國立清華大學 電機工程學系

實作專題研究成果摘要

TXF LOB mid-price prediction based on tree-based model

基於樹模型預測台指期限價單數據簿的中間價格

專題領域： 系統領域

組　　別： B460

指導教授： 翁詠祿 教授

組員姓名： 吳昕庭、蔡宜蒨

研究期間： 113 年 2 月 4 日至 113 年 11 月 12 日止，共 9 個月

摘要

隨著金融科技和高頻交易的發展，市場參與者愈發關注如何透過限價單數據簿（Limit Order Book, LOB）來預測期貨價格變動。本研究聚焦於台灣加權股價指數期貨（TXF），以 2023 年 9 月到期的合約為研究對象，探討其價格波動與市場趨勢之間的關係。研究的主要目標是利用 LOB 數據預測 TXF 的中間價格和漲跌走勢，並比較不同機器學習模型的預測性能，分析模型的預測結果異同。

在研究方法方面，數據來自元富期貨的高頻交易數據，經過預處理後，保留了第一檔至第五檔的買賣價格和交易量等關鍵資訊。我們選用了三種樹模型，分別是 Random Forest、Decision Tree 和 Gradient Boosting Tree，並使用均方根誤差（RMSE）、平均絕對誤差（MAE）和 R 平方等多項指標來評估模型的表現。此外，為了提升模型預測的準確度，採用 Moving Average 方式降低極端價格對於預測結果的影響，並改以 Sliding Window 方法生成一個隨時間更迭的模型，使模型能夠更依照近期資訊的變化來預測，藉此提升預測準確性。

研究結果顯示，Random Forest 在所有指標上表現最佳，其 RMSE 和 MAE 顯著低於其他模型。預測漲跌走勢時，Random Forest 達到 68.75%的準確率，表現優於其他模型。此結果證明 Random Forest 的集成學習特性和穩定性使其更適合處理高頻數據和捕捉市場的複雜模式。

本研究提供了一種基於機器學習的期貨價格預測方法，並為未來在金融市場中應用 AI 技術提供了參考。未來的研究方向包括演算法優化、硬體實現、跨學科合作及風險管理，期望能夠構建一個實時預測系統，為投資者提供更即時的市場走勢資訊，並協助其做出明智的決策。

一、研究背景與動機

金融科技的蓬勃發展，尤其是在高頻交易領域，使市場參與者愈加關注如何利用限價單數據簿（LOB）預測市場趨勢。本研究選擇台灣加權股價指數期貨（TXF）作為研究標的，探討利用高頻數據來預測價格變動，進一步提升投資決策的科學性和準確性。

