台灣股價預測

一、動機

股票市場趨勢預測一直是金融領域的一大研究焦點，在投資決策中扮演關鍵角色。傳統的預測方法主要仰賴技術分析和基本面分析，然而，這些方法存在一些限 制。技術分析往往過度依賴歷史數據，而基本面分析則需要大量財務數據和對市場的高度理解。

近年來，隨著機器學習技術的崛起，越來越多的研究者開始將機器學習應用於股票市場趨勢預測。機器學習方法具有學習非線性關係的能力，並能夠自動提取數據特徵，因而提高了預測的精確性。

這次的報告主要在建立一個基於機器學習的股票市場趨勢預測模型，以預測台灣五檔代表性股票未來十天的股價走勢。我們將利用這五檔股票所有過去上市的股價資料來訓練模型，並使用2023年1月1日之後的資料進行測試。

二、資料集敘述

我們選擇台灣五家具有代表性的上市公司，以深入研究不同產業領域的股價變動。包括日月光半導體製造股份有限公司(3711)、元大台灣50(0050)、台灣積體電路製造公司(2330)、聯發科技股份有限公司(2454)和聯華電子股份有限公司(2303)。涵蓋了半導體製造、大型股票指數基金以及其他電子產品製造領域。

而資料時間範圍分為兩個主要部分，以確保模型的準確性和適應性。首先是訓練資料，包含從過去上市開始追溯至2023年1月1日的資料。具體而言，日月光的訓練資料時間範圍為2018年4月30日至2023年1月1日，元大台灣50為2003年6月30日至2023年1月1日，台灣積體電路製造公司為1994年9月5日至2023年1月1日，聯發科技為2001年7月23日至2023年1月1日，而聯華電子則為1985年7月16日至2023年1月1日。

其次是測試資料，涵蓋從2023年1月1日至目前的資料，這部分資料將用來評估我們建立的預測模型對從今天開始未來十天股價的預測效果。這樣的時間範圍旨在考慮歷史資料的重要性，同時也確保模型的應用能夠適應最近的市場情境。

三、分析工具

本次作業使用三種 model來進行預測：

1. LSTM：

   LSTM，或者長短時記憶（Long Short-Term Memory），是一種特殊類型的循環神經網絡（Recurrent Neural Network, RNN），用於處理和學習序列數據。LS TM的主要特點是能夠捕捉和保持長期依賴性，這使得它在處理具有長距離相依性的序列數據時表現出色。

2. XGBoost：

   XGBoost（eXtreme Gradient Boosting）是一種梯度提升樹（Gradient Boosti ng Tree）的機器學習算法。它在預測建模和機器學習競賽中獲得了廣泛的應 用，因為它在效能和準確性上都表現出色。

3. ARIMA：

   ARIMA（AutoRegressive Integrated Moving Average）是一種時間序列分析和預測的統計模型。ARIMA模型結合了自回歸（AR）模型、整合（I）模型和移動平均（MA）模型的特點，它的應用範圍涵蓋了多種時間序列數據，包括經濟、金融、氣象、銷售等領域。

四、實作與評估方法

實作方法：

首先先用爬蟲將選擇之股票與其股價爬下來，每次挑選其中一支股票進行預測，以 LSTM 來說，輸入資料有七個，分別是 stockNo, TrainDateStart, TrainDateEnd, TrainDateEnd, year, month, day，輸出資料有四個，分別是 date_actual, LSTMdate_predict, actual_price, LSTM_predict；以 XGBoost來說，輸入資料有七個，分別是 stockNo, TrainDateStart, TrainDateEnd, T rainDateEnd, year, month, day，輸出資料有四個，分別是 date_actual, XG Boostdate_predict, actual_price, XGBoost_predict。以 ARIMA 來說，輸入

