式表達：

(1) 當今日收盤價>昨日收盤價

$$U = \text{今日收盤價} - \text{昨日收盤價}$$

$$D = 0$$

(2) 當今日收盤價<昨日收盤價

$$U = 0$$

$$D = \text{昨日收盤價} - \text{今日收盤價}$$

(3) 當今日收盤價=昨日收盤價

$$D = U = 0$$

在 N 天期間內，可以算出各自的平滑或修正後的移動平均(Smoothed or Modified Moving Average; SMMA)[5]，將其兩數值相除得到相對強度(Relative Strength; RS)。

$$RS = \frac{SMMA(U,N)}{SMMA(D,N)} \quad (4.22)$$

[5] 為一種指數型的移動平均，當權重選取為 1/期間。

DOI:10.6814/NCCU201900623

從 RS 計算方式可知，當 D 值的平均數越接近 0，會導致整個 RS 值趨近於無限大。故可用一些手段將其值壓到 100 內，即為 RSI 的計算公式：

$$\text{RSI} = 100 - \frac{100}{1+\text{RS}} \quad (4.23)$$

一般而言，RSI>80，表示市場過熱，為超買訊號。反之，RSI<20，表示市場過冷，為超賣訊號。另外，當短天期的 RSI 向上穿越長天期的 RSI，可視為買進訊號，反之，當短天期的 RSI 向下穿越長天期的 RSI，則為賣出訊號。

2. 順勢指標(Commodity Channel Index; CCI)

由美國股市分析家 Donald R. Lamber 所發明的，在判斷「非正常」走勢上相當準確的一個技術指標，其針對股價異常波動而設計的，引入價格與固定期間的股價平均區間之偏離程度概念，並強調股價平均絕對偏差的重要性。多數超買超賣的指標皆有 0 至 100 之上下界，因此會有發生指標鈍化的可能。而 CCI 指標的波動可從負無限大至正無限大，不以 0 為中心點，因此不會產生鈍化的情況。

第一步、典型價格(Typical Price; TP)：

$$\text{TP}_t = \frac{\text{最高價}_t + \text{最低價}_t + \text{收盤價}_t}{3} \quad (4.24)$$

第二步、典型價格的移動平均：

$$\text{MA}_t = \frac{\text{TP}_t + \text{TP}_{t-1} + \cdots + \text{TP}_{t-N+1}}{N} \quad (4.25)$$

第三步、MA_t與TP_t離差絕對值之 N 日加總：

$$\text{MD}_t = \frac{|\text{MA}_t - \text{TP}_t| + |\text{MA}_{t-1} - \text{TP}_{t-1}| + \cdots + |\text{MA}_{t+N-1} - \text{TP}_{t+N-1}|}{N} \quad (4.26)$$

DOI:10.6814/NCCU201900623

第四步、CCI 計算：

$$\mathrm{CCI_t} = \frac{\mathrm{TP_t - MA_t}}{0.015 \times \mathrm{MD_t}} \quad (4.27)$$

CCI 的數值可分成三區間：(1) -100 至+100 為常態區。(2) +100 以上為超買區。(3) -100 以下為超賣區。然而當常態區的 CCI，此時參考意義並不大，可採取其他超買或超賣的指標判斷。當 CCI>100 時，表示價格差正向異常，買氣異常旺盛，可視為買進訊號；當 CCI<-100 時，表示價格逆向異常，買氣異常冷清，可視為賣出訊號。但也有一說是 CCI>100 為超買，因此為賣出訊號；反之，CCI<-100 為超賣，因此為買入訊號。本研究將採取前者論點。

3. 簡易波動指標(Ease of Movement Value; EMV)

簡易波動指標是由 Richard W. Arms, Jr.一個將價格與成交量的變化結合在一起的指標，其原理在於價格在上升趨勢的保持過程中不會耗用太多能力，僅當趨勢發生轉折的時候成交量才會轉大。一般而言，當 EMV>0，可視為買進訊號；當 EMV<0 時，則可視為賣出訊號。公式如(4.28)至(4.30)：

$$\text{Distance Moved} = \frac{\text{今日最高價} - \text{今日最低價}}{2} - \frac{\text{前一交易日最高價} - \text{前一日交易日最低價}}{2} \quad (4.28)$$

$$\text{Box Ratio} = \frac{\text{成交量}}{100{,}000{,}000} / (\text{今日最高價} - \text{今日最低價}) \quad (4.29)$$

$$\text{EMV} = \frac{\text{Distance Moved}}{\text{Box Ratio}} \quad (4.30)$$

4. 資金流動指標(Money Flow Index; MFI)

資金流動指標的理念與 RSI 指標類似，除了構成成分有所不同。其原理在於追蹤股票資金的流入或流出。若流入大於流出即為漲，若流入小於流出即為跌。

$$\text{Money Flow Index} = 100 - \frac{100}{1 + \text{Money Flow Ratio}} \quad (4.31)$$

$$\text{Money Flow Ratio} = \frac{\text{近 14 日正現金流}}{\text{近 14 日負現金流}} \quad (4.32)$$

DOI:10.6814/NCCU201900623

$$\text{Raw Money Flow} = \text{Typical Price} \times \text{成交量} \quad (4.33)$$

$$\text{Typical Price} = \frac{\text{最高價}+\text{最低價}+\text{收盤價}}{3} \quad (4.34)$$

若這一期到下一期的 Raw Money Flow 是正值，會被納入正現金流的計算；反之，若這一期到下一期的 Raw Money Flow 是負值，會被納入負現金流的計算。本研究設置一個門檻，若 MFI 值大於此門檻，即為上漲訊號。

(四) 量能指標

1. 能量潮指標(On Balance Volume; OBV)

由 Joseph Granville 於 1963 年提出的能量潮指標是一種依據行情的漲跌，以累計或刪去市場的成交量值，而以此累積值作為市場行情動能變化趨勢的指標。Joseph Granville 的看法在於市場動能應是反應在交易量的變化，至於價格則是另外一個面向的呈現。量應是價的先行指標，先有量才有價。成交量的多寡本質上應反應市場交易的活絡程度，同時也代表市場上人氣聚集的程度。

$$\text{OBV}_t = \text{OBV}_{t-1} + \text{Volume}_t, \text{ if 今日收盤價} > \text{昨日收盤價}$$

$$\text{OBV}_t = \text{OBV}_{t-1} - \text{Volume}_t, \text{ if 今日收盤價} < \text{昨日收盤價} \quad (4.35)$$

OBV 若在上升中，表示買氣轉強，為買進訊號。即使當時股價還是下跌，下跌力道逐漸趨緩，仍是買進訊號。換言之，OBV 若在下降中，表示買氣轉弱，為賣出訊號，即使當時股價還是上升，上漲力道逐漸趨緩，仍是賣出訊號。

2. Chaikin Volatility

Chaikin volatility 為 Marc Chaikin 所開發的一個指標，透過觀察價格在特定時間段內的高點和低點以衡量波動性，公式如(4.36)及(4.37)。當盤整被突破時，CV 值通常處於底部，故可設一門檻，當 CV 值低於此門檻，即為買進訊號。

DOI:10.6814/NCCU201900623

第一步、計算每日高低點差之指數移動平均(通常區間通常為 10 日)。

$$\text{EMA}[\text{H}-\text{L}] = \text{EMA of } (\text{最高點} - \text{最低點}) \quad (4.36)$$

第二步、計算此段期間之移動平均改變比率(通常區間為 10 日)。

$$\text{CV} = \frac{\text{EMA}[\text{H}-\text{L}]-\text{EMA}[\text{H}-\text{L}]_{10\text{ Days ago}}}{\text{EMA}[\text{H}-\text{L}]_{10\text{ Days ago}}} \times 100 \quad (4.37)$$