二、研究流程圖

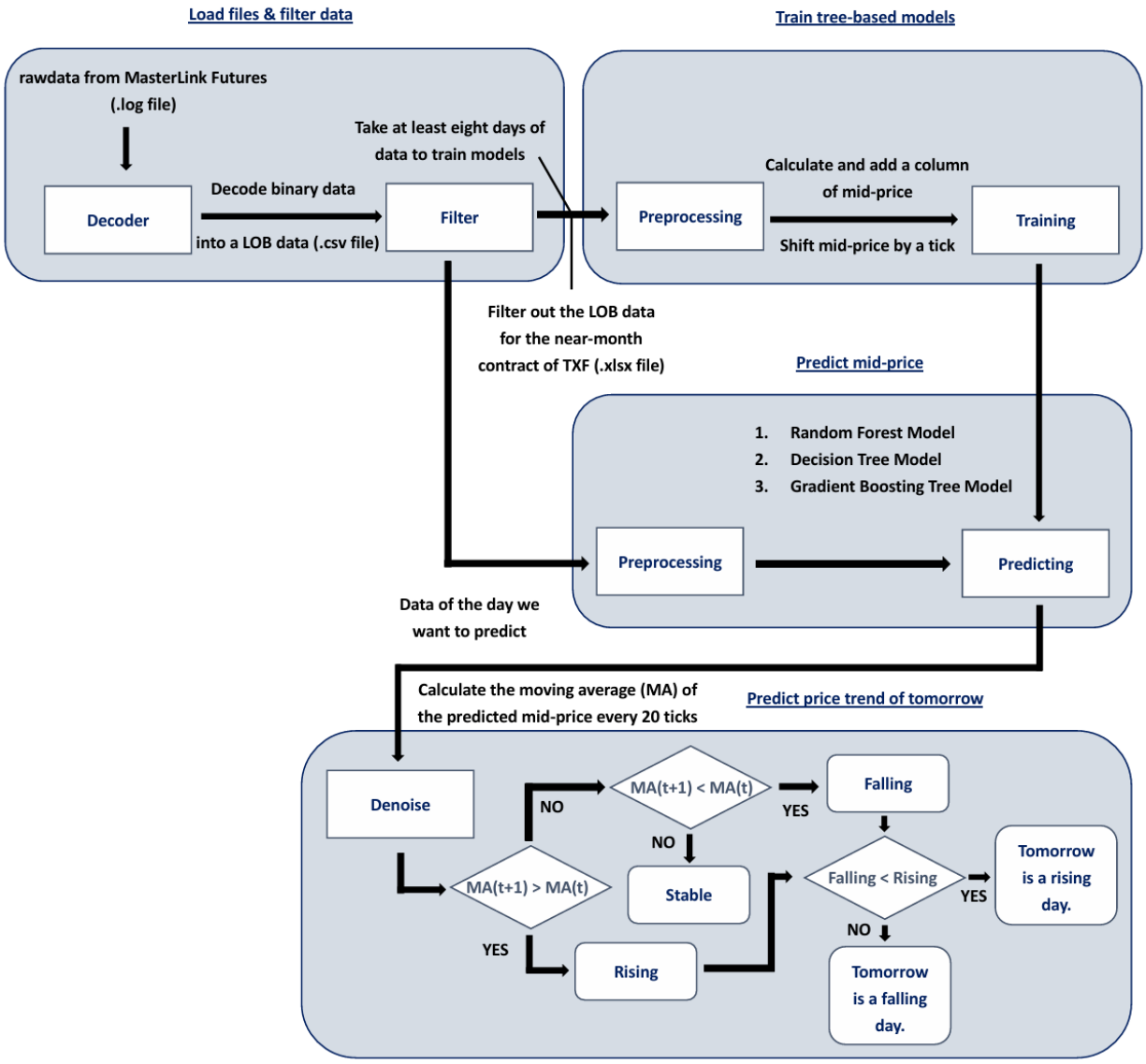

Fig.1 研究流程圖

三、研究目的

- 利用 LOB 數據預測台指期近月合約商品的中間價格及其漲跌走勢。
- 分析模型預測結果的差異，探索不同模型在高頻數據上的適用性。
- 比較傳統訓練方法與改良版 Sliding Window 方法的模型性能與表現。

四、研究方法

（一）數據來源與預處理

採用來自元富期貨的限價單資料，涵蓋早盤與夜盤的高頻交易數據，並以 2023 年 9 月到期的台指期合約作為研究對象。首先，將原始數據中包含的各個檔次的買賣價格及交易量進行解碼，並轉換成易於分析的 CSV 格式資料，以便後續的數據處理。接著，過濾其他期貨商品，篩選出台指期 2023 年 9 月到期的商品數據，並移除不完整的數據行，確保訓練模型的資料能完整且有意義。為了更好地訓練模型，我使用至少八日的高頻數據來進行模型訓練，總數量超過百萬筆交易紀錄。這些數據提供了豐富的信息，如第一檔至第五檔的買賣價格、交易量等，使得模型能夠從中提取到有用的特徵來預測中間價格的變化。

（二）模型選擇

選擇三種樹模型，分別是 Decision Tree、Random Forest 和 Gradient Boosting Tree，以預測台指期商品的中間價格，並對這些模型的表現進行對比。

