**Multivariate Time Series Prediction for Stock Market Data**

資科一 資科一 資科二

曹昱維 鄭詠儒 謝政彥

110753201 110753126 109753207

1. **Introduction**

　　時間序列資料是由按照時間發生先後順序進行排列的數據點序列，以股票市場資料為例，從2000年至2020年為止，股票A的每日收盤股價就是一種時間序列資料。我們希望透過單一變數的時間序列資料(如股票A的每日收盤價)以及多變數的時間序列資料(如所有能源類股的每日收盤價)來訓練各種模型，來比較單變數與多變數對於預測效果的影響，也同時比較相同類型的變數下各種模型的預測效果。

股票市場以波動性、動態性和非線性著稱。由於政治、全球經濟狀況、突發事件、公司財務業績等多重（宏觀和微觀）因素，準確預測股價極具挑戰性。但是，所有這一切也意味著有大量數據可供尋找模式。因此，金融分析師、研究人員和數據科學家不斷探索分析技術來檢測股市趨勢。股票分析基本上可利用基本面分析與技術分析。

ARIMA (p, d, q)模型被廣泛使用在時間序列分析，在了解ARIMA模型之前可以了解相關的模型，(1)自我迴歸模型(AR)，它用前期的資料來預測本期的資料，越接近本期的資料，對預測結果的影響力就越大，設定一筆資料會與他過去p期的資料相關，(2)移動平均模型(MA)，方法在於本期的隨機誤差會與過去產生的隨機誤差有關，接著設定要計算q期移動平均，這些模型的缺點是只能處理穩定的資料，時間序列資料是否為定態資料，會影響後續預測的結果，因此，ARIMA模型改進了以上缺點，使用ADF-test計算差分次數d將時間序列資料處理成定態資料。

LSTM模型是近期時間序列資料常見的深度學習模型，屬於RNN (Recurrent Neural Network) 的一種，適合在輸入特徵空間中提取模式，其中輸入數據跨越長序列，可以從多個輸入變量的問題進行建立many to many 或 many to one 模型，在建模問題方面提供了很大的靈活性，包括可以很好地控制時間序列的幾個參數。

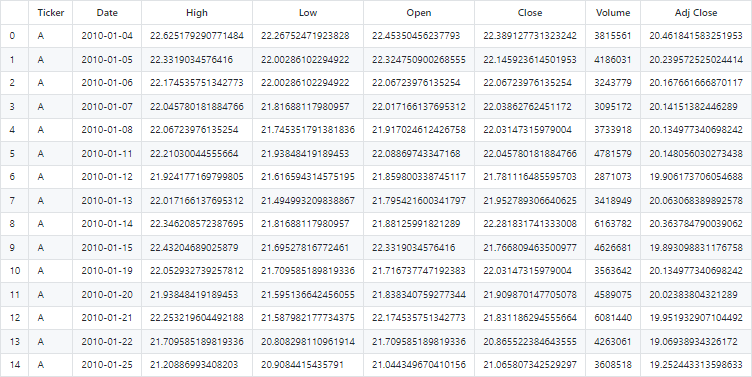
本專案嘗試針對S&P 500中的能源類股APA Corporation股價，利用python相關套件(arima, keras, sklearn,etc.)實作深度學習RNN模型LSTM以及傳統的時間序列模型ARIMA進行預測與比較。

1. **Related Work**

在早期有發展預測模型用來預測與分析股價，像是ANN模型[1]，近年來更是有相對穩定的ARIMA模型[2, 3]用來預測金融相關序列資料，而深度學習模型更是盛行，深度學習模型的效能仰賴參數的設定，以LSTM模型為例重要的參數有activation function (sigmoid, tanh, softmax等等)、optimizer (Adam, Adadelta, RMSprop 等等)、batch size、epoch數量以及hidden layers數量等等[4]。