第四節　特徵標準化過程

在產生特徵的過程中，因單位或衡量的方式不同，易使得原始資料的尺度影響了模型的解釋能力，故對於尺度具有差異的特徵值作標準化的動作。不論訓練集以及驗證集，皆須經過標準化的處理。對於第 i 筆觀測值，第 j 個特徵，標準化的公式如(4.38)：

$$\text{Standardization}(x_{i,j}) = \frac{x_{i,j} - \min(x_{i,j})\text{in training data}}{\text{Max}(x_{i,j})\text{ in training} - \min(x_{i,j})\text{in training data}} \quad (4.38)$$

在此研究中，台股技術指標之連續型特徵則全部經過標準化調整。此外，雖然其他國際重要股市指數以及台灣總經變數的連續型數值皆有與過去不同天期的期間先作比值處理，但仍作標準化調整。經過一連串資料預先處理後，全部特徵變數之資料彙整可見表 5。

DOI:10.6814/NCCU201900623

表 5　特徵資料彙整

資料	型態	特徵個數	簡易說明	特殊資料處理
台股技術指標 (日頻資料)	離散	23	四大類技術指標之變化趨勢	(無)
	連續	35	四類技術指標之數值	※資料經標準化處理。
其他國際重要股市指數 (日頻資料)	連續	24	五、二十日以前與今日收盤價之比值	※資料經標準化處理。 ※因各國股市時差有異，採取資料「遞延一日」的方式處理。 ※為了配合台股開盤的時間，往往面臨到各國的休市時間不一，而一有遺漏值即無法使用該筆觀測值。故遺漏值皆沿用前一交易日之數值。
台灣總體經濟指標 (月頻資料)	連續	110	月、年比值	※月頻資料轉日頻資料，並經標準化處理。 ※因公開資料需統計時間，故往往無法在該月底或下個月初即取得該月資料，故採取資料「遞延兩個月」的方式處理。

註：詳細台股技術指標特徵列表請參見附錄 3。

DOI:10.6814/NCCU201900623

第五章　模型建構與績效

第一節　混合模型架構

大多集成學習的方法，皆使用了相同的弱學習器，集合各自的結果產生一個強大的學習器。然而，本研究欲使用異質性的模型達到如同 stacking 的概念，並與單一隨機森林模型預測的效果比較，以觀察是否可以再由一次的集成學習效果使模型更加優秀。混合模型作法如圖 7，需要分成兩個階段建構：

第一階段：先將整體資料依照時間先後切割成訓練集與原始驗證集，再將訓練集中，切割成子訓練集與子驗證集，作為第一階段建模使用。首先挑選兩種集成學習的方法，第一種為隨機森林(預測結果為報酬率，為連續型數值)，第二種為 AdaBoost(預測結果為上漲與否，為離散型數值)，為第一層模型的基礎學習器。兩種模型使用相同的子訓練集樣本，利用最佳的模型各自對子驗證集作出預測，同時，也對原始驗證集作一次的模型預測。故在第二階段之前，資料會保留子驗證集與原始驗證集之兩個模型預測結果與原本的目標變數(上漲與否)。

第二階段：又稱為 Meta 階段。於第一階段萃取出來的預測結果，將會是 Meta 階段的輸入，而仍然是以監督式方式學習，目標變數仍然是其樣本點的目標變數(上漲與否)。第二階段僅會剩下兩個特徵，分別是由隨機森林與 AdaBoost 的預測結果，而因第二階段剩下變量少，故以羅吉斯迴歸作為簡單的分類器，而羅吉斯迴歸在此也稱作混合學習器或 Meta 學習器。

以下將簡述隨機森林、AdaBoost 及羅吉斯迴歸這三種於混合模型中的基本學習器。

一、第一層基礎模型一：隨機森林模型

在 1995 年，第一個隨機決策森林的演算法是由貝爾實驗室的何天琴(Ho Tin Kam)女士使用了隨機子空間理論所提出。而後，於 2001 年，隨機森林的技術首

DOI:10.6814/NCCU201900623

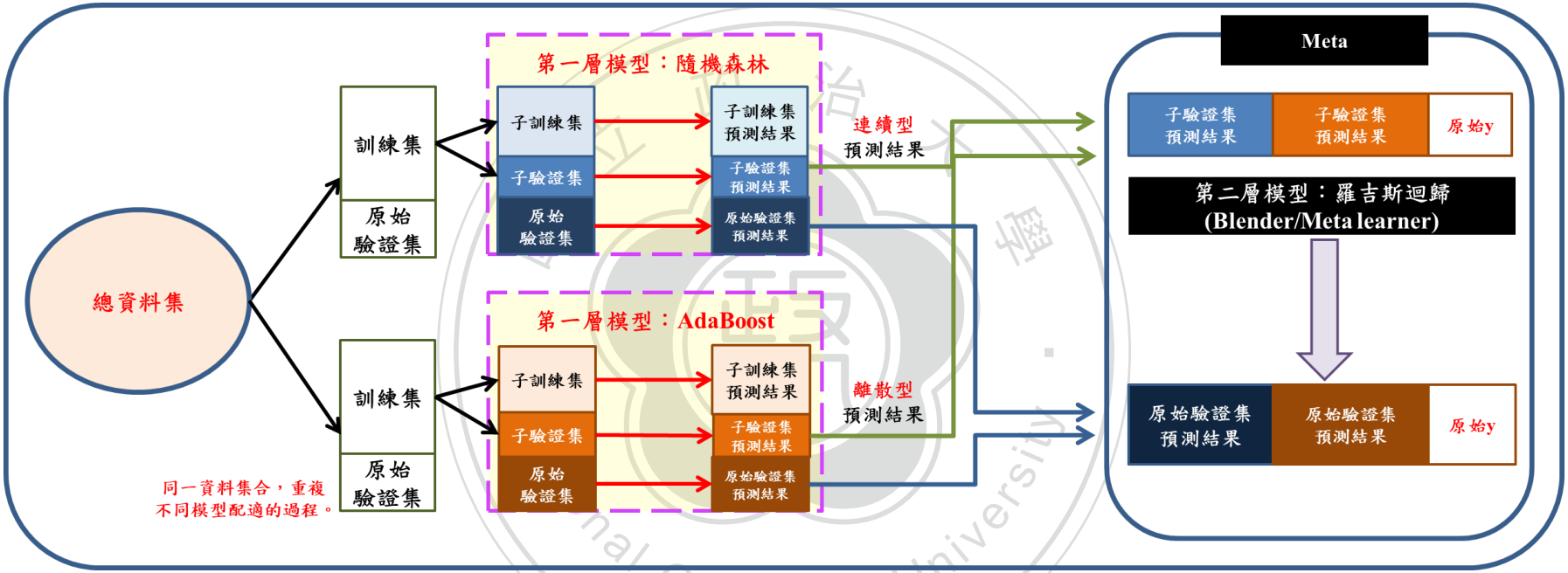

圖 7 混合模型示意圖

DOI:10.6814/NCCU201900623

次被 Leo Breiman 發表在期刊論文上。論文中敘述如何使用不具相關的決策數，透過類似於 CART(Classification And Regression Trees)的過程與結合了隨機節點的最佳化及裝袋(bagging)的概念，以建構隨機森林模型。同時，也提及隨機森林實務上的基本概念：(1) 使用袋外誤差(Out-of-bag error)以估計樣本外誤差、(2)藉由排列的概念測量變數的重要性。因此，Leo Breiman 和 Adele Cutler 發展了一個延伸版本的隨機森林演算法，其融合了 Breiman 的裝袋概念、何天琴女士以及後來由 Amit 和 Geman 所提出的隨機特徵選擇，建構一群決策樹，以達到控制其變異數。

