機器學習運用在股價漲跌預測
—以台灣加權指數為例

第 6 組

劉昱辰、徐緯澔、童思謙
傅宗皓、趙佑霖、姜品威、文詳惠

目錄

內容

壹、摘要

本研究旨在探討臺灣股市中常用技術指標對於臺灣加權指數（TAIEX）漲跌之預測能力，研究中比較了四種預測模型：極限梯度提升（XGBoost）、隨機森林（Random Forest）、類神經網路（ANN）、單純貝氏（Naïve Bayes），並對這些模型使用兩種變數輸入方式進行比較，第一種輸入數據方式是使用股票交易數據（開盤價、最高價、最低價、收盤價）計算十種技術指標（連續型變數），第二種方式是將這些技術指標轉換為確定趨勢參數（離散型變數）。實證結果是基於 2013 年至 2023 年的歷史數據，而實證結果表明：上述四種模型對於臺灣加權指數之漲跌皆沒有良好的預測能力，轉換為確定趨勢參數後，預測能力也未有顯著提升。

貳、簡介及文獻探討

股票價格走勢預測一直是學術界和金融業中有趣且具有挑戰性的課題。隨著資訊技術的進步，仍被認為是時間序列預測中最具挑戰性的應用之一。本研究的核心目的是利用機器學習技術來提高股票價格走勢預測的準確性，通過詳細研究台灣股票市場，驗證和展示股票價格走勢預測的可行性，並提供市場參與者更有效的交易策略。

鑒於台股屬於新興市場，我們選擇參照一篇印度市場的研究(Patel et al., 2015)，他們提出了一種趨勢確定數據(trend deterministic)處理方法，並將其應用於多種機器學習技術中來預測股價的漲跌，包括人工神經網絡（ANN）、支持向量機（SVM）、隨機森林（Random Forest）和單純貝氏（Naive-Bayes）。

在人工神經網絡（ANN）方面，Patel et al. (2015)的研究證實了其在預測股票價格回報和走向上的優越性。其他研究也表明，ANN 在金融建模和預測方面具有很強的能力（Avcı, 2007；Chen et al., 2003；Kara et al., 2011；Karaatli et al., 2005；Olson and Mossman, 2003）。單純貝氏（Naive-Bayes）分類器假設類別條件獨立性，利用貝氏定理來計算數據屬於特定類別的機率。雖然單純貝氏模型在某些情況下表現不如其他複雜的模型，但 Patel et al. (2015)研究表明，通過使用確定趨勢數據，單純貝氏模型的預測性能也得到了顯著提升。隨機森林（Random Forest）則通過構建多個決策樹來進行分類和回歸，其在處理複雜數據集方面具有顯著優勢。Patel et al. (2015)的研究表明，隨機森林在使用趨勢確定數據時，預測精準度顯著提高。隨機森林的這一特性使其成為股票市場預測中一種強有力的工具。

總結來說，Patel et al. (2015)的研究強調了技術指標在股票市場預測中的重要性，並證明了通過確定趨勢數據方法可以顯著提升機器學習模型的預測能力。這一發現對於制定更有效的市場交易策略具有重要意義，並為未來的研究提供了新的方向。我們將參照 Patel et al. (2015)所用的十個技術指標與數據方式進行台股市場的實證，使用四個機器學型模型(XGBoost、Random Forest、ANN、Naïve-Bayes)來預測台灣加權平均指數的漲跌。

參、研究方法及資料

一、資料介紹

本研究的資料取自台灣經濟新報(TEJ)資料庫，取樣期間從 2013 至 2023 年，本研究取日資料作為實證研究之頻率，資料樣本共 2691 筆日資料。

二、變數介紹

本研究探討之自變數有以下十項技術指標：簡單移動平均 (SMA)、加權移動平均 (WMA)、動能 (Momentum)、隨機指標 K 值 (Stochastic K%)、隨機指標 D 值 (Stochastic D%)、威廉指標 R 值 (Larry William's R%)、指數平滑異同移動平均線(MACD)、相對強弱指標 (RSI)、ADO 聚散震盪指標 (A/D Oscillator)及順勢指標 (CCI)，自變數相關之運算公式詳見下表 （表 3-1），研究中，我們會將自變數分成連續型資料及離散型資料兩種資料形式，以進行後續的模型應用與分析。

表 3-1、 選用變數與變數運算公式

變數名稱	運算公式
SMA	$\frac{C_t + C_{t-1} + \cdots + C_{t-9}}{10}$
WMA	$\frac{n * C_t + (n-1) * C_{t-1} + \cdots + C_{t-9}}{n + (n-1) + \cdots + 1}$
Momentum	$C_t - C_{t-9}$
Stochastic K%	$\frac{C_t - LL_{t-(n-1)}}{HH_{t-(n-1)} - LL_{t-(n-1)}}$
Stochastic D%	$\frac{\sum_{i=0}^{n} K_{t-i}}{10}$
Larry William's R%	$\frac{H_n - C_t}{H_n - L_n} * 100$
MACD	$MACD(n)_{t-1} + \frac{2}{n+1} * (DIFF_t - MACD(n)_{t-1})$
RSI	$10 - \frac{100}{1 + (\frac{\sum_{i=0}^{n-1} UP_{t-i}}{n}) / (\frac{\sum_{i=0}^{n-1} DW_{t-i}}{n})}$
A/D Oscillator	$\frac{H_t - C_{t-1}}{H_t - L_{t-1}}$
CCI	$\frac{M_t - SM_t}{0.015 D_t}$

C_t, H_t, L_t分別代表在時間點 t 時的收盤價、最高價及最低價，$DIFF_t = EMA(12)_t - EMA(26)_t$，EMA 表指數平滑移動平均線，$EMA(k)_t = EMA(\mathrm{k})_{t-1} - \alpha * \left(C_t - EMA(k)_{t-1}\right)$，$\alpha$為平滑因子，$HH_t, LL_t$分別表示過去 t 天的最高及最低價，$M_t = \frac{H_t + L_t + C_t}{3}$, $\mathrm{SM}_t = \frac{\sum_{i=1}^{n} M_{t-i+1}}{n}$, $D_t = \frac{\sum_{i=1}^{n} |M_{t-i+1} - SM_t|}{n}$，$UP_t$表在 t 時點價格上漲，$DW_t$表在 t 時點價格下跌。

連續型資料

由表 3-1 中的運算公式，可以發現十項技術指標自變數已經為連續型資料，而本研究中，我們對這十項自變數做歸一化處理(Min-Max Normalization)，將所有自變數壓縮至$(-1, 1)$之間，以防止任一自變數數值過大，對於後續模型的預測能力有壓到性的影響力。

離散型資料

表 3-1 中各變數計算出的數值，本研究遵照 Patel et al. (2015)中的分類標準對各個自變數進行處理，將各變數資料調整為"-1"或"1"，其中"-1"代表預期明日股價將會下跌，而"1"代表預期明日股價將會上漲，以下為各變數的調整標準：

