國立成功大學
National Cheng Kung University

基於 K-NN 分群與隨機森林模型對股票價格預測

Stock Price Prediction Using K-NN Clustering and Random Forest Model

黃群翔 R26121081

研究動機

- 股票市場的不確定性:
  - •股票價格的波動性由於多種因素而變得無法預測:
    - 宏觀經濟因素：如利率變動、經濟增長數據、貨幣政策等。
    - 公司基本面：例如企業財報、管理層變動、產業趨勢等。
    - 市場情緒與投資者行為：投資者的情緒與市場心理影響股價，這些因素很難用數據量化。
    - 突發事件：例如自然災害、政治事件、戰爭等因素常常使市場短期劇烈波動。
  - •這些不確定性使得傳統的統計模型（如 ARIMA）或基於簡單趨勢的模型（如 SMA）難以準確捕捉到市場的真實波動。

研究方法

- 過往方法：

  Arroyo, González-Rivera, and Maté（n.d.）提出了一種基於 k-NN 的方法，用於預測股票價格的上下區間。

  - 核心概念：

    該方法通過 k-NN 在歷史數據中找到 k 個最相似的樣本，並利用這些樣本來預測未來股價的範圍（即上下區間）。

    k-NN 會計算當前數據點與過去歷史數據點之間的距離，然後選擇 k 個最接近的數據點，根據這些相似的數據來預測未來價格的走勢。

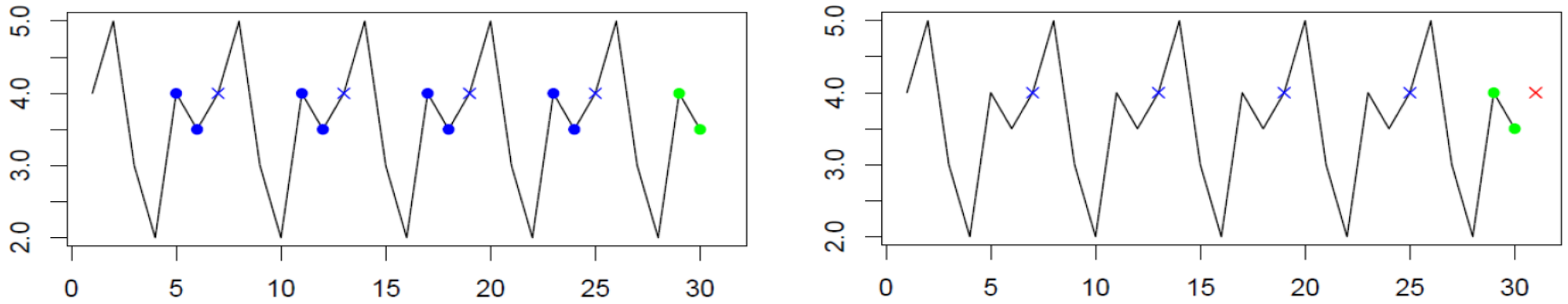

圖1：k-NN 方法概述

研究方法

- 改良的 k-NN 方法： ：
  - ATR 作為特徵：
    - ATR 是一種衡量市場波動性的指標，能夠有效捕捉價格變動的範圍和幅度。
    - 該方法將 ATR 值作為主要特徵，幫助模型在市場波動性高或低的情況下進行調整。
  - k-NN ：
    - 使用 k-NN 在歷史數據中找到 k 個與當前市場情況最相似的樣本。
    - 距離計算基於 ATR 和其他選定的市場特徵，以確保選出的樣本具有代表性。
  - 動態調整 k 值：
    - ATR 值的大小直接影響 k-NN 中 k 的選擇：
      - 當 ATR 值較高時（波動性大），增大 k 值以增加樣本數量，提高模型的穩定性。
      - 當 ATR 值較低時（波動性小），減小 k 值以提高模型對局部市場走勢的靈敏度。
  - 隨機森林進行預測：
    - 將選出的 k 個相似樣本的 ATR 值和其他特徵作為輸入，使用 隨機森林模型（Random Forest） 進行未來 ATR 數值的預測。