隨機森林為一種集成學習(Ensemble learning)的模型，以決策樹為基礎模型(Base model)。在隨機森林模型訓練期間，藉由一次建構多棵決策樹，依照分類或迴歸的目的，輸出的結果可能為所有決策樹的眾數或平均數。雖然單一一棵決策樹已可捕捉變數之間非線性的關係，但如果沒有限制決策樹的生成，則往往會面臨過度配適(Overfitting)的嚴重問題。這會使得決策樹模型在訓練集的表現相當優秀，但在驗證集的表現不如預期。此外，決策樹對於雜訊的資料也相當敏感。隨機森林演算法之概念如表 6 所示。圖 8 與圖 9 則表示在隨機森林模型下之訓練與驗證階段流程圖。

一般而言，在統計與機器學習領域中，模型的建構皆涉及「偏誤與變異數的抵換關係」。雖然可藉由一個很複雜的模型盡可能捕捉到樣本的樣貌，但同時也帶來比較高的不確定性；反之亦然。理想上，能同時讓偏誤和變異數同時降低是最好的。然而，隨機森林模型卻可達到如此的優點：藉由相互投票或平均的形式，可在建構複雜的模型下，同時降低變異數。

如何調整模型的參數以使得模型達到最佳化是必經的過程，以下的參數皆可以在使用交叉驗證的階段調整各別的值。大多演算法的實作皆有給定預設值，故可以此為出發點調整參數。

DOI:10.6814/NCCU201900623

表 6　隨機森林演算法

一、訓練階段
(1) 隨機抽取觀測值之子樣本，抽後放回，與標準拔靴法(Bootstrap)概念相同。
(2) 隨機抽取部分特徵以決定每次分裂的準則，抽後不放回。
(3) 依前兩步驟建立一棵決策樹模型。
(4) 重複前三步驟，直到預設的決策樹數量為止。
二、驗證階段
(1) 輸入一組新的特徵集觀測值，使用在訓練階段建構的模型進行預測。
(2) 平均其預測值(迴歸問題)或取眾數(分類問題)。

1. 決策樹的數量

基本上，越多樹的預測效果應該越好，尤其資料集合相當大或者是預測變數越多，因需確保每一筆觀測值與每一個特徵至少在一個決策樹被使用過。理論上，隨機森林不應該會過度配適，所以如果上無任何限制或是計算時間的壓力，可以使用越多決策樹建構隨機森林。Segal (2004)提出在某些具有噪音(Noisy)的資料集合當中，可能會發生過度配適的問題。在任何情況下，可以依照交叉驗證的結果以決定最佳的決策樹數量。

2. 觀測值的比例

在建構每一棵決策樹的時候，須考量到多少觀測值被納入。當越少的觀測值樣本被採用，相對而言需要越多棵決策樹使得每筆觀測值都至少被整個隨機森林模型使用到一次。

3. 特徵的比例

每一次分裂需使用某比例的特徵作為分裂準則，要選擇的比例與觀測值的比例概念相似，當越少比例的特徵被選取，即需要更多的決策樹以確保每個特徵值

DOI:10.6814/NCCU201900623

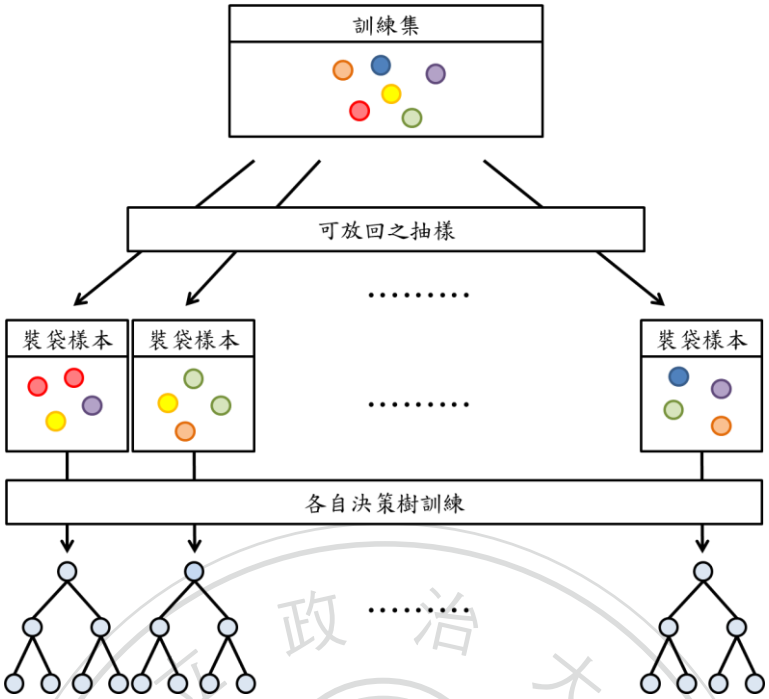

圖 8　隨機森林演算法之訓練階段圖示

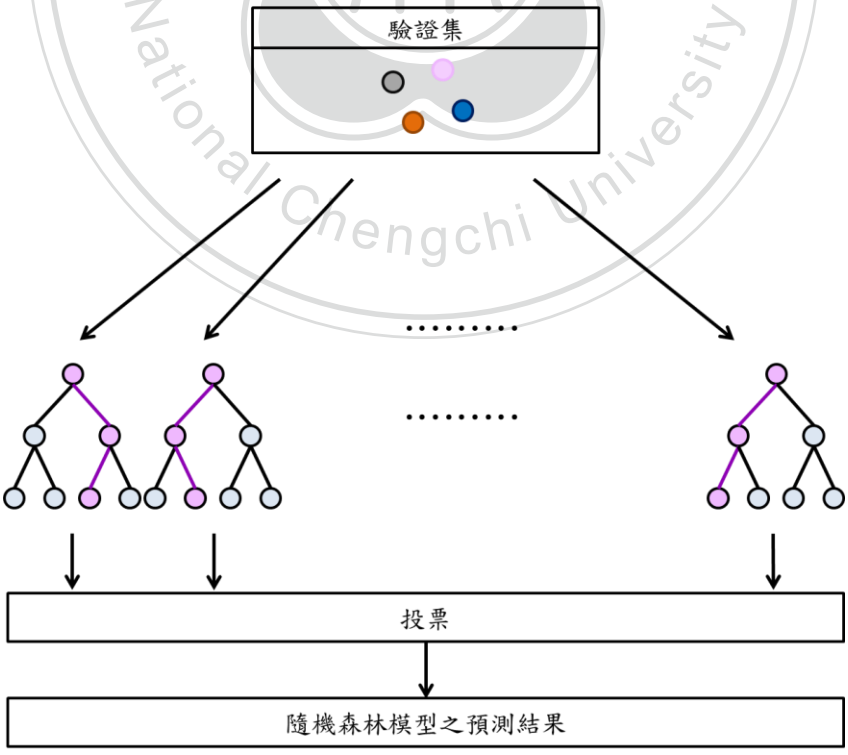

圖 9　隨機森林演算法之驗證階段圖示

DOI:10.6814/NCCU201900623

至少被使用一次。此外，當越大的特徵比例被使用，將會降低 bagging 的效果，會使得每一棵決策樹的模樣長得差不多。此參數通常在模型配適的過程當中時常被調整。

4. 決策樹的參數

決策樹是隨機森林的基礎模型。在隨機森林的概念上，因 bagging 的過程中會使得變異數被大幅減低，故期盼各自的決策樹越複雜越好。理論上，對於各個決策數，以下的參數是可以控制的：

(1) 樹的深度(Depth)