本研究使用 10 天簡單移動平均線（SMA）和加權移動平均線(WMA)來判斷趨勢方向。當收盤價高於移動平均線時，表示趨勢上漲（"1"）；反之則表示趨勢下跌（"-1"）。

MACD、隨機指標、威廉指標等技術指標，其數值上升通常代表股價可能上漲；數值下降則可能代表股價下跌。本文將這些指標在當前時刻（t）和前一時刻（t-1）的數值進行比較，如果當前數值高於前一數值，則表示趨勢上漲（"1"）；反之則表示趨勢下跌（"-1"）。

RSI 指標介於 0 到 100 之間，用於判斷股票是否超買超賣。RSI 高於 70 代表超買，可能下跌（"-1"）；RSI 低於 30 代表超賣，可能上漲（"1"）。若 RSI 在 30 到 70 之間，則比較當前（t）和前一（t-1）的 RSI 值，當前值高於前一值則表示趨勢上漲（"1"）；反之則表示趨勢下跌（"-1"）。

CCI 指數用於衡量股價波動幅度。CCI 大於 200 表示可能超買，未來可能下跌（"-1"）；CCI 小於 -200 表示可能超賣，未來可能上漲（"1"）。CCI 在 -200 到 200 之間，如果時間點"t"的值高於時間點"t−1"的值，則代表未來趨勢可能上漲("1")；反之則代表趨勢可能下跌("-1")。

三、模型介紹

本研究採用四種機器學習模型進行股票漲跌之預測：極限梯度提升（XGBoost）、隨機森林（Random Forest）、類神經網路（ANN）、單純貝氏（Naive Bayes），以下將針對各模型作詳細說明。

1. 極限梯度提升（XGBoost）

極限梯度提升模型透過訓練多個弱學習器，逐步提升模型的預測能力，並使用二階近似方法，使得模型在優化過程中提高精確度。此外，本方法透過正規化標準來防止模型過度擬合，提高模型的泛化能力，日後能更廣泛地應用於預測未來數據上。關於本方法的參數設置見表 3-2。

表 3-2、XGBoost 的參數設置

參數名稱	設置參數作用
booster	指定欲使用的模型類型，如：gbtree、gblinear
objective	指定欲優化的目標函數，如：binary、multi
eta	設定模型的學習率，通常介於 0.01 到 0.3 之間
gamma	設定懲罰項，控制節點分裂的最小損失減益
max_depth	設定樹的最大深度，防止過度擬合

XGBoost 適用於處理大規模數據集和特徵數量多的情境，如金融風控、醫療診斷和營銷預測等高精度應用領域。其優勢在於高效能、運算速度快，以及具有良好的預測精度。

2. 隨機森林（Random Forest）

隨機森林模型最主要的運作原理為 Bagging，採用取後放回的方式建立資料子集，並用不同的資料子集建立森林決策樹。並採用 Bootstrap 的方式對樣本以及特徵作取後放回的抽樣，以此建立一棵棵的決策樹。在兩個隨機因子之下，降低本方法產生過度配飾的機率。當建構好決策樹之後，最終以多數決投票的方式決定模型判定結果。若樣本資料為離散型，採取個別決策樹結果的眾數；若樣本資料為連續型，則採取個別決策樹結果的平均值。關於本方法的參數設置見表 3-3。

表 3-3、隨機森林的參數設置

參數名稱	設置參數作用
n_estimators	森林中樹木的數量
max_depth	設定樹的最大深度，防止過度擬合

隨機森林模型適用於處理離散值的分類與連續值的迴歸問題。其優勢在於，可處理高維度的特徵資料，以及在兩種隨機因子的抽取下，有效防止過擬合的狀況產生。另外須特別留意，過多決策樹將導致計算成本如時間和空間成本的提高，若樣本資料本身雜訊過多，依舊會讓隨機森林模型出現過度配飾的結果。

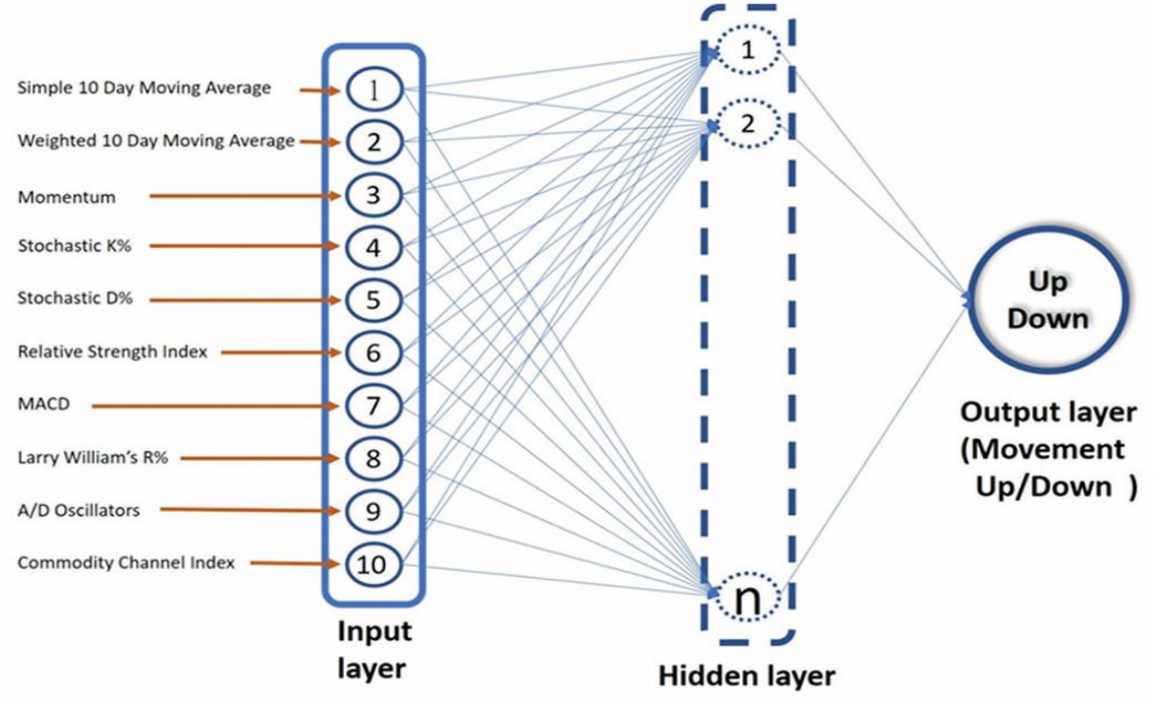
3. 類神經網路（ANN）

多層神經網路模型（MLP）是一種模仿生物神經系統運作的計算模型，由多層神經元所組成，包含輸入層、隱藏層以及輸出層。每一層透過權重連接，形成複雜的網絡結構，其運作機制見表 3-4 和圖 3-1。

表 3-4、類神經網路模型的結構組成

名稱	機制
輸入層	接收大量原始數據，輸入層的神經元數量等於輸入特徵的數量，每個神經元將輸入值傳遞給下一層
隱藏層	每個隱含層由多個神經元組成，可設定為一至多層，位於輸入層與輸出層之間，每個隱含層的神經元接受來自前一層的輸出信號，經加權求和以及激活函數的非線性轉換，將結果傳遞至下一層
輸出層	在神經元鏈結中傳輸、分析，輸出最終的預測結果