流程圖

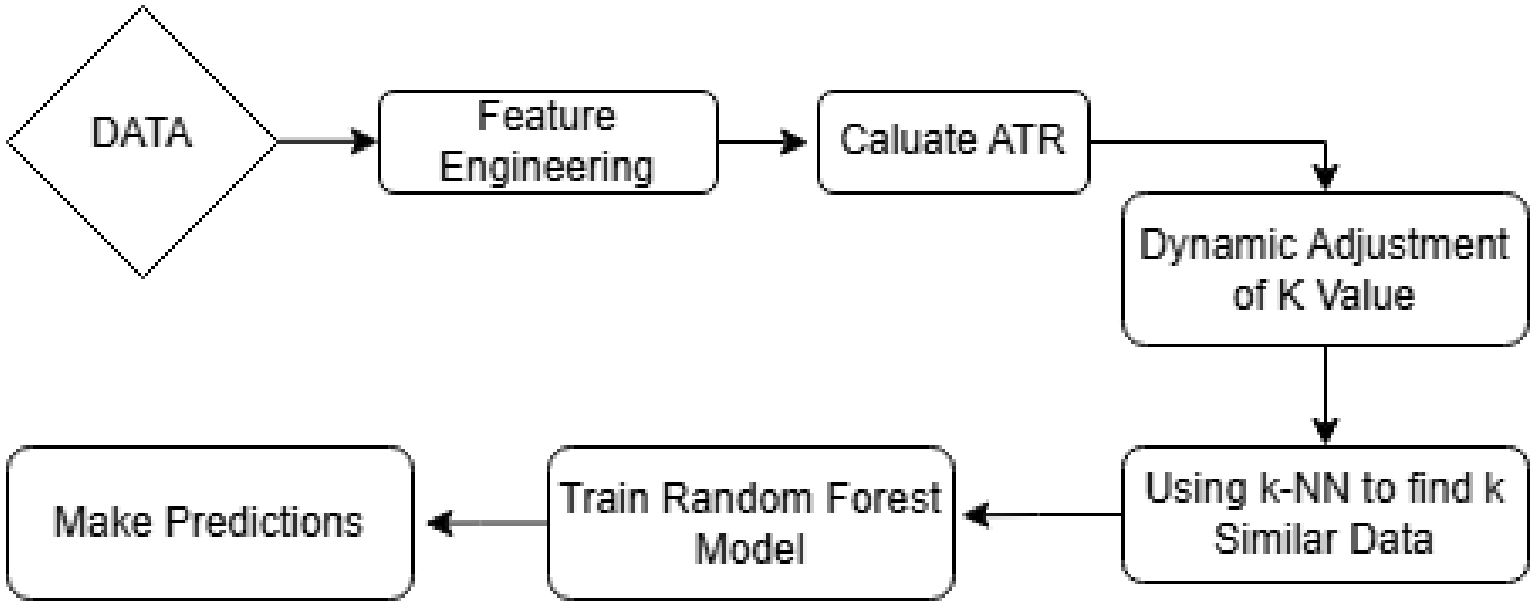

圖2: 股票價格預測流程圖

實例分析

- 資料取自 Kaggle 上的一筆 BMW 的股價。
- 資料範圍 ： 1996 － 2024
- 此筆資料包含了
  - 每日交易的最高價
  - 每日交易的最低價
  - 收盤價
  - 調整後收盤價 ： 是經過股票拆股（split）或股利（dividends）等因素調整後的收盤價。它反映了這些 corporate actions（公司行為）對股票價格的影響，使得股價能夠與過去的數據進行比較，而不會因為這些行為而產生誤差。

Date	Adj_Close	Close	High	Low	Open	Volume
1996-11-08	8.100290	18.171000	18.209999	18.171000	18.209999	767000
1996-11-11	8.078445	18.122000	18.200001	18.082001	18.190001	260000
1996-11-12	8.139520	18.259001	18.327999	18.091999	18.160999	1066000
1996-11-13	8.126592	18.230000	18.344000	18.190001	18.344000	793000
1996-11-14	8.152893	18.289000	18.289000	18.132000	18.205000	351000