(2) 每次分裂時的最低改善率

(3) 在分裂之前最低要求之觀測值數量

(4) 在分裂之後最低要求之觀測值數量

(5) 分裂方法

決策樹中分裂的原則是以不純度指標(Impurity)衡量是何種特徵可以使得純度提升最多。常見的不純度指標有數種，例如熵(Entropy)、Gini 指標、分類錯誤率指標(Classification error)。其對於每個節點的不純度計算方式(5.1)至(5.3)：

$$\text{Entropy(N)} = -\sum_{i=1}^{c} p_i \log_2 (p_i) \quad (5.1)$$

$$\text{Gini(N)} = 1 - \sum_{i=1}^{c} {p_i}^2 \quad (5.2)$$

$$\text{Classification error(N)} = 1 - \max_{i=1,\ldots,c} p_i \quad (5.3)$$

c 為分類的個數，以上漲與否的二元問題下，c=2。p_i則為在該節點下，各分類種類的占比。依不同的不純度指標計算之後，可以算得由父節點到子節點改善的幅度，進而選取最佳的特徵。改善幅度的指標稱為「資訊獲得(Information Gain)」，公式如(5.4)：

$$\text{Information Gain} = \text{Impurity(Parent)} - \sum_{k=1}^{K} \frac{N_k}{N_p} \text{Impurity}(\text{Child}_k) \quad (5.4)$$

DOI:10.6814/NCCU201900623

綜合以上所提及四種模型參數通常是在交叉驗證以評估模型績效的時候可以進行調整，因此可允許嘗試任何的組合以達到最佳的模型。

二、第一層基礎模型二：AdaBoost 模型

AdaBoost 為 Adaptive Boosting 的縮寫，是一種迭代的演算法。在每一回合加入一個新的弱分類器，直到達到某個預定足夠小的錯誤率。在每一個訓練樣本當中都會給予一個權重，表示被某個分類器選入訓練集的機率。如果當某個樣本點已經被很準確的分類，在下一個訓練集中，則其權重將會被調低；反之，如果某個樣本點沒有被很準確的分類，在下一個訓練集中，其權重則會被提高。因此，AdaBoost 的原理是聚焦在較難分類出來的樣本。由於將注意力放在分類錯誤的資料點上，故通常對噪聲相當敏感。如圖 10 所示。

三、第二層基礎模型：羅吉斯迴歸模型

羅吉斯迴歸模型可使用在當目標變數為二元類型的時候。作法如下：為了要使得目標變數的區間不超過[0,1]區間，故先以羅吉斯函數將其壓縮到[0,1]間。

$$p(X) = Pr(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \quad (5.5)$$

再將其化簡，可得羅吉斯迴歸模型，如公式(5.6)。若要使得模型的預測結果為二元類型，即可在 log-odds[6]或 logit 設定一個門檻。一般而言皆設定為 0.5。

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \quad (5.6)$$

[6] $\ln\left(\frac{p(X)}{1-p(X)}\right)$又可稱作 log-odds 或 logit。

DOI:10.6814/NCCU201900623

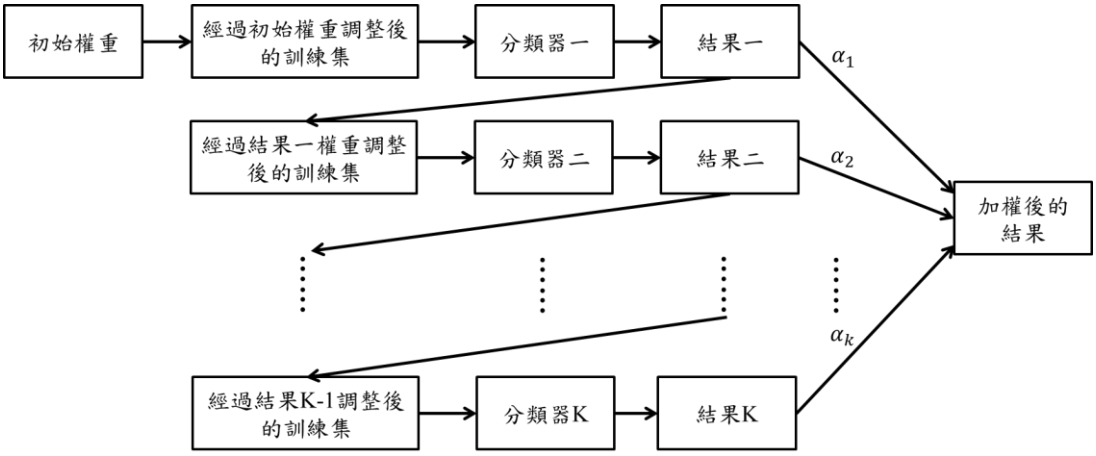

圖 10　AdaBoost 模型示意圖

第二節　績效評估

為了瞭解混合模型的績效表現，故本研究以單純的隨機森林模型作為基準。

一、隨機森林模型績效評估

本研究欲以集成學習方法以準確預測未來台股未來趨勢，提供對於不同天期、不同門檻之準確率進行分析。因股價資料為時間性資料，故將不以傳統 K-fold 交叉驗證的方式預估模型績效以避免使用到未來資料，而是以增長式視窗滾動法(Increasing Window Rolling)交叉驗證，如圖 11 所示。增長式視窗滾動法交叉驗證是將訓練期間的起始點固定，隨著時間的推移，會有更多的訓練資料進入模型學習，並建構出各段時間的集成學習模型，進而預測未來一年內的台股趨勢。因技術指標生成的緣故，會犧牲 1999 年的部分早期資料，故整體建構模型與驗證的時間點皆為 2000 年後。以每年為一個區間，經過每次學習，就會往後加一年資料進入學習，故從 2000 至 2018 年總共可建構 18 次模型。藉由增長式視窗滾動法交叉驗證得到的 18 個準確率結果，取得其平均值後，可做為預估整體模型的預測能力。為了盡可能提升準確率，可回頭調整模型的參數，使其達到最佳化。

DOI:10.6814/NCCU201900623

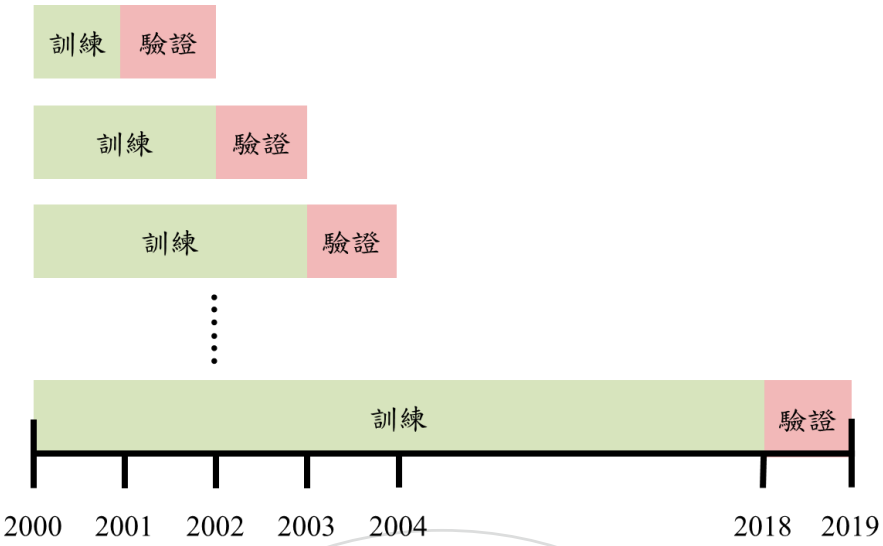

圖 11　隨機森林模型之增長式視窗滾動法圖示

預測股價的漲跌是屬於二元分類問題，根據所訓練集所建立的模型當作分類原則，將所有驗證集分成漲、跌兩類，如圖 12 所示。而分類後的結果應有四種組合：(1) 實際上為漲，預測亦為漲，此分類結果正確；(2) 實際上為跌，預測結果亦為跌，此分類結果正確；(3) 實際上為漲，但預測為跌，此分類結果錯誤；(4) 實際上為跌，但預測為漲，此分類結果錯誤。故此四種結果可以「混淆矩陣(Confusion Matrix)」表示，見表 7。