圖 3-1、類神經網路模型運作機制示意圖



4. 單純貝氏（Naïve Bayes）

單純貝氏模型是以貝式定理為基礎的分類算法，透過機率統計判斷未知的資料類別，核心概念是假設特徵之間相互獨立，計算每個類別的後驗機率，並進行分類。透過表 3-5 的參數設定，本方法能夠快速計算出樣本屬於某類別的機率，根據最大後驗機率原則進行分類。

後驗機率： $\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i|y)$

表 3-5、單純貝氏模型的參數設置

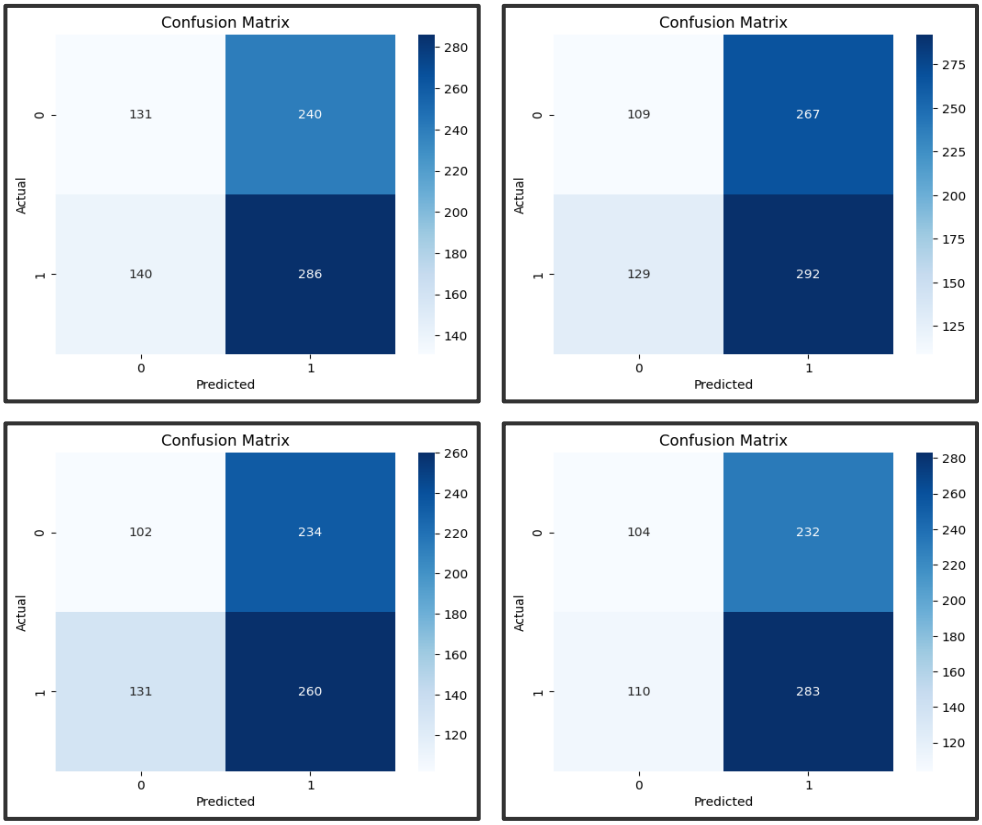
參數名稱	設置參數作用
條件機率	$P(A\|B) = \frac{P(B\|A)\cdot P(A)}{P(B)}$ 描述在給定類別的情況下，特徵取某一值的可能性
先驗機率	在沒有任何特徵信息時，各類別的初始機率

肆、實證結果

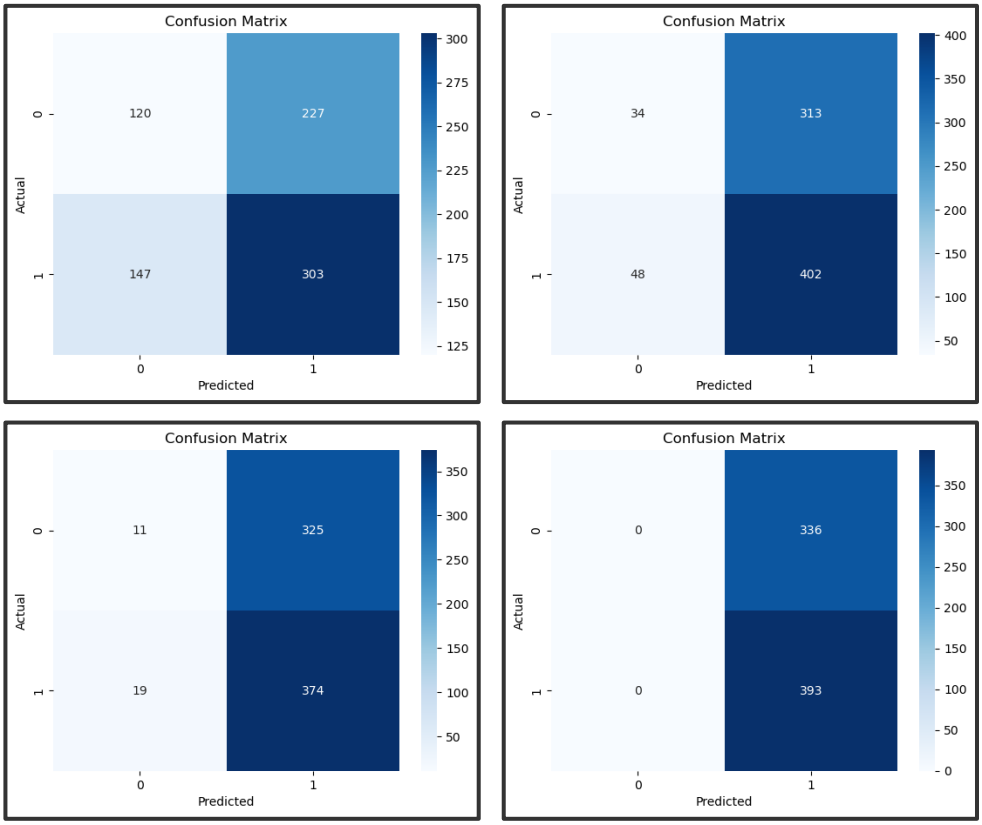
本組遵照前述所提及之研究方式，建構出十大研究指標，並使用四種機器學習模型預測臺灣加權指數的每日漲跌。模型的評估方式則採用 AUC (Area Under Curve)以及 accuracy score。AUC 為 ROC (receiver operating characteristic curve)曲線下的面積，accuracy score = (TP+TN)/total data。AUC 與 accuracy score 越接近 1，其模型效能越好，反之則越低。