表1: BMW 股票資料

實例分析

- k-NN

透過繪製收盤價看到收盤價的時序圖並不平穩，為次我們對資料進行一階差分。差分過後可以看到資料相對穩定。我們再依據差分過後的收盤價用k-NN 的方式進行預測。

圖3: BMW 股票收盤價時序圖

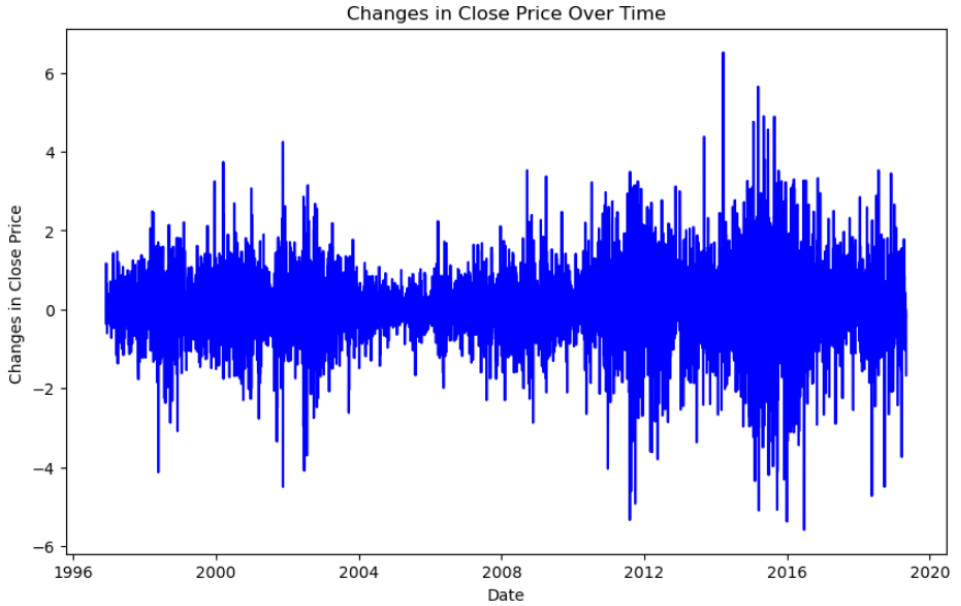

圖4: BMW 股票收盤價差分後時序圖

實例分析

- k-NN + Random Forest Model

透過繪製 ATR 的時序圖可以看出此時序圖並不平穩，為次我們對 ATR 進行一階差分。差分過後可以看到資料相對穩定。我們再依據差分過後的收盤價用k-NN 的方式找相似的資料點，在加入特徵放入隨機森林模型進行預測。