依照混淆矩陣可知道，正確的分類是f_{11}及f_{00}，錯誤的分類是f_{10}及f_{01}。此外，混淆矩陣也可以藉由計算其中的比值以表達不同模型的績效。

(1) 準確率(Accuracy)：計算整體模型預測的正確率。公式如(5.7)：

$$\text{Accuracy} = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad (5.7)$$

(2) 敏感率(Sensitivity)：亦稱為召回率(Recall)。計算模型在實際漲的股價中的預測正確率。公式如(5.8)：

$$\text{Sensitivity} = \text{Recall} = \frac{f_{11}}{f_{11}+f_{10}} \quad (5.8)$$

DOI:10.6814/NCCU201900623

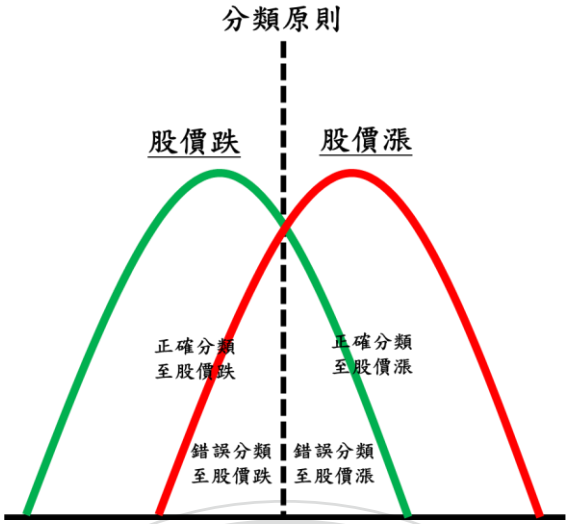

圖 12　股價漲跌之二元分類結果示意圖

表 7　漲跌二元分類問題之混淆矩陣

		預測分類	
		跌：類別「0」	漲：類別「1」
實際分類	跌：類別「0」	f_{00}	f_{01}
	漲：類別「1」	f_{10}	f_{11}

(3) 特異率(Specificity)：計算模型在實際跌的股價中的預測正確率。公式如(5.9)：

$$\text{Specificity} = \frac{f_{00}}{f_{01}+f_{00}} \quad (5.9)$$

以上三個績效指標，皆介於 0 到 1 之間的數值。越接近 1，表示模型在該績效指標中表現很好；反之越接近 0，表示模型在該績效指標中表現很差。

經過隨機森林模型建模與驗證，結果從表 8 至表 12 可觀察，在較短天期內之準確率雖有較高的趨勢，但也受限於樣本的不平衡(Imbalanced)，另外在表 11 與表 12 中有幾筆 NA 的數值是表示 1 日內要達到門檻維 1.5%及 2.0%的情況，此時資料平衡度已經相當偏差，因此不再作任何績效指標的評估。整體而言，對於長天期來說，AUC 整體有向上的趨勢，不過皆略大於 50%，模型效果有限。

DOI:10.6814/NCCU201900623

表 8　隨機森林模型績效(報酬率門檻 0%)

報酬率門檻 0%				
預測區間(日)	準確率	敏感率	特異率	AUC[7]
1	50.79%	54.19%	46.88%	50.70%
5	50.91%	62.81%	37.62%	51.04%
10	51.17%	62.17%	39.42%	51.42%
20	50.96%	60.42%	41.37%	52.37%

表 9　隨機森林模型績效(報酬率門檻 0.5%)

報酬率門檻 0.5%				
預測區間(日)	準確率	敏感率	特異率	AUC
1	69.04%	3.67%	97.26%	50.46%
5	51.83%	36.33%	66.69%	51.73%
10	51.31%	49.28%	55.63%	52.64%
20	48.46%	47.23%	53.21%	52.86%

表 10　隨機森林模型績效(報酬率門檻 1.0%)

報酬率門檻 1.0%				
預測區間(日)	準確率	敏感率	特異率	AUC
1	83.50%	1.31%	99.08%	50.20%
5	51.83%	36.33%	66.69%	51.73%
10	53.25%	36.35%	68.48%	53.27%
20	49.97%	40.03%	62.78%	53.79%

表 11　隨機森林模型績效(報酬率門檻 1.5%)

報酬率門檻 1.5%				
預測區間(日)	準確率	敏感率	特異率	AUC
1	NA	NA	NA	NA
5	69.94%	8.71%	94.10%	51.40%
10	59.72%	22.20%	83.43%	53.67%
20	55.45%	35.02%	69.74%	53.73%

[7] AUC 為 Area Under the Curve of ROC 之縮寫，用以評估模型之優劣標準。

DOI:10.6814/NCCU201900623

表 12　隨機森林模型績效(報酬率門檻 2.0%)

報酬率門檻 2.0%				
預測區間(日)	準確率	敏感率	特異率	AUC
1	NA	NA	NA	NA
5	77.32%	3.59%	97.06%	50.33%
10	66.15%	13.51%	89.99%	51.75%
20	56.02%	26.88%	76.56%	51.72%

二、混合模型績效評估

因混合模型需要兩層學習的步驟，故第一層的訓練集之起始點皆固定於 2000 年，隨著每次的學習皆會多加一年學習樣本。而第二層的訓練集(亦等同於第一層驗證集的結果)將皆以一年為學習區間，以驗證下一年的結果，如圖 13。

模型準確率表現的部分，以預測區間為 20 日、報酬率門檻為 2.0%為例(見圖 14)，經過不同的參數嘗試後，預測準確度約落在將近六成，與單純隨機森林而言相比，似無太大的改善。但可發現，模型的準確率整體是受到 08 年附近影響，加上許多特徵資料有時間遞延以及混合模型第二層學習皆為跌的緣故，故在 09 年模型準確率是最差的。然而，此模型雖然整體效果沒有太大改善，但以近五年來模型準確率有回穩，且可以逐漸爬升到六成以上，可推測此模型若無面對重大國際金融情勢的影響，準確率應可表現得不錯。

表 13 至表 17 則為在不同報酬率門檻下之模型預測績效。同樣地，在較短天期內之準確率雖有較高的趨勢，但也受限於樣本的不平衡(Imbalanced)，另外在表 16 與表 17 中有幾筆 NA 的數值是表示 1 日內要達到門檻維 1.5%及 2.0%的情況，此時資料平衡度已經相當偏差，因此不再作任何績效指標的評估。整體而言，混合模型的集成效果與單純之隨機森林模型相差不遠，故是否能在不同的混合模型組合提升準確率則為尚可探討之議題。