一、混淆矩陣

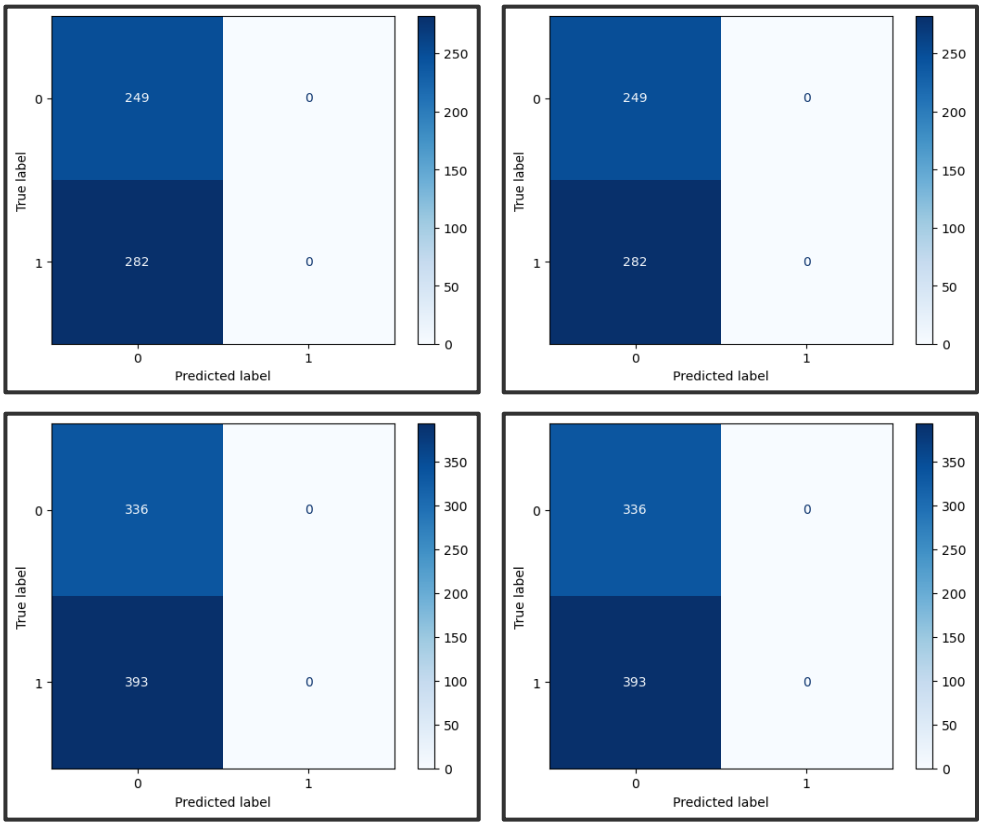
1. XGBoost



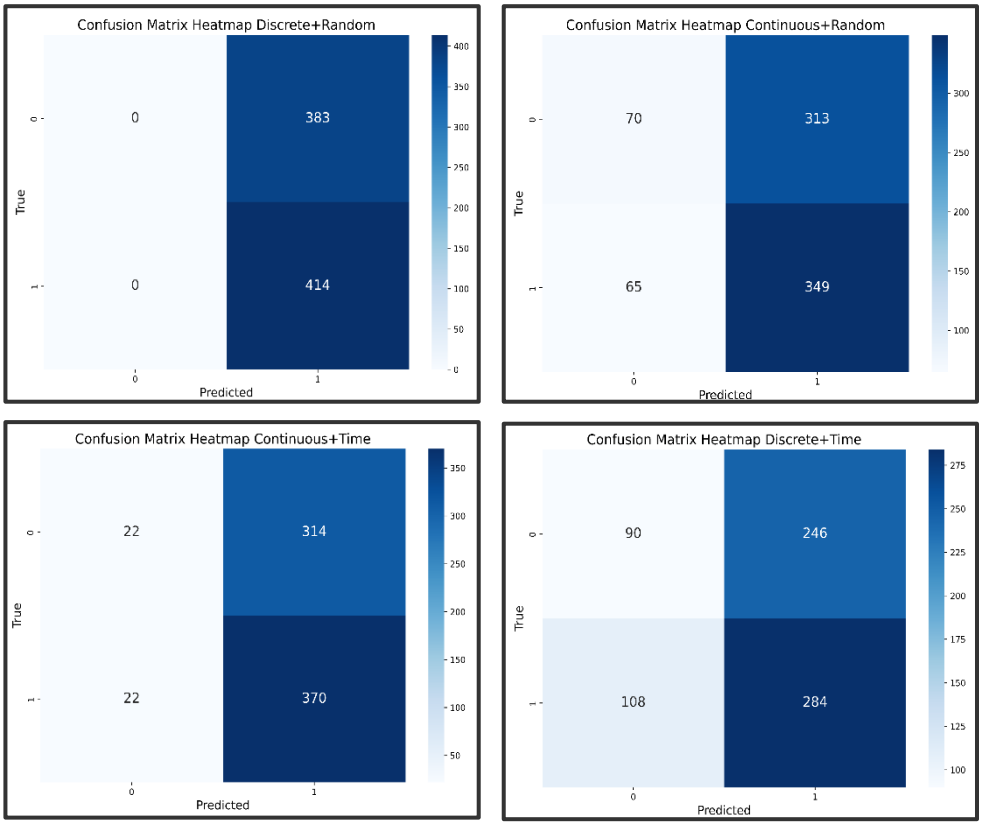
2. Random Forest



3. ANN



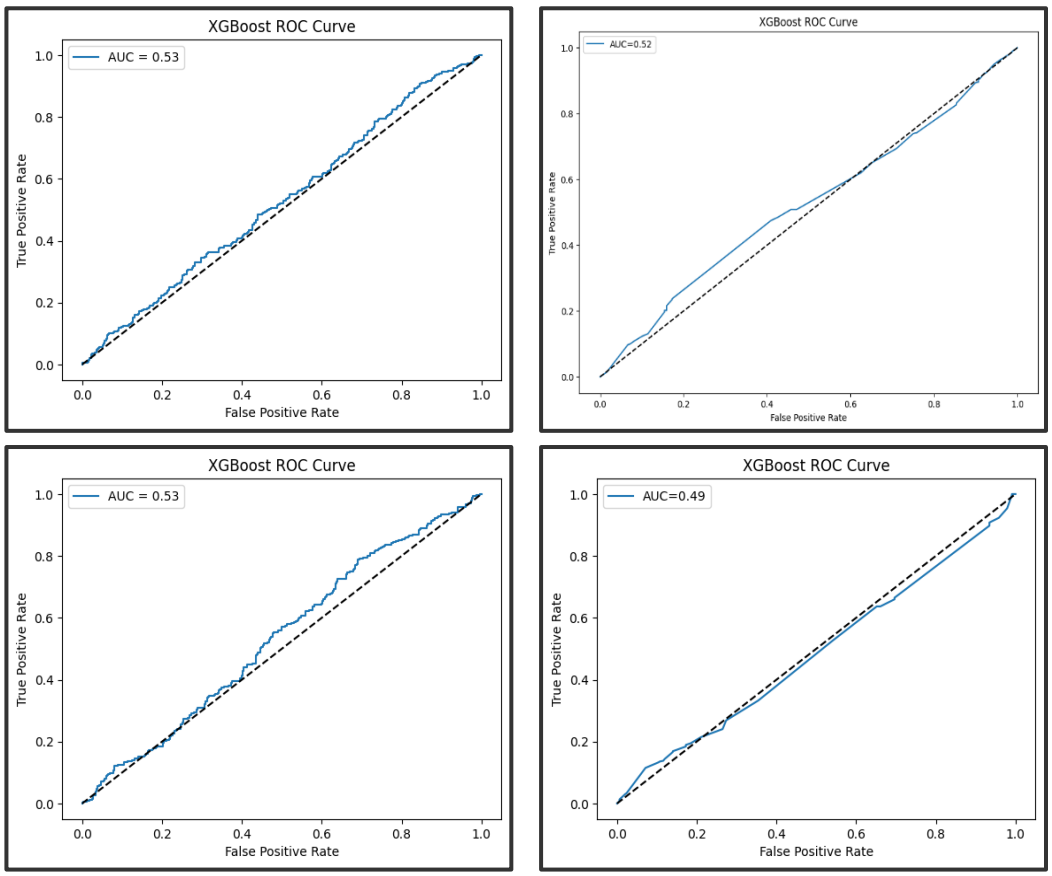
4. Naïve Bayes



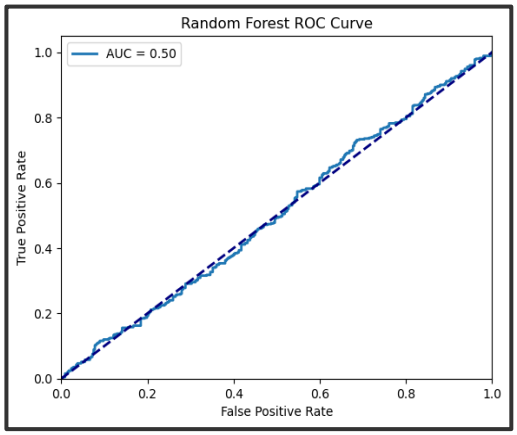
從上面四種模型的混淆矩陣可以看出，accuracy 幾乎都介於 0.5 上下。無論是資料是離散或連續型態、隨機分組或是時間序列分組。這代表模型的預測能力低落，這點同樣也能在 ROC 看出來。值得一提的是，在 ANN 模型及部分 Random Forest 與 Naïve Bayes 的模型中，出現了模型預測為全漲或全跌的現象。