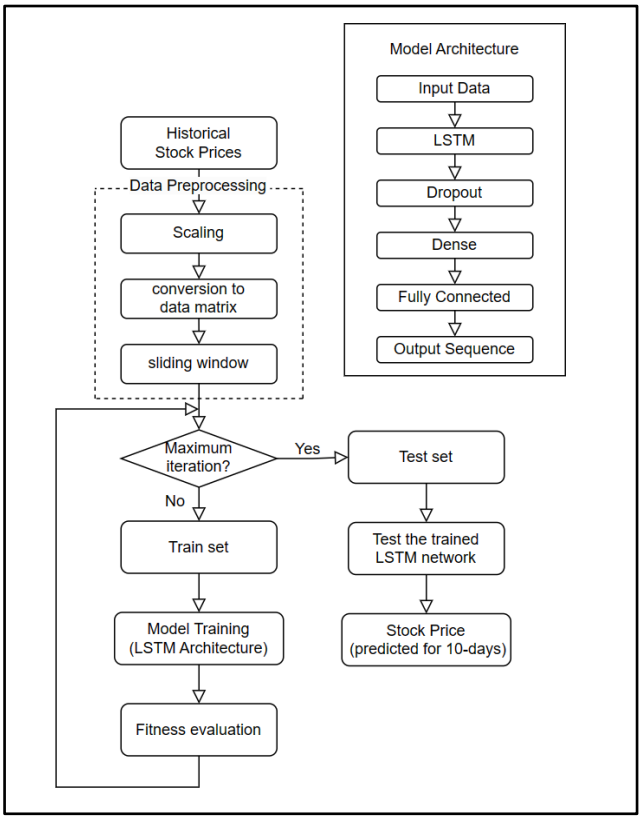
資料有七個，分別是 stockNo, TrainDateStart, TrainDateEnd, TrainDateEnd, year, month, day，輸出資料有一個，是預測值。

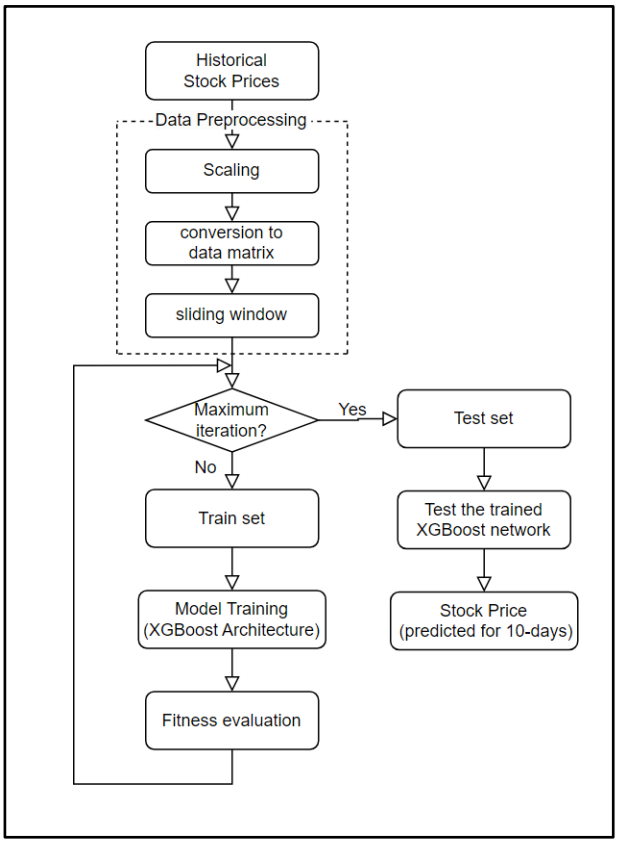
評估方法：

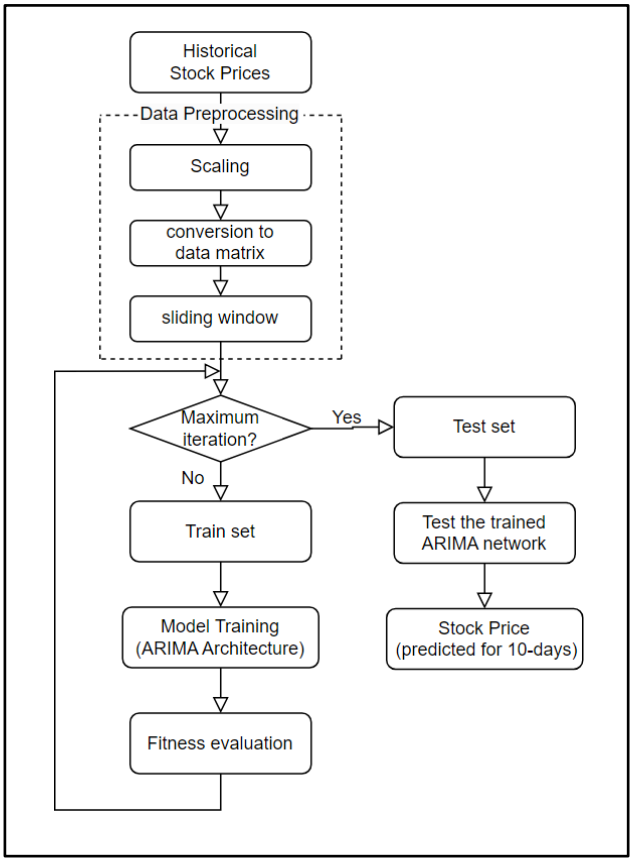
本次作業使用 mae 以及 mse 作為評估標準，以下為其介紹： MAE（Mean Absolute Error）是一種用於評估預測模型性能的指標，特別適用於回歸問題。它衡量模型預測值與實際觀測值之間的平均絕對誤差，即預測值和實際值之間的差異的絕對值的平均值。