DOI:10.6814/NCCU201900623

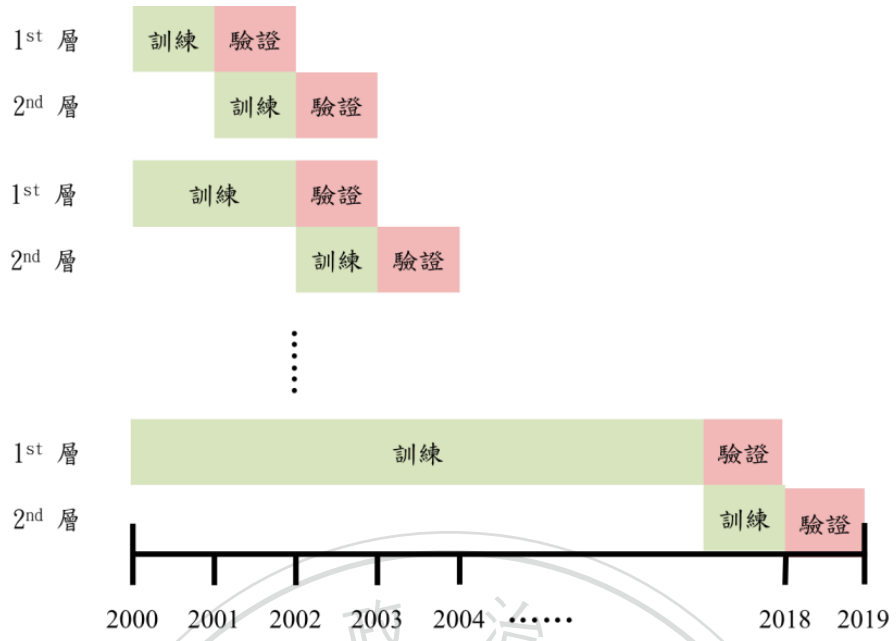

圖 13　混合模型之增長式視窗滾動法圖示

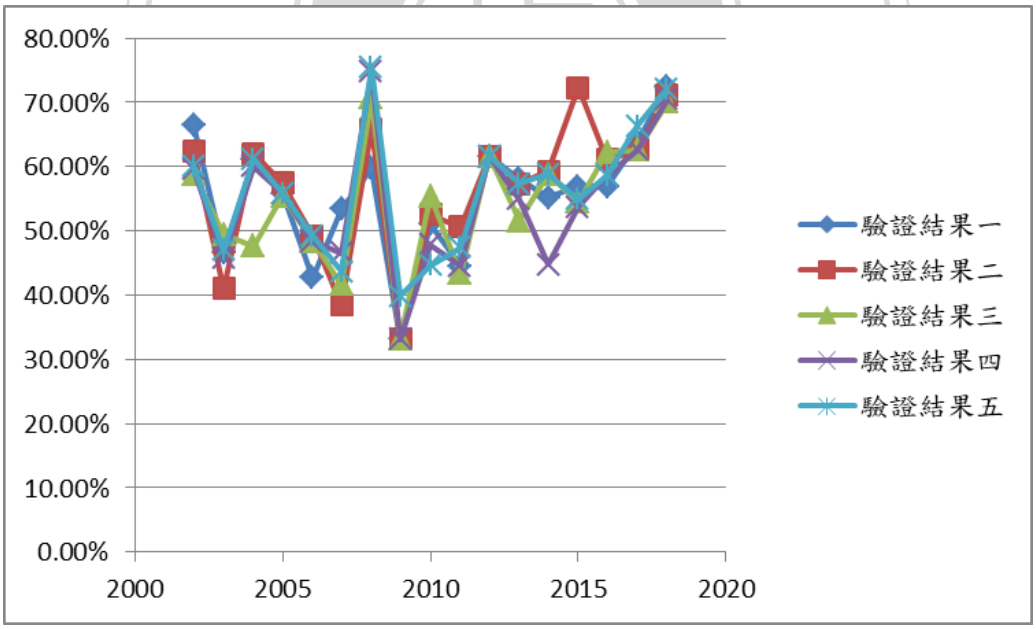

圖 14　混合模型準確率驗證(以預測區間 20 日、報酬率門檻 2.0%為例)

DOI:10.6814/NCCU201900623

表 13　混合模型績效(報酬率門檻 0%)

報酬率門檻 0%				
預測區間(日)	準確率	敏感率	特異率	AUC
1	52.05%	45.95%	58.05%	52.00%
5	46.46%	37.87%	62.70%	50.29%
10	44.33%	38.92%	60.54%	50.75%
20	43.94%	46.09%	52.36%	51.43%

表 14　混合模型績效(報酬率門檻 0.5%)

報酬率門檻 0.5%				
預測區間(日)	準確率	敏感率	特異率	AUC
1	70.30%	0.94%	99.65%	50.29%
5	51.18%	35.42%	65.51%	50.46%
10	49.38%	41.63%	59.94%	51.70%
20	44.73%	43.72%	57.67%	50.69%

表 15　混合模型績效(報酬率門檻 1.0%)

報酬率門檻 1.0%				
預測區間(日)	準確率	敏感率	特異率	AUC
1	82.44%	0.55%	99.72%	50.13%
5	59.63%	12.53%	87.50%	50.01%
10	53.21%	23.88%	77.26%	51.53%
20	50.20%	37.14%	68.49%	53.22%

表 16　混合模型績效(報酬率門檻 1.5%)

報酬率門檻 1.5%				
預測區間(日)	準確率	敏感率	特異率	AUC
1	NA	NA	NA	NA
5	68.42%	8.60%	92.59%	50.59%
10	59.01%	18.31%	84.45%	51.38%
20	51.93%	38.07%	65.47%	53.69%

DOI:10.6814/NCCU201900623

表 17　混合模型績效(報酬率門檻 2.0%)

報酬率門檻 2.0%				
預測區間(日)	準確率	敏感率	特異率	AUC
1	NA	NA	NA	NA
5	77.68%	2.77%	97.76%	50.27%
10	65.64%	5.14%	94.21%	49.68%
20	56.27%	27.19%	77.51%	52.35%

DOI:10.6814/NCCU201900623

第六章　結論與建議

本研究的貢獻在於提升台灣加權股票指數於未來數日內為上漲的準確率，而上漲的定義由預設的門檻值訂定。使用了台股技術指標、其他國家重要股市指數以及台灣總體經濟指標，共 192 個特徵。在資料處理層面，考量到資料時間取得的可行性，故有經過遞延處理；而因各個測量變數之指標單位不一，故也經過標準化處理。因蒐集之資料皆具有時間性，故使用增長式視窗滾動法(Increasing Window Rolling)以驗證模型績效表現。

使用集成學習的益處在於可不用過於擔心受到離群值的影響及可以藉由模型間平均或投票的結果以降低模型變異，甚至還可以處理非線性可分之結構資料。以單純隨機森林的結果顯示，雖在短天期的預測準確率高，但易受門檻訂定標準的影響，使得樣本呈現樣本分類失衡的現象；反之，在長天期的預測準確率較低，但對於門檻值也較為穩定，同時 AUC 指標也呈現較佳的表現。然而在提出的混合模型，在整體的模型效果並未獲得太大的改善，但細看模型於各年度的驗證表現，可以發現模型預測波動較為明顯的時期為國際金融事件之後，故若可避開這段時間而使用此模型，應能達到不錯的準確率效果。同時，在台灣總體經濟變數較為被動的訊號下，是否能可提供短天期的上漲趨勢判斷仍有待考量。

本研究在許多地方尚有探討與發展的空間，故提出下列數點以供後續研究的參考方向：

(一) 技術指標的參數選取上，本研究皆以經驗法則挑選是否具有上漲的訊號。故對於任一模型而言，是否為最佳參數則需以各種組合作為嘗試。

(二) 台灣總體經濟指標資料一直存有公布時間遞延的問題，會使得資訊無法與現實中相符。若在預測當下，景氣從壞轉好，已有復甦跡象，但因遞延的緣故，仍顯示在景氣不好的跡象，此時會導致錯過良好看漲時機。因此，是否能提出一個在總體經濟對於未來預測的模型可能為一種解決方法。同時，每個月皆只有一個數值的情況，亦較無法立即反映現況。

DOI:10.6814/NCCU201900623

(三) 提出之混合模型為一種藉由各個模型結果的集合，可以嘗試不同的模型應該可以有不同的效果。

(四) 對於投資人而言，最重要的是如何在實務上得到良好的投資績效以及較低的波動，若可將模型代入實務的應證上，配合交易策略，更可表現出模型的實質效用。

(五) 在此研究選取的特徵大多為較長時間的指標。特徵選取上可更為多樣化，例如消息面、情緒面資訊，可更及時反應股價的走勢。

DOI:10.6814/NCCU201900623

參考文獻

Ahmed Imran Hunjra, Muhammad Irfan Chani, Muhammad Shahzad, Muhammad Farooq and Kamran Khan (2014). “The Impact of Macroeconomic Variables on Stock Prices in Pakistan,” International Journal of Economics and Empirical Research, 2(1), 13-21.