ARIMA模型處理數據的非平穩性收集和建模，以類似的方式，作為基於深度學習的算法的代表 – LSTM模型是儲存和訓練於較長時間內給定的數據特徵，因此[5]對ARIMA和LSTM做了比較，提供了我們比較依據與方法。

1. **Method**
   1. **Dataset**

****數據集為S&P 500其中的APA股價數據，為一檔美國能源類股，APA Corporation是一家是從事油氣勘探的美國公司，我們從Kaggle[6]上抓取S&P 500股價，並將其APA的股價抓出來，時間序列從2010年一月至2021年十一月，原始資料集如圖一。

圖一、S&P 500股價資料集

* 1. **ARIMA (Autoregressive Integrated Moving Average)**

1. ARIMA（p，d，q）

ARIMA模型是ARMA（p，q）模型的擴展。ARIMA（p，d，q）模型可以表示為：

* p是AR(Autoregressive)的落後期數，為過去資料的加權平均，今天的股價會是過去股價的加權平均值
* d是非季節性差異的數量，使其成為平穩序列所做的差分次數，公式如下：



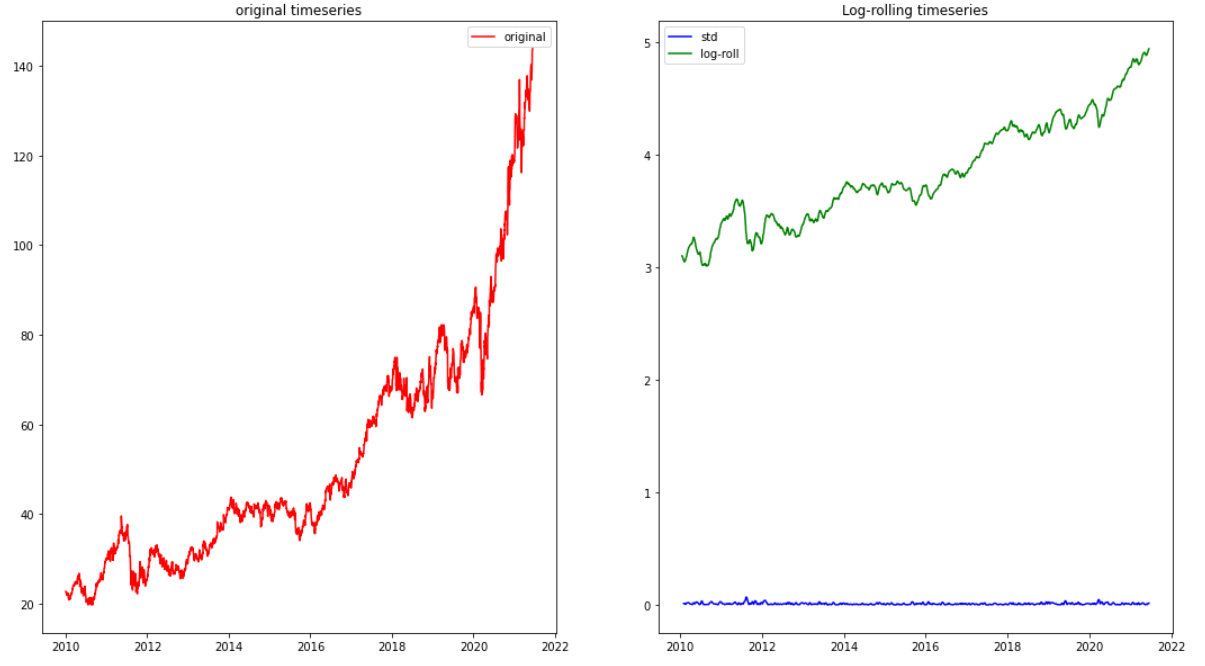
其中L為Lag operator，

*資料來源：wikipedia*

* q是MA(Autoregressive)的落後期數，隨機誤差的加權平均，今天的股價的隨機誤差會與過去產生的隨機誤差有關

1. 原始資料分析

我們使用使用EDA對資料作分析

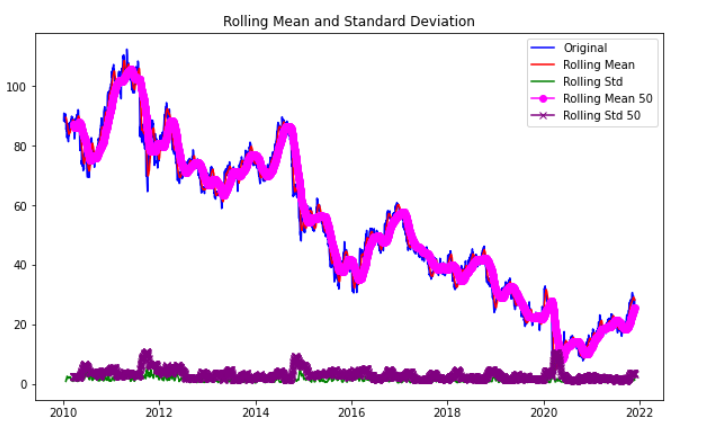
* 將原始資料轉為Log-rolling格式，如圖二。