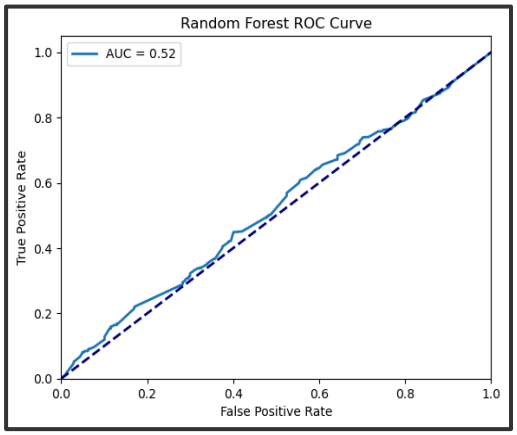
二、ROC 曲線

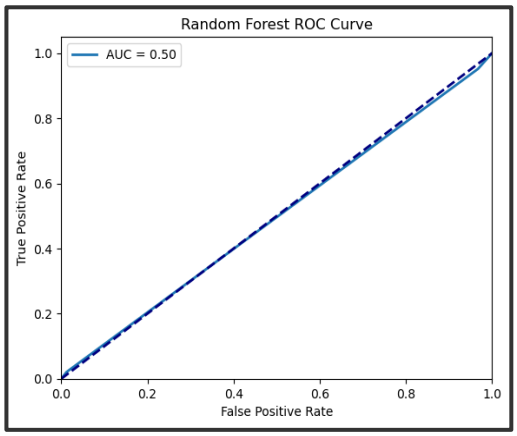
1. XGBoost

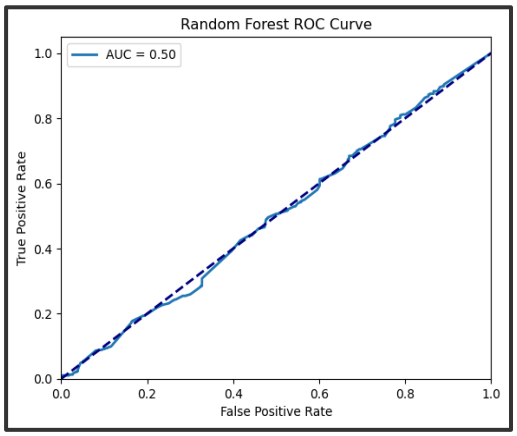


2. Random Forest

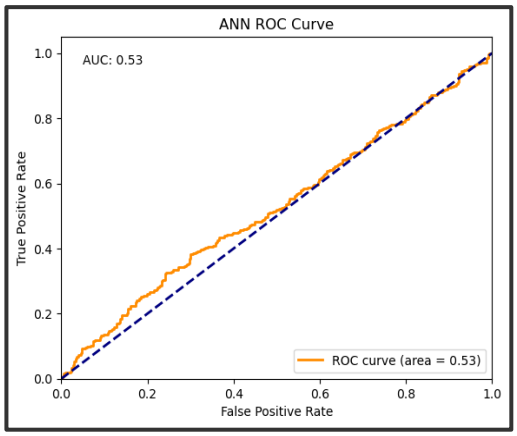

Random Forest ROC Curve
AUC = 0.50
True Positive Rate
False Positive Rate

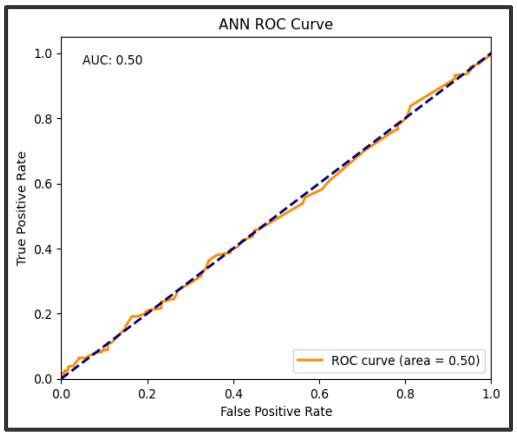

Random Forest ROC Curve
AUC = 0.52
True Positive Rate
False Positive Rate

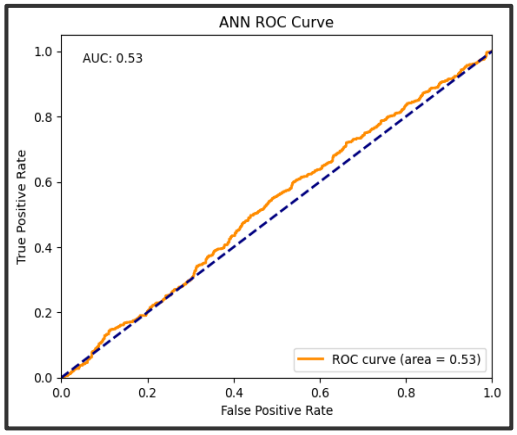

Random Forest ROC Curve
AUC = 0.50
True Positive Rate
False Positive Rate

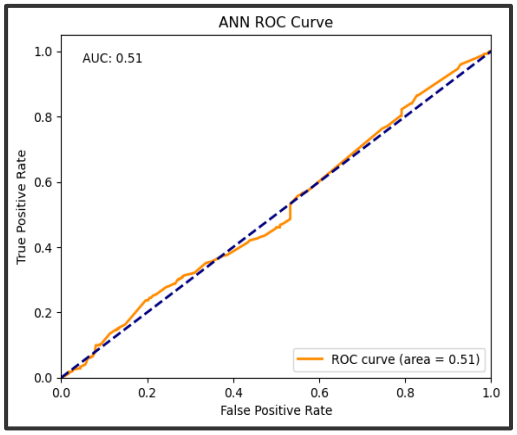

Random Forest ROC Curve
AUC = 0.50
True Positive Rate
False Positive Rate

3. ANN

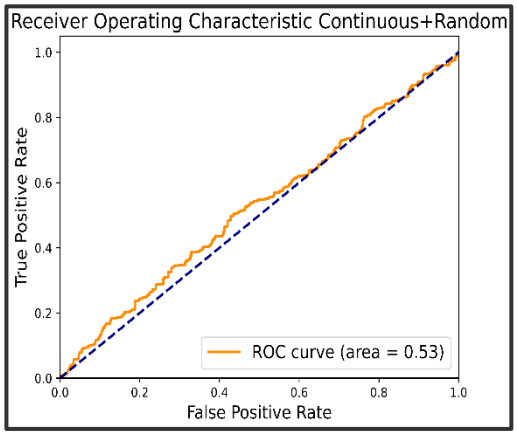

ANN ROC Curve
AUC: 0.53
ROC curve (area = 0.53)
True Positive Rate
False Positive Rate

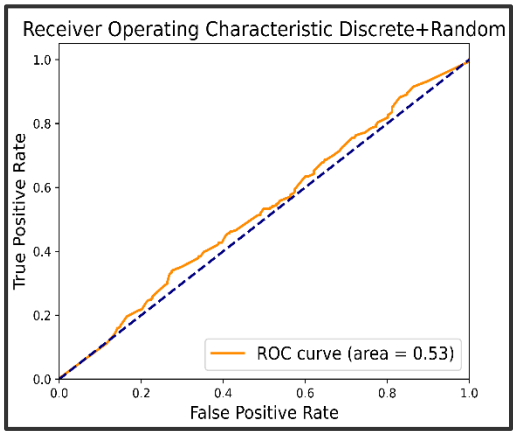