Allan Timmermann and Clive William John Granger (2004). “Efficient Market Hypothesis and Forecasting,” International Journal of Forecasting, 20(1), 15-27.

Amith Vikram Megaravalli, Gabriele Sampagnaro and Louis Murray (2018). “Macroeconomic Indicators and Their Impact on Stock Markets in ASIAN 3: A Pooled Mean Group Approach,” Cogent Economics and Finance, 6, 1-14.

Berninger, Jordan (2018). “Forecasting the Time Series of Apple Inc.'s Stock Price,” UCLA Electronic Theses and Dissertations.

Christopher N. Avery, Judith A. Chevalier and Richard J. Zeckhauser (2016)."The "CAPS" Prediction System and Stock Market Returns," Review of Finance, European Finance Association, 20(4), 1363-1381.

Depei Bao and Zehong Yang (2008). “Intelligent Stock Trading System by Turning Point Confirming and Probabilistic Reasoning,” Expert Systems with Applications, 34(1), 620-627.

Eugene F. Fama (1970). “Efficient Capital Markets: A Review of Theory and Empirical Work,” The Journal of Finance, 25(2), 383-417.

Felipe Giacomel, Renata Galante and Adriano Pareira (2015). “An Algorithmic Trading Agent Based on A Neural Network Ensemble: A Case of Study in North American and Brazilian Stock Markets,” IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

DOI:10.6814/NCCU201900623

Haoming Li, Tianlun Li and Zhijun Yang (2014). “Algorithmic Trading Strategy Based on Massive Data Mining,” Stanford University.

Jan Ivar Larsen (2010). “Predicting Stock Prices Using Technical Analysis and Machine Learning,” Thesis, Norwegian University of Science and Technology.

Jawad Khan and Imran Khan (2018). “The Impact of Macroeconomic Variables on Stock Prices: A Case Study of Karachi Stock Exchange,” Journal of Economics and Sustainable Development, 9(13), 15-25.

Joseph Tagne Talla (2013). “Impact of Macroeconomic Variables on the Stock Market Prices of the Stockholm Stock Exchange (OMXS30),” Master´s Thesis within International Financial Analysis.

K. Nirmala Devi, V. Murali Bhaskaran and G. Prem Kumar (2015). “Cuckoo Optimized SVM for Stock Market Prediction,” IEEE Sponsored 2nd International Conference on Innovations in Information, Embedded and Communication systems (ICJJECS).

Leo Breiman (1994). “Bagging Predictors,” Machine Learning 26(2), 123-140.

Luckyson Khaidem, Snehanshu Saha and Sudeepa Roy Dey (2016). “Predicting the Direction of Stock Market Prices Using Random Forest,” arXiv preprint arXiv:160500003.

Ludmila Kuncheva and Chris Whitaker (2003). “Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy,” Machine Learning 51(2), 181-207.

Michael Jensen (1978). “Some Anomalous Evidence Regarding Market Efficiency,” Journal of Financial Economics, 6, Nos. 2/3 95-101.

Ramazan Gencay (1999). “Linear, Non-Linear and Essential Foreign Exchange Rate Prediction with Simple Technical Trading Rules,” Journal of International Economics 47(1), 91-107.

DOI:10.6814/NCCU201900623

Segal and Mark R (2004). “Machine Learning Benchmarks and Random Forest Regression,” Center for Bioinformatics and Molecular Biostatistics, UC, San Francisco, California.

Snehanshu Saha, Swati Routh and Bidisha Goswami (2014). “Modeling Vanilla Option Prices: A Simulation Study by An Implicit Method,” Journal of advances in Mathematics, 6(1), 834-848.

Suryoday Basak, Saibal Kar, Snehanshu Saha, Luckyson Khaidem and Sudeepa Roy Dey (2019). “Predicting the Direction of Stock Market Prices Using Tree-Based Classifiers,” The North American Journal of Economics and Finance, Volume 47, 552-567.

Xinjie (2014). “Stock Trend Prediction with Technical Indicators Using SVM,” Stanford University.

Yoav Freund and Robert E. Schapire (1996). “Experiments with a New Boosting Algorithm,” Machine Learning: Proceedings of the Thirteenth International Conference, 148-156.

Yuqing Dai and Yuning Zhang (2013). “Machine Learning in Stock Price Trend Forecasting,” Stanford University.

DOI:10.6814/NCCU201900623

附錄

附錄 1　MSE 之期望值分解推導

已知 $E\{MSE\} = E\left\{\frac{1}{n}\sum_{i=1}^{n}\left(y_i' - \hat{f}(x_i')\right)^2\right\} = \frac{1}{n}\sum_{i=1}^{n} E\{[y_i' - \hat{f}(x_i')]^2\}$

其中，期望值的部分可展開為：

$$E\left[y_i' - \hat{f}(x_i')\right]^2$$

$$= E\left[y_i' - f(x_i') + f(x_i') - \hat{f}(x_i')\right]^2$$

$$= E\{\left(y_i' - f(x_i')\right)^2\} + E\{\left(f(x_i') - \hat{f}(x_i')\right)^2\} + 2E\left\{\left(y_i' - f(x_i')\right)\left(f(x_i') - \hat{f}(x_i')\right)\right\}$$

$$= E\{\epsilon^2\} + E\left(f(x_i') - \hat{f}(x_i')\right)^2$$

$$+2(E\{y_i' \times f(x_i')\} - E\{f(x_i')^2\} - E\{y_i' \times \hat{f}(x_i')\} + E\{f(x_i') \times \hat{f}(x_i')\})$$

$$= E\{\epsilon^2\} + E\{\left(f(x_i') - \hat{f}(x_i')\right)^2 + 2({y_i'}^2 - {y_i'}^2 - E\{y_i' \times f(x_i')\} + E\{y_i' \times f(x_i')\})$$

$$= E\{\epsilon^2\} + E\left\{\left(f(x_i') - \hat{f}(x_i')\right)^2\right\} + 0$$

$$= E\{\epsilon^2\} + E\{\left(f(x_i') - E\left(\hat{f}(x_i')\right) + E\left(\hat{f}(x_i')\right) - \hat{f}(x_i')\right)^2\}$$

$$= E\{\epsilon^2\} + E\left\{\left(f(x_i') - E\left(\hat{f}(x_i')\right)\right)^2\right\} + E\left\{\left(E\left(\hat{f}(x_i')\right) - \hat{f}(x_i')\right)^2\right\}$$

$$+2E\left\{(f(x_i') - E\left(\hat{f}(x_i')\right)) \times \left(E\left(\hat{f}(x_i')\right) - \hat{f}(x_i')\right)\right\}$$

$$= E\{\epsilon^2\} + E\left\{\left(f(x_i') - E\left(\hat{f}(x_i')\right)\right)^2\right\} + \left\{\left(E\left(\hat{f}(x_i')\right) - \hat{f}(x_i')\right)^2\right\}$$

$$+2(E\left\{f(x_i') \times E\left(\hat{f}(x_i')\right)\right\} - E\{f(x_i') \times \hat{f}(x_i')\}$$

$$-E\left\{E\left(\hat{f}(x_i')\right)^2\right\} + E\{E\left(\hat{f}(x_i')\right) \times \hat{f}(x_i')\})$$

$$= E\{\epsilon^2\} + E\left\{\left(f(x_i') - E\left(\hat{f}(x_i')\right)\right)^2\right\} + \left\{\left(E\left(\hat{f}(x_i')\right) - \hat{f}(x_i')\right)^2\right\}$$

$$+2\{f(x_i') \times E\left(\hat{f}(x_i')\right) - f(x_i') \times E\left(\hat{f}(x_i')\right) - E\left(\hat{f}(x_i')\right)^2 + E\left(\hat{f}(x_i')\right)^2\}$$