圖二、原始資料與平滑後資料視覺化呈現

* 將原始資料收盤價轉為rolling格式，如圖三、圖四。



圖三、原始資料收盤價趨勢圖



圖四、原始資料收盤價平滑平均與標準化視覺化呈現

1. ARIMA參數設定：

* 模型參數：

|  |
| --- |
| model\_autoARIMA = auto\_arima(train\_data, start\_p=0, start\_q=0,  test='adf', # use adftest to find optimal 'd'  max\_p=10, max\_q=10, # maximum p and q  m=1, # frequency of series  d=None, # let model determine 'd'  seasonal=False, # No Seasonality  start\_P=0,  D=0,  trace=True,  error\_action='ignore',  suppress\_warnings=True,  stepwise=True) |

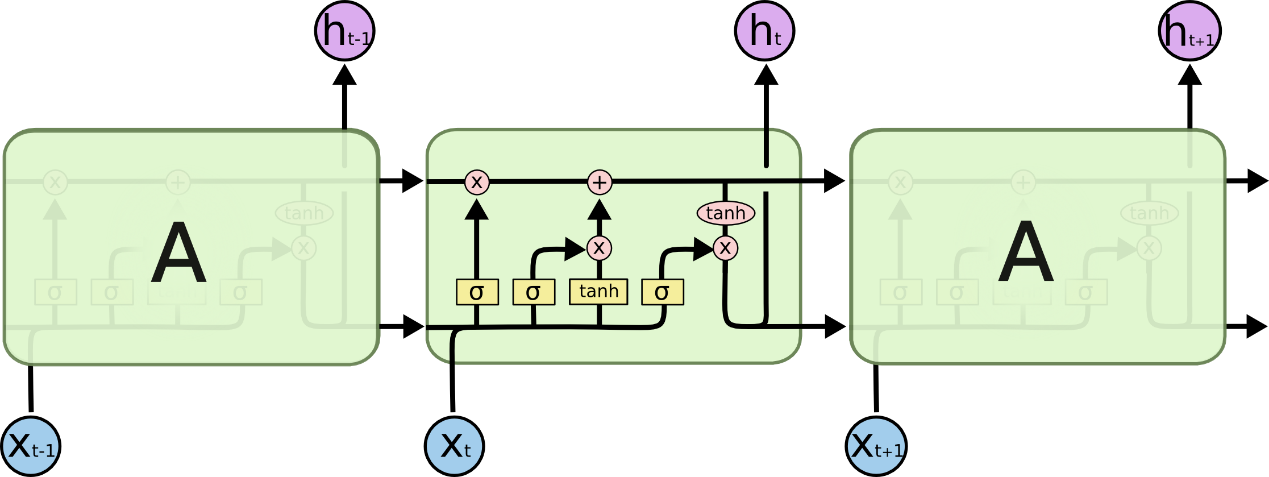
* 建立模型：ARIMA(3,0,2)

|  |
| --- |
| model = ARIMA(train\_data, order=(3, 0, 2)) |

* 1. **LSTM (Long Short-Term Memory)**

1. 模型架構

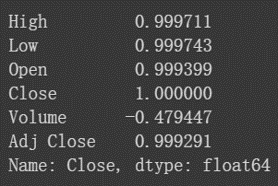
LSTM是RNN架構的其中一種模型，不同於RNN的一層架構，主要解決時間序列的問題，是由四種架構構成如圖五：