MSE（Mean Squared Error）是一種用於評估預測模型性能的指標，特別適用於回歸問題。它衡量模型預測值與實際觀測值之間的平均平方誤差，即預測值和實際值之間差異的平方的平均值。

五、流程圖




Historical
Stock Prices
Data Preprocessing
Scaling
conversion to
data matrix
sliding window
Maximum
iteration?
Yes
No
Test set
Train set
Test the trained
XGBoost network
Model Training
(XGBoost Architecture)
Stock Price
(predicted for 10-days)
Fitness evaluation


Historical
Stock Prices
Data Preprocessing
Scaling
conversion to
data matrix
sliding window
Maximum
iteration?
Yes
No
Test set
Train set
Test the trained
ARIMA network
Model Training
(ARIMA Architecture)
Stock Price
(predicted for 10-days)
Fitness evaluation

六、分析結果與結論

1. 預測模型：LSTM, XGBoost, ARIMA
2. 預測時間：測試集中2023/12/27至2024/01/09的10天（非假日）
3. 預測結果分析-訓練集資料量及預測結果：

表一為五家公司股價預測結果與實際股價的MAE，按照公司上市的時間升冪排列，代表開始採用為訓練集的年份，也代表訓練集資料量從大到小的排列。
從表一可以觀察到股價**資料量最大**的兩支股票：2023及2330，LSTM model在第二大訓練資料量的項目預測結果比XGBoost model的還要差，在第一大訓練資料量的項目預測結果雖較好，但跟XGBoost model的預測結果只有一點差距。
股價**資料量最小**的三支股票：2454, 0050, 3711，則是由**LSTM model**預測的結果最好。
股價訓練集最小的情況下（股票代號3711），LSTM model為表現最好的模型，跟XGBoost model相差甚遠，可推測資料量不多情況下適合使用的LSTM model。

	LSTM	**XGBoost**	**ARIMA**
2303 (1985)	1.09	1.13	8.00
2330 (1994)	10.74	5.49	122.95
2454 (2001)	30.47	44.07	327
0050 (2003)	0.89	1.11	21.80
3711 (2018)	1.68	7.29	36.33

(表一)誤差值MAE

4. 預測結果分析 - 股票走勢圖、誤差值：

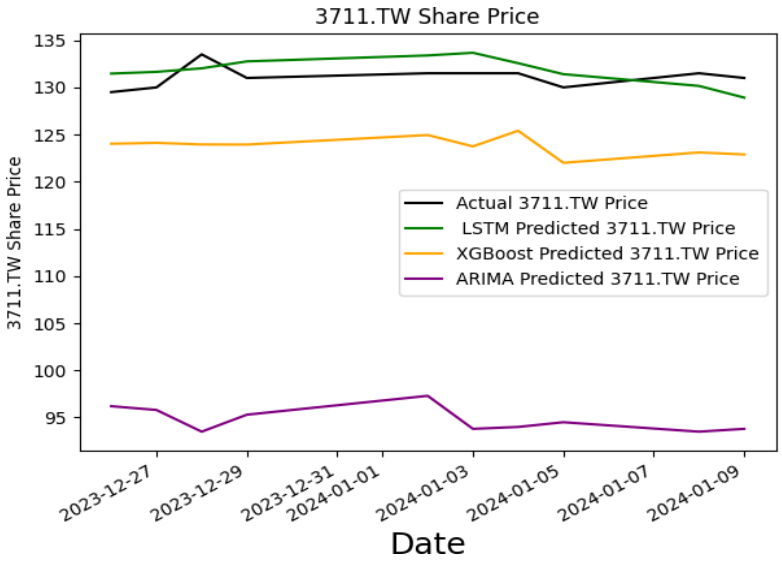
(1). 日月光半導體製造股份有限公司(3711)：

依據下圖一（股票走勢圖）可以發現模型預測表現最差的為ARIMA model，低估了股票價格，依據下圖二（誤差值表格）ARIMA model預測值與與實際價格的誤差平均大約為36元。