ANN ROC Curve
AUC: 0.50
ROC curve (area = 0.50)
True Positive Rate
False Positive Rate

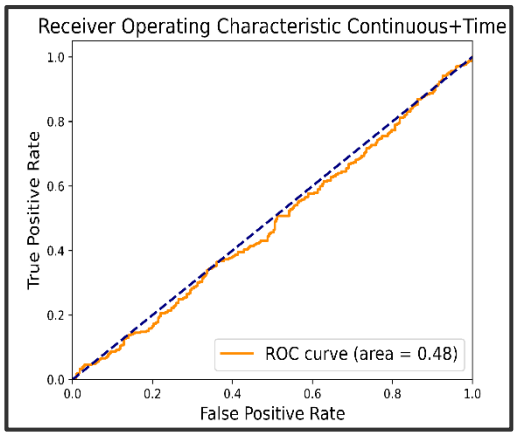

ANN ROC Curve
AUC: 0.53
ROC curve (area = 0.53)
True Positive Rate
False Positive Rate

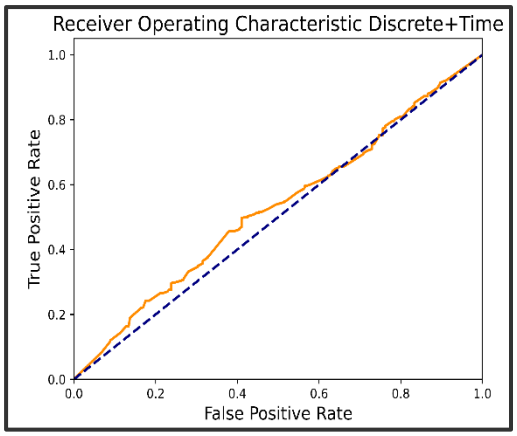

ANN ROC Curve
AUC: 0.51
ROC curve (area = 0.51)
True Positive Rate
False Positive Rate

4. Naïve Bayes









三、AUC 及 Accuracy

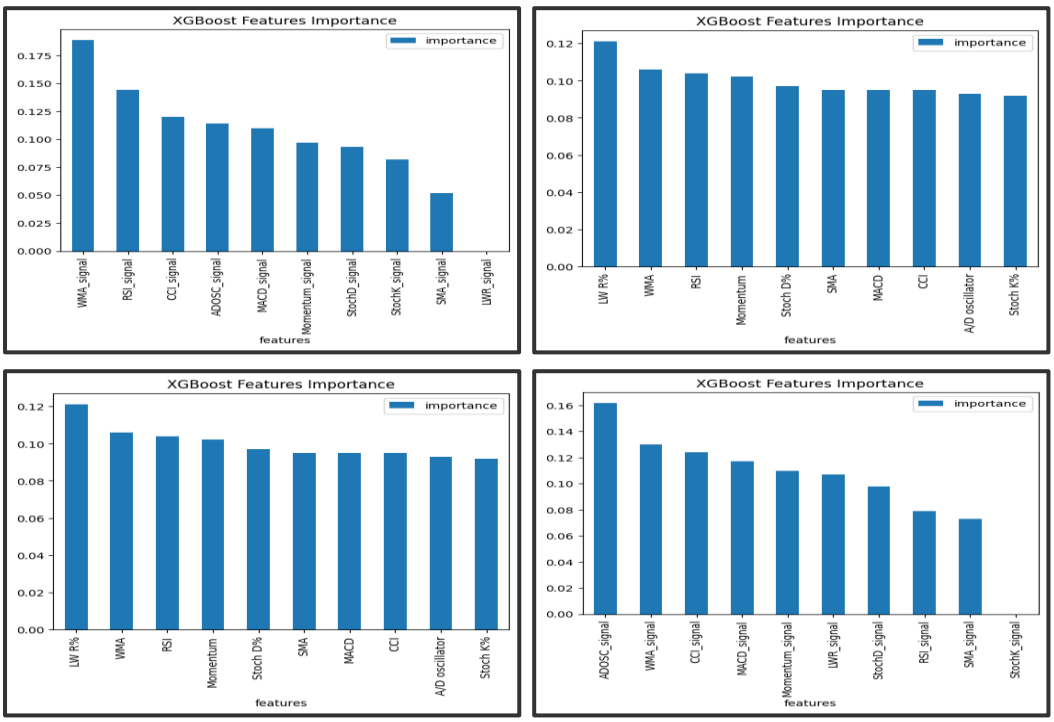
	ANN Random Continuous	ANN Time Period Continuous	ANN Random Discrete	ANN Time Period Discrete	Naïve Bayes Random Continuous	Naïve Bayes Time Period Continuous	Naïve Bayes Random Discrete	Naïve Bayes Time Period Discrete
Accuracy	0.5437	0.5206	0.5192	0.5206	0.5257	0.5385	0.5194	0.5137
AUC	0.5373	0.5178	0.5060	0.5044	0.5261	0.4828	0.5268	0.5274