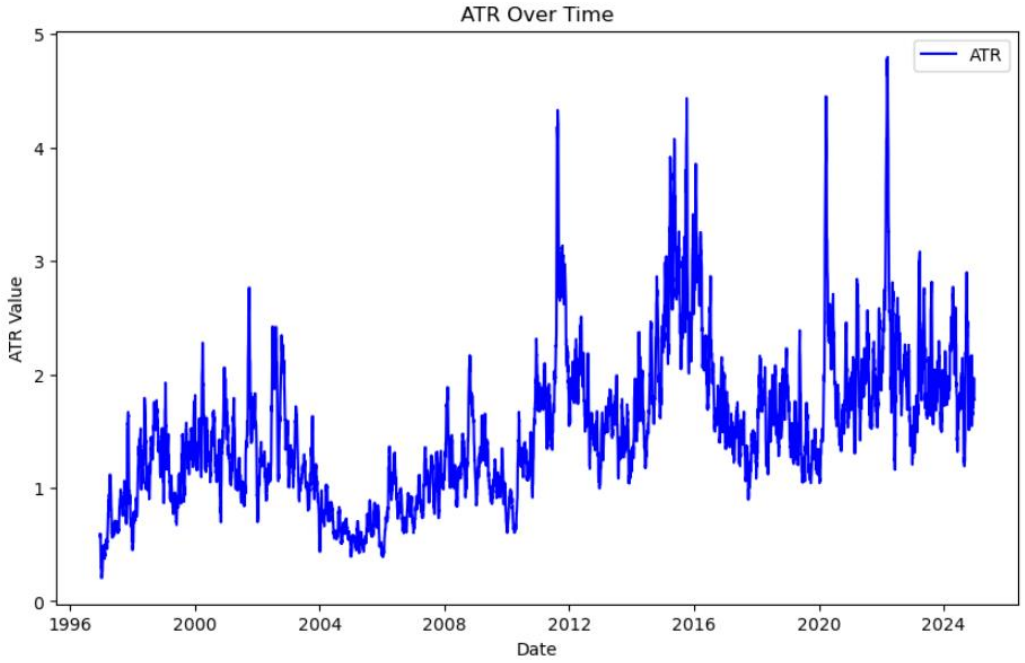

圖5: BMW 股票 ATR 時序圖

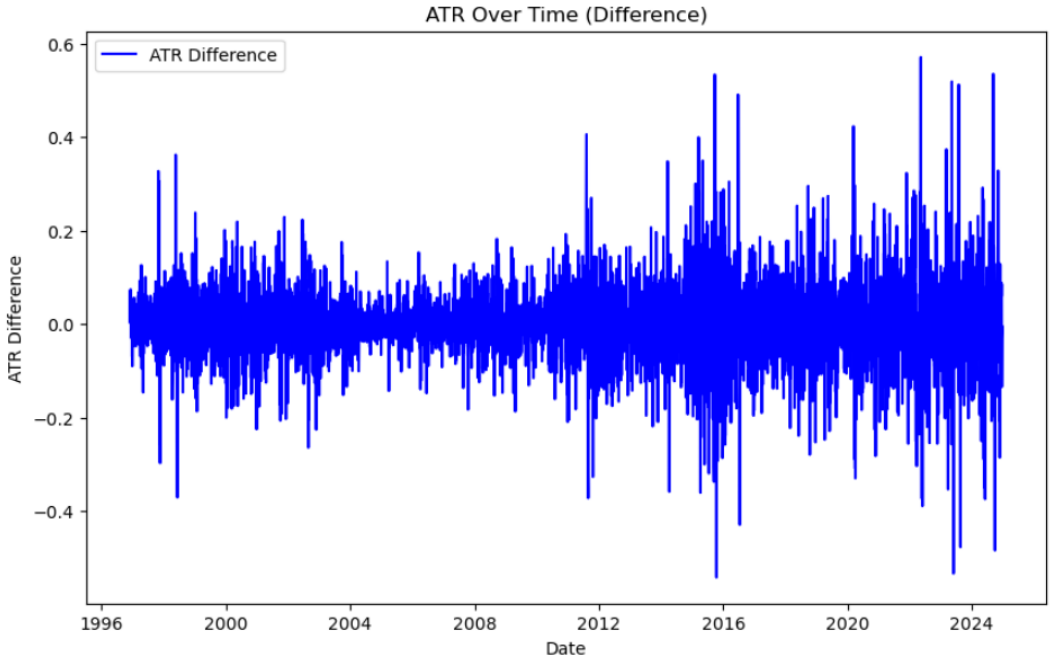

圖6: BMW 股票 ATR 差分後時序圖

實例分析

- 預測結果

圖7 和圖8 分別展示了預測結果。可以發現，原先使用 k-NN 的方法預測結果較為準確，對應的 RMSE 分別為 1.66 和 4.75。我們推測新方法表現較差的原因可能是參數 k 值的設定方法仍需改進，這將是未來研究的重要方向。此外，目前在特徵選擇上僅使用了前一天的收盤價和 ATR 值，這可能在引入額外特徵時增加了數據中的噪音，進而影響模型的預測精度。未來可以嘗試引入更多有意義的特徵，同時進一步調整隨機森林模型的參數，以優化現有方法並提升預測準確性。