決策樹（Decision tree）是一種用於決策和預測的模型，其結構由節點、分支和葉節點組成。節點表示對某些屬性進行的測試或決策，分支代表測試的結果，葉節點則對應最終的預測結果。在建立決策樹時，核心步驟是選擇最佳屬性來分割數據，而這通常依賴於以下兩種常見的指標和公式：

1. Gini 不純度 $= 1 - \sum_{i=1}^{k} {p_i}^2$

   其中，p_i表示屬於第 i 類的樣本比例，k 是類別數量。Gini 不純度越低，表示分割後的數據越純，故決策樹會採取讓 Gini 不純度越低的決策。

2. 熵 $= 1 - \sum_{i=1}^{k} p_i \log_2 p_i$

   同樣，p_i是第 i 類的樣本比例，熵值越低，分割後的數據越純，故決策樹會採取讓熵值越低的決策。

決策樹的優勢在於其易於解釋和可視化，使用這些公式選擇分割屬性，使其能有效地處理分類和回歸問題，並直觀地呈現決策過程。

隨機森林（Random Forest）是一種集成學習方法，用於解決分類和回歸問題。它通過結合多棵決策樹來進行預測。在訓練過程中，隨機森林會構建多棵決策樹，對於分類任務，輸出為這些樹的預測類別中出現次數最多的值；對於回歸任務，則取這些樹預測值的平均值作為最終結果。

其每棵樹是基於訓練數據的一個隨機子集構建的，且每次分割時僅考慮隨機選取的一部分特徵。這種隨機性有助於減少過擬合問題，並提升模型的泛化能力。通過在數據選擇和特徵選擇中的隨機性，隨機森林能夠在保證穩健性的同時提高預測準確性。

梯度提升樹（Gradient Boosting Tree, GBT）也是一種基於決策樹的集成學習方法，透過逐步減少預測誤差來提升模型性能。其核心思想是以損失函數的負梯度 $-\nabla L(y,\hat{y})$作為殘差目標，在每次迭代中訓練一棵新的弱學習器，通常稱為淺層決策樹，並將其按學習率(η)加權累加至模型，即 $\hat{y}_t = \hat{y}_{t-1} + \eta h_t(x)$，其中$\hat{y}_t$表示第 t 次迭代後的模型預測值，$h_t(x)$為當前弱學習器的預測結果。這種逐步優化的方式使梯度提升樹能有效處理非線性關係，適應分類和回歸等多種任務，在金融市場預測和搜索引擎排序等領域表現出色。然而，梯度提升樹對於學習率(η)、樹的數量和深度等超參數較為敏感，且因訓練過程依賴於逐次迭代，其計算成本較高。此外，對於異常極端值的穩定性較弱，需依賴適當的數據預處理和超參數調整來確保模型性能。

（三）模型訓練與評估

基於前一筆 Tick 的資料，以上述模型預測下一筆 Tick 的中間價格，避免 Data Leakage 以提高模型穩定性。將預測結果與正確的中間價格比較，使用 RMSE、MAE 和 R 平方來評估預測精度，最終發現 Random Forest 模型表現最佳，Decision Tree 模型次之，Gradient Boosting Tree 模型表現較差。

（四）模型改良

為了提升預測準確性，目標是讓模型不易受到極端價格所影響，也期望預測結果更貼近近期價格漲跌的變化趨勢，而採用 Moving Average 和 Sliding Window 方法。考量 Random Forest 模型的最佳表現，以其作為基礎延伸，利用前 8 天的數據進行模型訓練，預測隔日的價格，並以每 20 筆的平均價格的變化來預測後天的漲跌趨勢，並在預測新一天數據時，將最舊的一天數據剔除，加入最新一天的數據，依然保持以 8 天的數據訓練模型，形成隨時間不斷更新的模型，也讓模型能更靈敏地適應近期資訊的變化。