* Input Gate: feature輸入時，input gate會控制是否輸入
* Memory Cell: 將計算出的值儲存，使下個階段能使用
* Output Gate: 控制是否將這次計算出來的值輸出
* Forget Gate: 控制是否將Memory清除

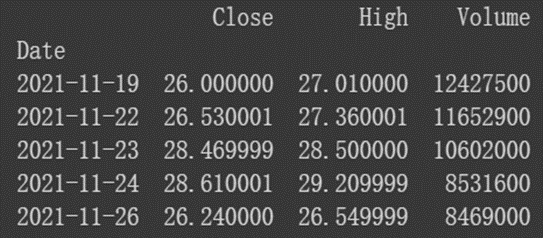
圖五、LSTM模型架構

1. 實驗過程

我們使用LSTM中many-to-one方式來訓練資料來預測2021.7.1 ~ 2021.11.26的收盤股價：

* 關聯係數分析

圖六、Correlation Analysis

* 實驗特徵：Close(收盤價)、High(股票高點)、Volume(成交量)如圖七。

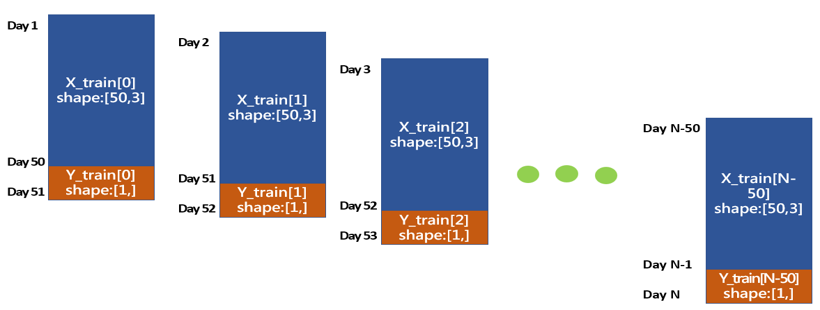
圖七、LSTM模型取用特徵

* 訓練資料

我們將資料切分成train、validation以及test資料做訓練，資料日期：2010.1.4 ~ 2021.11.26總共2997筆，訓練資料切分方式如表一。

表一、訓練資料切分方式

|  |  |  |
| --- | --- | --- |
| **資料集** | **時間** | **Shape** |
| Training Data | 2010.1.4 ~ 2019.12.31 | (2516, 3) |
| Validation Data | 2020.1.1 ~ 2020.12.31 | (377, 3) |
| Testing Data | 2021.7.1 ~ 2021.11.26 | (104, 3) |

Timestep設定：輸入資料的切分方式為104筆資料(Testing Data)中假設每50天為一筆輸入資料來預測下一天資料，例如第一天至第五十天預測第五十一天，方式如圖八。