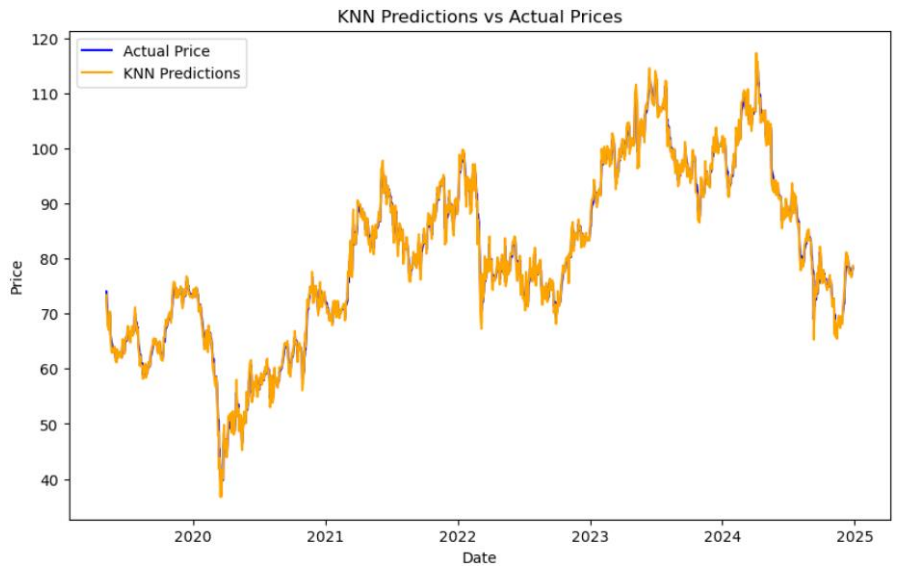

圖7: k-NN 預測結果

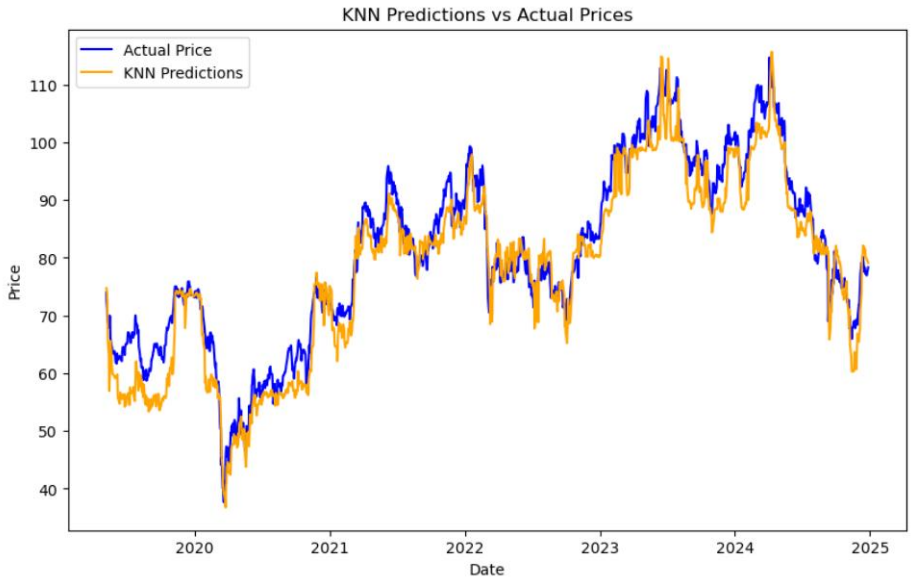

圖8: k-NN + Random Forest Model 預測結果

結論

本研究結合 k-NN 的加權機制與隨機森林模型，提出了一種動態局部預測框架，為股票價格預測問題提供了一種新穎的解決方案。本研究的主要貢獻如下：

1. 動態鄰域選擇

透過自適應 k-NN ，根據市場波動性（ATR 值）的變化動態調整鄰域範圍，靈活篩選局部訓練集，提高了對局部數據的利用效率。

2. 非線性建模能力

隨機森林模型在處理局部數據的非線性關係時展現出優越的性能，能捕捉市場數據中複雜的短期趨勢，從而增強預測的精確度。

雖然結果顯示該方法的 RMSE 較原始 k-NN 模型有所上升，但這可能是由於動態調整 機制所引起。在未來的研究中，可以考慮以下幾個方向進一步改進：

1. 引入更多市場指標與外部數據，以豐富模型的特徵選擇，提升預測能力。
2. 探索其他加權策略或核函數，進一步提高 k-NN 的適應性和準確性。
3. 將該框架應用於其他金融市場或非金融領域，以驗證其通用性與實際應用價值。

參考文獻

- Arroyo, Javier, Gloria González-Rivera, and Carlos Maté. n.d. *Forecasting with Interval and Histogram Data*. *Some Financial Applications*.
- Wilder, J. W. (1978). *New concepts in technical trading systems*. Trend Research.