五、研究結果

本研究結果顯示，Random Forest 模型在預測市場波動和趨勢變化方面的表現最佳，尤其是在高頻數據下具備極高的穩定性。如表一所示，Random Forest 的均方根誤差（RMSE）為 30.79，平均絕對誤差（MAE）為 13.85，並且 R 平方值高達 0.9999，顯示出模型在捕捉價格變動特徵上的精準度。相比之下，Decision Tree 模型因過擬合問題導致預測效果不如理想，而 Gradient Boosting Tree 模型則可能受到極端值的影響，未達到預期表現。此外，使用 Random Forest 模型預測隔日中間價格的移動平均，進而判定漲跌走勢，其準確率達到 62.5%，顯示出該模型在趨勢預測上的有效性。

樹模型	隨機森林 (Random Forest)	決策樹 (Decision Tree)	梯度提升樹 (Gradient Boosting Tree)
均方根誤差	30.79339	38.098049	64.3533157
R^2	0.9999964	0.99999461	0.999968302
性能表現	最佳	次之	最差

Table.1 不同樹模型之間的性能表現差異

採用 Sliding Window 方法後，成功避免模型因過時數據而導致的準確性下降，同時提升了對近期趨勢的敏感度。如表二，可發現相較於傳統方法訓練的模型，以 Sliding Window 方法訓練出來的 Random Forest 模型具有較低的均方根誤差，表示此模型較貼近真實的價格走勢。此外，如表三、表四及圖二所示，在預測準確率上顯著提高，由原本的 62.5%提升至 68.75%，顯示此方法在改善模型性能和捕捉市場動態方面的有效性。這種改良方式不僅優化了模型的預測能力，還展現了在高頻交易環境中應用的潛力，為市場趨勢分析提供了更可靠的工具。

採用隨機森林模型所預測之日期	8/30	8/31	9/1	9/4	9/5	9/6
均方根誤差	12.002	4.375	7.635	4.441	9.756	4.6601
R^2	0.9903	0.9985	0.9965	0.9984	0.9906	0.9978
比較	相較於傳統訓練的模型，以 Sliding Window 方法訓練出來的模型性能較好，具有較低的誤差。					

Table.2 Sliding window method 不同日期對應不同模型的表現

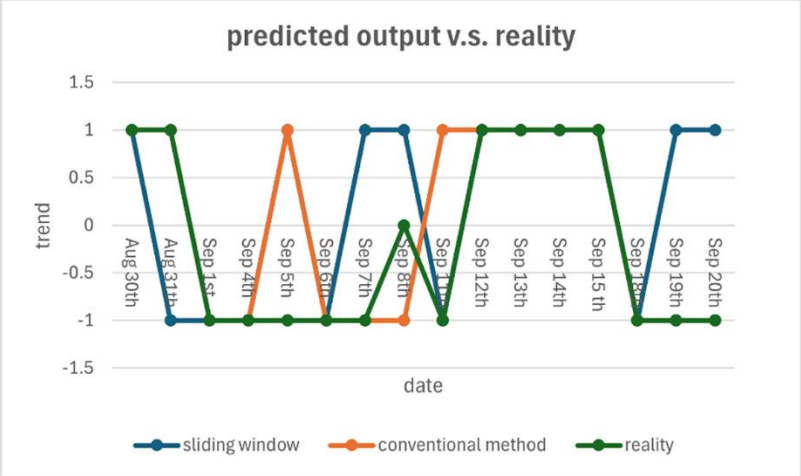

Fig.2 漲跌走勢預測結果

傳統模型			
Precision	Recall	F1-score	Accuracy
62.5%	71.4%	66.6%	62.5%

Table.3 傳統模型預測 9 月台指期商品漲跌結果

Sliding Window 模型			
Precision	Recall	F1-score	Accuracy
62.5%	71.4%	66.6%	68.75%

Table.4 Sliding Window 模型預測 9 月台指期商品漲跌結果

六、結論與討論

（一）模型性能

1. Random Forest 在預測中間價格和市場趨勢方面表現最佳
2. Decision Tree 簡單直觀，但處理高維數據有局限性
3. Gradient Boosting Tree 受到極端數據影響，需要優化與超參數調整

（二）訓練方法

1. 傳統方法的不足

   傳統方法基於前一筆 Tick 的資料預測下一筆 Tick 的中間價格，能在一定程度上反映價格變化。但由於固定的數據範圍無法及時適應市場的波動，可能導致對近期價格走勢的敏感度下降，尤其在市場快速變動的情況下容易降低預測準確性。