ARIMA model為是唯一的統計模型，與其他兩種機器學習模型表現差異甚大。

XGB oost model也是普遍低估了實際股價，但表現比ARIMA好，其預測值與實際股價平均差距為大約7元。

LSTM為3種模型中表現最好的，與其與實際股價平均差距為約1.6元。

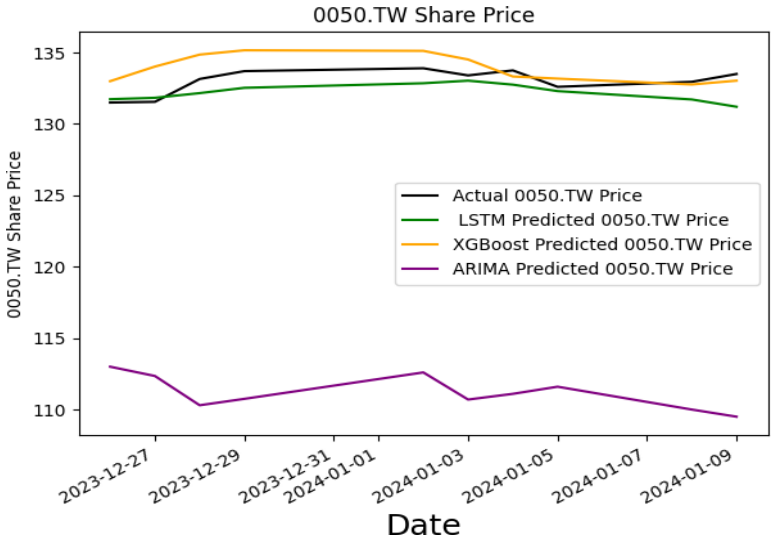



```
mae:    LSTM_mae:  1.6800247192382813 / XGBoost_mae:  7.285836791992187 / ARIMA_mae:  36.32999877929687
mse:    LSTM_mse:  2.9363353295484558 / XGBoost_mse:  54.585785614571066 / ARIMA_mse:  1323.8289110717828
```

(2). 元大台灣50(0050)：

股票代碼0050的訓練集為這次實驗中第二小的。

XGBoost model能比較準確的預測較新的股價。

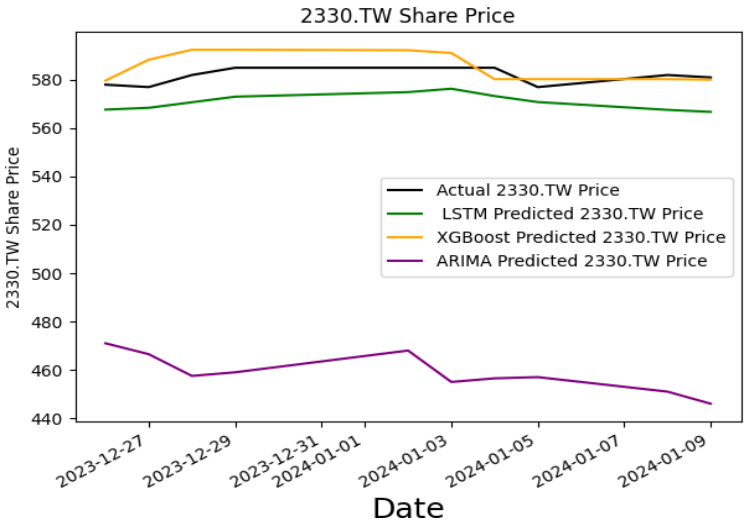



```
mae:    LSTM_mae:  0.8918548583984375 / XGBoost_mae:  1.1120010375976563 / ARIMA_mae:  21.809999084472658
mse:    LSTM_mse:  1.1590881284326315 / XGBoost_mse:  1.6871695907087996 / ARIMA_mse:  478.54195332338566
```