	XGBoost Random Continuous	XGBoost Time Period Continuous	XGBoost Random Discrete	XGBoost Time Period Discrete	Random Forest Random Continuous	Random Forest Time Period Continuous	Random Forest Random Discrete	Random Forest Time Period Discrete
Accuracy	0.5232	0.5308	0.5031	0.4979	0.5307	0.5281	0.5471	0.5391
AUC	0.5260	0.5197	0.5214	0.4869	0.5033	0.4964	0.5214	0.4981

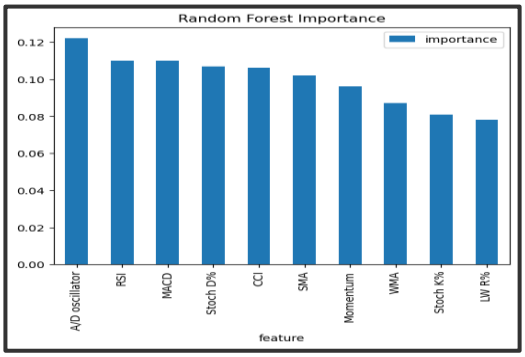
綜合 Accuracy score 和 AUC 的比較，本組認為最佳模型應是 Time period 的離散型變數 ANN 模型，而 Time period 的離散型變數 XGBoost 模型則是表現最差的模型，準確率略低於 50%，存在訓練結果為反向預測的可能。但從混淆矩陣來看，本組發現 ANN 的預測結果皆為全漲或全跌，再加上 Accuracy 和 AUC 依舊相當接近 0.5，本組認為即使是表現最佳的模型，也無法用來精準預測臺灣加權指數的漲跌。接著，本組也試圖透過特徵重要性分析找出技術指標如何影響各模型的預測能力。