2. Sliding Window 方法的改進

   Sliding Window 方法有效解決了傳統方法的局限性，通過動態更新數據範圍來確保模型始終基於最新的市場信息進行訓練。該方法在模型訓練時剔除過時數據，並加入最新數據，使模型能更靈敏地適應近期趨勢，成功提升了預測準確率，顯示其在高頻交易環境中的應用潛力。

（三）研究意義

1. 為投資者提供有效的 TXF 價格走勢預測方法
2. 證明機器學習模型在捕捉複雜市場模式的優勢
3. 為金融科技研究提供重要參考

（四）研究局限

1. 僅聚焦單一期貨合約和有限時間範圍
2. 金融市場還存在許多影響因素，需要盡可能地考慮到

（五）文獻驗證

在訓練數據集相同的條件下，即使用 8/17、8/18、8/21、8/22、8/23、8/24、8/25 和 8/28 共八日的限價單數據進行模型訓練，結果顯示隨機森林模型在決定係數(R^2)和均方根誤差(RMSE)兩個指標上的表現均優於決策樹和梯

度提升樹。具體而言，模型性能排序為 Random Forest 最佳，Decision Tree 次之，Gradient Boosting Tree 表現相對較差。該結果與我們參考的文獻《Machine Learning Techniques for Price Change Forecast Using Limit Order Book Data》[5]中所提到的模型性能優劣排序結論一致，從而驗證了本研究所選用模型的有效性和可靠性。

七、心得感想

在本次研究中，我們深刻體會到金融科技與機器學習結合的重要性與挑戰。從數據的預處理到模型的選擇，每一步都需要經過反覆試驗與調整。在預測期貨價格的過程中，我們發現高頻數據具有極強的波動性，這使得模型很容易過擬合。為了提高準確度，我們在數據清洗和特徵選擇上投入了大量精力，並嘗試了多種模型。最終，Random Forest 模型的表現最佳，這也驗證了集成學習在處理複雜數據時的優勢。

此外，我們也發現研究的局限性。金融市場的變動不僅受限於限價單數據，還有其他外在因素的影響。由於研究時間和資源的限制，我們只能聚焦於單一合約和有限的數據，這限制了模型的普適性。未來，如果能結合更多的外部市場數據，如新聞和社交媒體情緒，或許能提升模型的準確度。

最後，這次研究增強了我對金融科技的興趣，也讓我們更加了解機器學習在實際應用中的挑戰。我們希望未來能夠繼續深入探討該領域，並且有機會將研究成果應用於實際交易中，以為市場參與者提供更科學的決策支持。

八、參考文獻

[1] Manojlović, T., & Štajduhar, I., “Predicting stock market trends using random forests: A sample of the Zagreb stock exchange,” IEEE 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),May 2015,pp. 25–29.
[2] Meher, B. K., Singh, M., Birau, R., & Anand, A., “Forecasting stock prices of fintech companies of India using random forest with high-frequency data,” Journal of Open Innovation: Technology, Market, and Complexity,March 2024.
[3] Viviana Arrigoni , Giuseppe Masi , Emanuele Mercanti , Novella Bartolini , Svitlana Vyetrenko, “Stock Shocks Modelling and Forecasting,” IEEE 43rd International Conference on Distributed Computing Systems Workshops, 2023.
[4] Deepan Palguna and Ilya Pollak, “Non-Parametric Prediction of the Mid-Price Dynamics in a Limit Order Book,” School of Electrical and Computer Engineering Purdue University, West Lafayette, IN 47907, USA
[5] James Han∗, Johnny Hong†, Nicholas Sutardja‡, Sio Fong Wong§, “Machine learning techniques for price change forecast using the limit order book data,” 2015.