DOI:10.6814/NCCU201900623

$$= E\{\epsilon^2\} + E\left\{\left(f(x_i') - E\left(\hat{f}(x_i')\right)\right)^2\right\} + \left\{\left(E\left(\hat{f}(x_i')\right) - \hat{f}(x_i')\right)^2\right\} + 0$$

$$= VAR(\varepsilon) + [Bias(y_i')]^2 + VAR(y_i')$$

故$E\{MSE\} = \text{Average of}\{\ VAR(\varepsilon) + [Bias(y_i')]^2 + VAR(y_i')\}$

DOI:10.6814/NCCU201900623

附錄 2　台灣總體經濟指標選取列表

分類	指標名稱	單位
人口就業薪資統計	台灣-平均每月工時-製造業	小時
	台灣-總人口數	千人
	台灣-就業人口數	千人
	台灣-勞動人口數	千人
	台灣-失業人口數	千人
	台灣-勞動參與率	%
	台灣-失業率	%
	台灣-就業人口-農林漁牧業	千人
	台灣-就業人口-工業	千人
	台灣-就業人口-服務業	千人
	台灣-就業人口-非農業部門	千人
	台灣-受僱人數	人
	台灣製造業受僱員工平均薪資	元
	台灣-平均工時-製造業	小時
	台灣-每人每月薪資-各行業	元
	台灣-經常性薪資-各行業	元
批發零售商業營業額	台灣-營業額-批發業	百萬元
	台灣-營業額-零售業	百萬元
	台灣-營業額-綜合零售業	百萬元
	台灣-營業額-零售業-百貨公司	百萬元
	台灣-營業額-超級市場	百萬元
	台灣-營業額-連鎖式便利商店	百萬元
	台灣-營業額-量販店	百萬元

DOI:10.6814/NCCU201900623

分類	指標名稱	單位
批發零售商業營業額(續)	台灣-營易額-批發業-機械器具	百萬元
	台灣-營業額-批發業-食品飲料及菸草製品	百萬元
	台灣-營業額-批發業-五金及家庭日常用品	百萬元
	台灣-營業額-批發業-藥類化妝品	百萬元
	台灣-營業額-零售業-家庭器具及用品	百萬元
	台灣-營業額-零售業-藥品及化妝品	百萬元
	台灣-營業額-布疋及服飾品	百萬元
	台灣-營業額-零售業-車輛及車輛零件	百萬元
	台灣-營業額-餐飲業	百萬元
	台灣-營業額指數-批發業	指數
	台灣-營業額指數-零售業	指數
	台灣-營業額指數-餐飲業	指數
	台灣-營業額指數-綜合商品零售業	指數
景氣指標	台灣-存貨率-製造業	%
	台灣-景氣對策信號綜合分數	分數
	台灣-痛苦指數	指數
	台灣-領先指標綜合指數	指數
	台灣-落後指標綜合指數	指數
	台灣-平均薪資-製造業(經季節調整)	元
	台灣股價指數變動率-月均值	%
	台灣-CPI 總指數	指數
證券統計數據	台灣-證券市場-股票-上市股數	百萬股
	台灣-證券市場-股票-資本總額	百萬元
	台灣-證券市場-股票-總市值	百萬元

DOI:10.6814/NCCU201900623

分類	指標名稱	單位
證券統計數據(續)	台灣-證券市場-股票-成交股數	百萬股
	台灣-證券市場-股票-成交金額	百萬元
	台灣-平均每營業日股票成交股數	百萬股
	台灣-平均每營業日股票成交金額	百萬元
	台灣-證券市場-債券-成交金額	百萬元
	台灣-本益比-上市公司-大盤	倍
工業產銷用電量	台灣-IPI 總指數	指數
物價統計數據	台灣-WPI	指數

DOI:10.6814/NCCU201900623

附錄 3　技術指標特徵列表(離散型)

技術指標類型	特徵名稱	說明[8] (符合條件之觀測值設為 1)
擺動類指標	AR_signal (n=25)	(AroonUp > AroonDn) & (AroonUp > 70)
	BBands_signal	pctB > 1
	KD_signal	fastK > fastD
趨勢類指標	MACD_signal (12,26, 9)	(macd > 0) & (signal > 0) & (macd > signal)
	ADX_signal	(DIp > DIn) & (DIp > ADX) & (DIn > ADX)
	MA_signal1	MA5 > MA20
	MA_signal2	MA5 > MA60
	MA_signal3	MA20 > MA60
	EMA_signal1	EMA5 > EMA20
	EMA_signal2	EMA5 > EMA60
	EMA_signal3	EMA20 > EMA60
	EVWMA_signal1	EVWMA5 > EVWMA20
	EVWMA_signal2	EVWMA5 > EVWMA60
	EVWMA_signal3	EVWMA20 > EVWMA60
	MA5_slope	MA5 > 1.05×lag(MA5)
	MA20_slope	MA20 > 1.05×lag(MA20)
	MA60_slope	MA60 > 1.05×lag(MA60)
動量指標	CCI_signal	CCI > 100
	EMV_signal	EMV > 0
	MFI_signal	MFI > 70
	RSI_signal	(RSI14 < 20) & (RSI60 < 20) & (RSI14 < RSI60)
量能指標	OBV_signal	OBV > 1.05×lag(OBV)
	Chai_signal	Chai < 0

[8] 說明欄位中的變數，除了收盤價(close)之外，皆使用 R 語言 TTR 套件之函數。

DOI:10.6814/NCCU201900623

技術指標特徵列表(連續型)

技術指標類型	特徵名稱	說明 (數值經標準化處理)
擺動類指標	AR (n=25)	oscillator
	BBands	pctB
	KD (fastK)	fastK
	KD (fastD)	fastD
	KD (slowD)	slowD
趨勢類指標	MACD	MACD
	MACD (signal)	signal
	ADX (DIp)	DIp
	ADX (DIn)	DIn
	ADX (DX)	DX
	ADX	ADX
	MA5	MA5
	MA20	MA20
	MA60	MA60
	EMA5	EMA5
	EMA20	EMA20
	EAM60	EAM60
	EVWMA5	EVWMA5
	EVWMA20	EVWMA20
	EVWMA60	EVWMA60
	VolMA5	VolMA5
	P_MA5	MA5÷close
	P_MA20	MA20÷close

DOI:10.6814/NCCU201900623

技術指標特徵列表(連續型(續))

技術指標類型	連續型特徵名稱	說明 (數值經標準化處理)
趨勢類指標	P_MA60	MA60÷close
(續)	MA5_MA20	MA5÷MA20
	MA5_MA60	MA5÷MA60
	MA20_MA60	MA20÷MA60
動量指標	CCI	CCI
	EMV	EMV
	EMV (MAEMV)	MAEMV
	MFI	MFI
	RSI14	RSI14
	RSI60	RSI60
量能指標	Chai	Chai
	OBV	OBV

DOI:10.6814/NCCU201900623