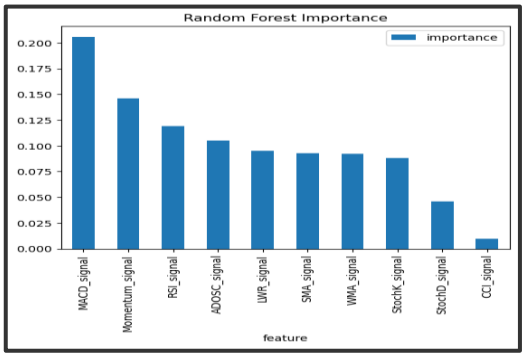
四、特徵根

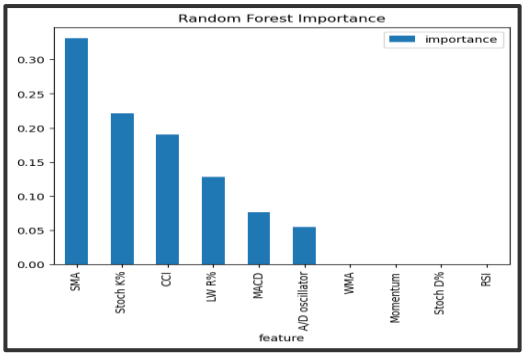
1. XGBoost

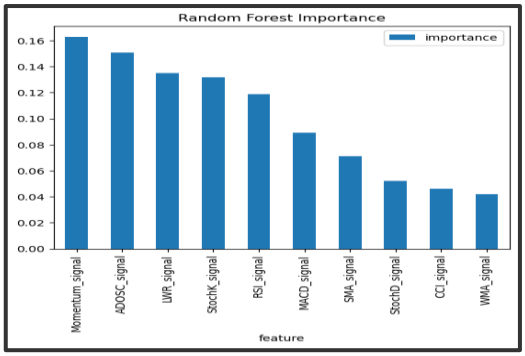


2. Random Forest

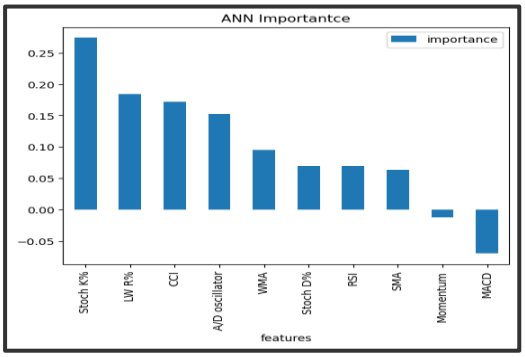


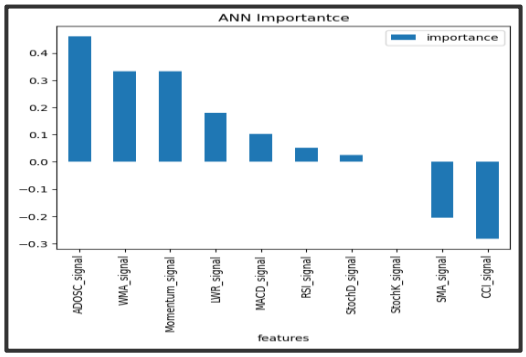


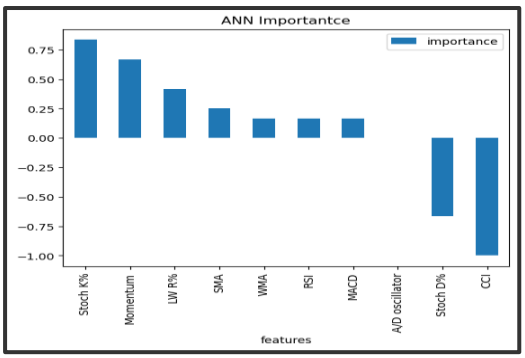


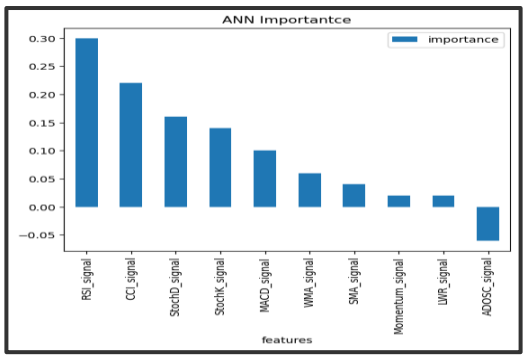


3. ANN

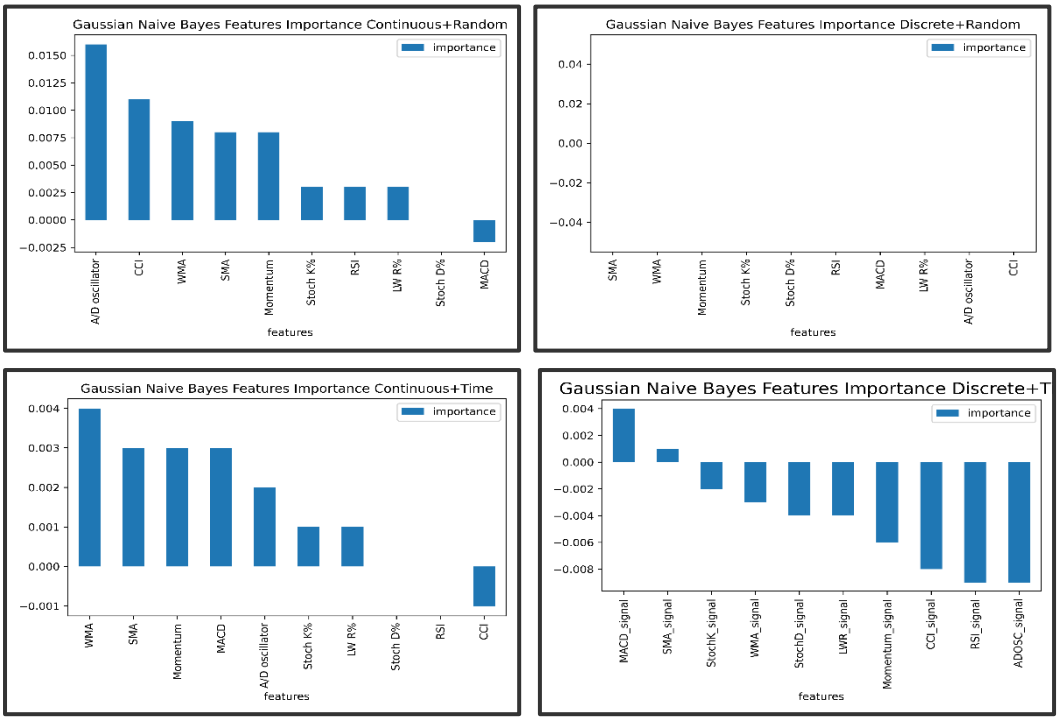








4. Naïve Bayes



從這四個結果可以看出 XGBoost 模型在連續資料的情況下，十種變數的特徵影響程度相差不大，而離散資料中並無哪個特徵在隨機或是時間趨勢取樣中有明顯的影響性。Random Forest 模型也呈現同樣的情況，在四種情形下並無哪個特徵根有明顯重要性，值得注意的是在連續資料且時間趨勢取樣下，有四個特徵根的重要性為零。若將在四個特徵根移除後重新訓練或許可以提高模型的準確率。在 ANN 模型中，連續資料進行訓練的情況下，stochastic K%的重要性相較於其他特徵根都是明顯較高的，離散資料則無。此外，ANN 模型與 Naïve Bayes 模型都出現有負的特徵根，這也代表出現所選特徵根對模型預測方向為負相關，也可以讓我們下次在訓練模型時重新考慮特徵根的選擇。綜上所訴，特徵根不明顯以及出現零或是負的特徵根也反映了我們模型準確率較低的問題。

伍、結論

根據本研究對 2013 年至 2023 年期間，台灣加權指數歷史數據的分析，結果顯示四種機器學習模型（XGBoost、隨機森林、類神經網路、Naïve Bayes）在預測加權指數漲跌方面的表現均不理想。無論是使用連續型數據還是離散型數據，模型的準確度與 ROC 曲線下面積均接近 0.5 ，顯示其預測效果與隨機猜測無異，模型沒有任何預測能力。此外，部分模型甚至出現預測結果為全漲或全跌的現象，顯示出模型可能存在過度擬合或是數據不足的問題。

出現與參考文獻研究大相徑庭之結果可以歸因於多種因素。首先，股市是一個高度複雜且動態的系統，受到眾多因素影響，包括經濟指標、政治事件、投資者情緒等，因素之間的相互作用非常複雜，技術指標沒辦法全面反映市場的真實情況，僅僅依賴技術指標來預測股價漲跌可能過於簡單化。其次，台灣與印度市場的結構與動態上存在差異，加上技術指標具有時效性與週期性，本研究所使用的機器學習模型未能

有效捕捉市場的非線性與隨機性變動。最後，機器學習模型需大量高品質的訓練數據來提升其預測能力，而本研究中的數據樣本可能不足以支持模型充分訓練，從而影響模型的預測精確度。

此外，探索特徵選擇及資料處理方法也是提升模型預測能力的重要方法。數據的預處理方式與變數選擇可能導致重要訊息的損失或是納入過多雜訊。另外，股市也常受到投資者情緒、非理性行為與內線交易嘿嘿的影響，而增加技術指標預測難度。另外，我們從特徵重要性分析可以觀察到，多數技術指標對模型預測結果的影響有限，甚至在某些情況下呈現負相關，顯示出選取適當的特徵及數據處理方法對提升模型預測能力至關重要。未來研究應考慮引入更多元的數據及更複雜的模型架構，例如：深度學習模型，期望能更準確的捕捉股票市場變化。

總結，本研究提供機器學習在台灣股市預測中的初步探索，儘管目前研究結果不如我們預期，但也為後續研究指引方向。我們發現，現有機器學習模型在預測股價漲跌方面的表現不理想，主要是由於股市的複雜與動態性，加上模型選擇上的限制。然而，該結果並不意味著機器學習在金融市場預測中沒有未來發展潛能。

未來研究預期會引入更多元的數據，例如：總經數據、行業指標、新聞事件、社交媒體情緒等，以期能更全面的捕捉市場變化。此外，使用更複雜的模型，例如：深度學習、強化學習模型等等，以獲得處理非線性、高維度數據時更強的預測能力。數據處理方法的改進也至關重要，包括更有效的特徵選擇與數據預處理技術，以提升整體模型的預測效能。透過不斷改進模型及數據處理方法，我們相信機器學習在金融市場預測中的應用仍具有廣闊的前景。

陸、參考文獻

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.

Avci, E. (2007). Forecasting daily and sessional returns of the ISE-100 index with neural network models. *Journal of Dogus University*, 8(2), 128-142.

Chen, A. S., Leung, M. T., & Daouk, H. (2003). Application of neural networks to an emerging financial market: Forecasting and trading the Taiwan Stock Index. *Computers and Operations Research*, 30(6), 901-923.

Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul stock exchange. *Expert Systems with Applications*, 38(5), 5311-5319.

Karaatli, M., Gungor, I., Demir, Y., & Kalayci, S. (2005). Estimating stock market movements with neural network approach. *Journal of Balikesir University*, 2(1), 22-48.

Olson, D., & Mossman, C. (2003). Neural network forecasts of Canadian stock returns using accounting ratios. *International Journal of Forecasting*, 19(3), 453-465