圖八、訓練資料的輸入資料切分

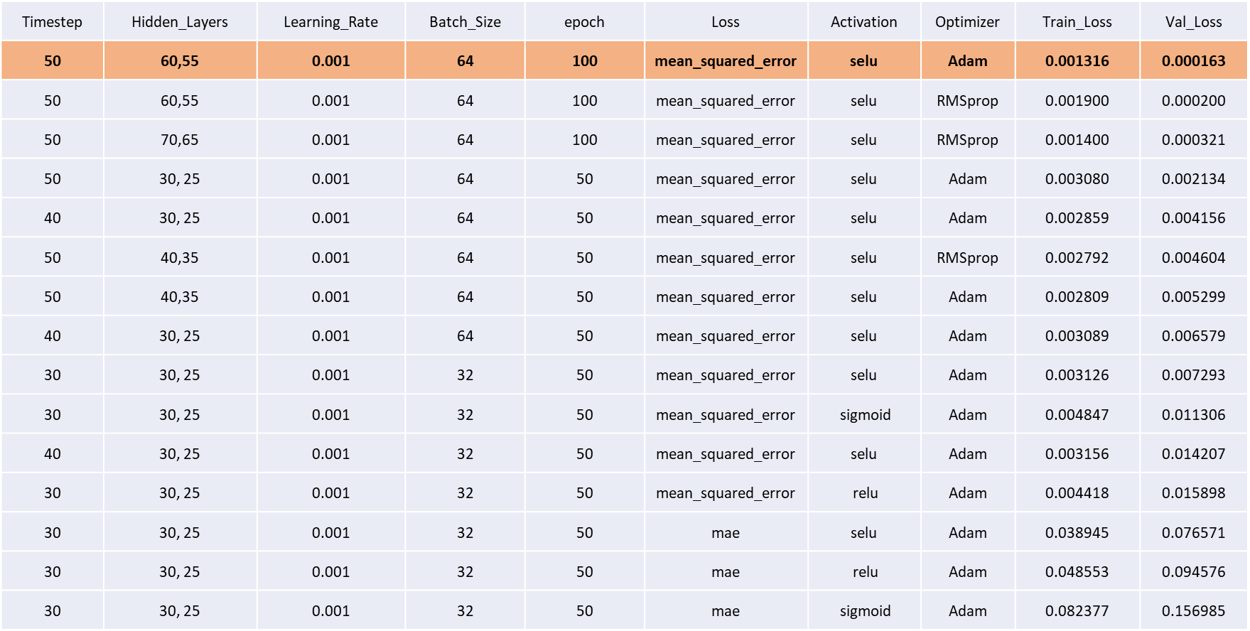
* 交叉驗證

使用時間序列的Cross Validation來進行驗證。

* LSTM模型參數

資料前處理使用了MinMaxScaler：

|  |
| --- |
| from sklearn.preprocessing import MinMaxScaler  sc = MinMaxScaler(feature\_range = (0, 1))  train = sc.fit\_transform(train\_data)  val = sc.fit\_transform(val\_data)  test = sc.transform(test\_data)  print(train.shape,val.shape,test.shape) |

而實驗過稱中透過參數的調整來找出最佳結果，參數有Activation function: selu、relu、sigmoid，Optimizer: Adam、RMSprop，Batch size、Epoch、Hidden layers以及Loss等等如圖九。

圖九、LSTM實驗參數

1. **Results**

我們對APA Corp. 50天的股票價格進行預測，LSTM模型的最佳參數設定如圖十。

圖十、LSTM最佳參數

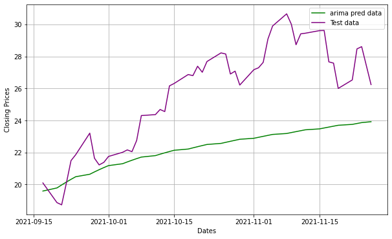
而評估指標計算使用了RSME、MSE以及R-Square Score，結果如表二。

* RMSE(Root Mean Square Error)：均方根誤差
* MSE(Mean Square Error)：均方誤差
* R-Squared：回歸模式之變異值與所有變異量之比例

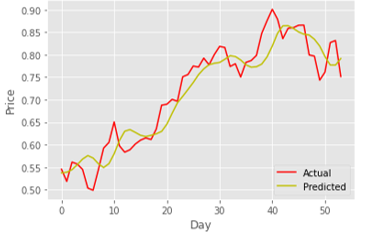
表二、ARIMA和LSTM評估指標結果

|  |  |  |
| --- | --- | --- |
| **評估指標** | **ARIMA(3,0,2)** | **LSTM** |
| RMSE |  | 0.0014981071162395126 |
| MSE | 16.362673218826885 | 0.03870538872352935 |
| R-Squared | -0.5659850824780199 | 0.8876417258075385 |