(3). 台灣積體電路製造公司（2330）：

訓練集為這次實驗中第二大的。

XGBoost model能比較準確的預測較新的股價，並在此支股票的預測中，表現得較LSTM model好。



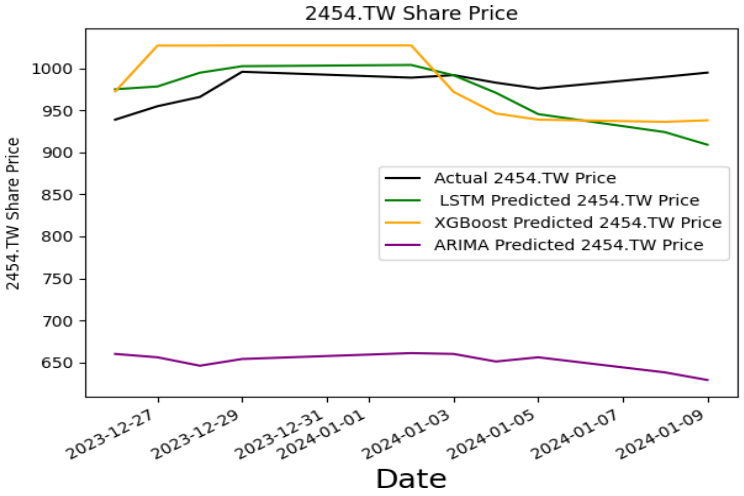

```
mae:    LSTM_mae:  10.742584228515625 / XGBoost_mae:  5.49176025390625 / ARIMA_mae:  122.95
mse:    LSTM_mse:  121.28657092638313 / XGBoost_mse:  42.24291765019298 / ARIMA_mse:  15192.275
```

(4). 聯發科技股份有限公司(2454)：

訓練集為這次實驗中訓練集中等的。

LSTM model跟XGBoost model在較新的股價預測中，都低估了股價。

LSTM model仍為表現最好的模型。

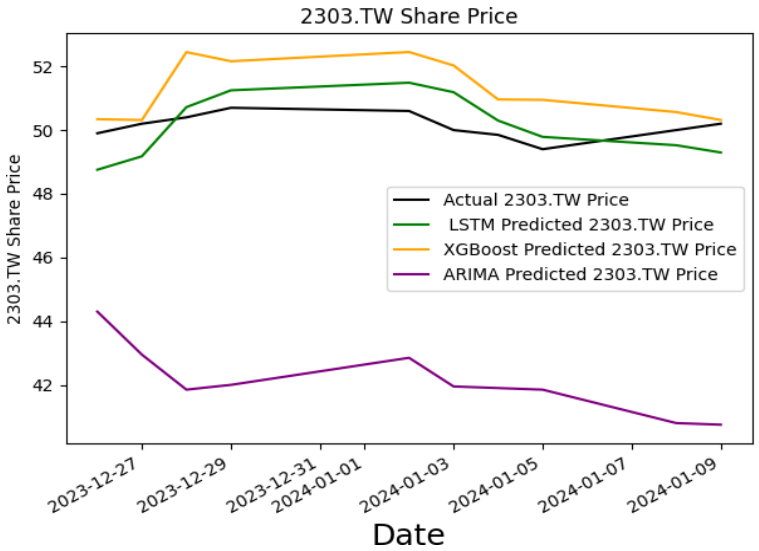



```
mae:    LSTM_mae:  30.468310546875 / XGBoost_mae:  44.07120971679687 / ARIMA_mae:  327.0
mse:    LSTM_mse:  1575.284791342169 / XGBoost_mse:  2175.0863592091946 / ARIMA_mse:  107489.8
```

(5). 聯華電子股份有限公司(2303)：

訓練集為這次實驗中最大的。

XGBoost model跟LSTM model普遍高估股價，LSTM model仍為表現最好的模型。




```
mae:    LSTM_mae:  0.732883071899414 / XGBoost_mae:  1.1292179107666016 / ARIMA_mae:  8.005000686645507
mse:    LSTM_mse:  0.6357854727466474 / XGBoost_mse:  1.8053007763475761 / ARIMA_mse:  65.17276048279163
```

5. 結論

LSTM model在此次的實驗中表現最佳，為最適合作為預測股票市場趨勢的模型。LSTM model表現比XGBoost model好的原因或許為其能找出影響股票的關鍵因素：利用forget gate（遺忘閥）篩選不重要的資訊，並記憶對未來有預測價值的資訊。