預測的結果趨勢圖如圖十一、圖十一二



圖十一、RIMA預測結果



圖十二、LSTM預測結果

1. **Discussion**

從實驗結果可以發現，比起傳統的ARIMA模型預測，因LSTM的參數調整彈性以及多樣性，預測出來的結果明顯筆ARIMA好很多，我們在實驗過程也有遇到幾項問題，如下：

1. 時間序列資料的validation set設置

* 實驗需要設置sliding window
* 日期很重要，我們希望拆分數據每個window都包含 X 天
* Test set必須在Train data之後

圖十三、Time-based cross-validation

1. 參數設定問題

一般LSTM多對一(many to one)模型，其參數 return\_sequences 設定為 False ，且不可使用 TimeDistribution。然而本專案，雖未使用 TimeDistribution，但 return\_sequences 卻設為 True，如果改成 False，反而出現錯誤訊息，值得再深入研究。

1. **Future Work**

從實驗的結果發現因少量資料丟進模型去跑，確實只有一個數字，或許many to many可當作本專案後續可以繼續精進的議題。

1. **Reference**
2. G.S. Atsalakis and P.V. Kimon, “Forecasting stock market short-term trends using a neuro-fuzzy methodology”, Expert Systems with Applications, vol. 36, no. 7, pp.10696–10707, 2009.
3. Debadrita Banerjee, “Forecasting of Indian Stock Market using Time-series ARIMA Model”, ICBIM 2014
4. Ayodele A. Adebiyi., Aderemi O. Adewumi and Charles K. Ayo, “Stock Price Prediction Using the ARIMA Model”, UKSim-AMSS 2014
5. Anita Yadava, C K Jhaa and Aditi Sharanb, “Optimizing LSTM for time series prediction in Indian stock market”, ICCIDS 2019
6. Sima Siami-Namini, Neda Tavakoli and Akbar Siami Namin, “A Comparison of ARIMA and LSTM in Forecasting Time Series”, IEEE 2018
7. S&P 500 stocks price with financial statement <https://www.kaggle.com/hanseopark/sp-500-stocks-value-with-financial-statement>
8. Predicting Stock Prices Using Machine Learning <https://neptune.ai/blog/predicting-stock-prices-using-machine-learning>
9. Prediction of price for ML with finance stats <https://www.kaggle.com/hanseopark/prediction-of-price-for-ml-with-finance-stats/data>
10. Time-Series Forecasting: Predicting Stock Prices Using An LSTM Model <https://towardsdatascience.com/lstm-time-series-forecasting-predicting-stock-prices-using-an-lstm-model-6223e9644a2f>
11. Berkshire Hathaway - Stock Time Series Analysis <https://www.kaggle.com/kalilurrahman/berkshire-hathaway-stock-time-series-analysis>
12. A Multivariate Time Series Modeling and Forecasting Guide with Python Machine Learning Client for SAP HANA <https://blogs.sap.com/2021/05/06/a-multivariate-time-series-modeling-and-forecasting-guide-with-python-machine-learning-client-for-sap-hana/>
13. Understanding LSTM Networks

<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

1. LSTM深度學習股價預測

<https://medium.com/data-scientists-playground/lstm-%E6%B7%B1%E5%BA%A6%E5%AD%B8%E7%BF%92-%E8%82%A1%E5%83%B9%E9%A0%90%E6%B8%AC-cd72af64413a>

1. **Team Work**
2. 110753201曹昱維: Data Preprocessing、ARIMA Model、Slides
3. 109753207謝政彥: LSTM Model、Slides
4. 110753126鄭詠儒: Transformer Model(未完成)、Slides、Presentation、Report