

HW4 郭盈均 312704004

1. Experiment with different window sizes and steps. Train the model using 3 different combinations of window size and step. Evaluate the Mean Squared Error (MSE) for each configuration. Report the MSEs using a table and analyze the results.

Configuration	window sizes	steps	MSE
0 (Original)	10	15	145.32
1	15	10	25.99
2	15	5	7.2
3	10	5	12.10
4	20	5	4.16

當 step size > window sizes，因為會有一些資料被跳過，使得 MSE 較大，我們將 window size = 15，steps = 10 (如 2)，MSE 明顯變小，且隨著 window sizes 變大，step size 變小，會有不錯的效果。以下實驗設 window sizes=20, steps=5

2. (i) Include 'Volume' as an additional input feature in your model. Discuss the impact of incorporating 'Volume' on the model's performance.

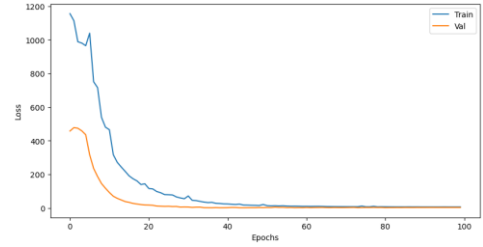
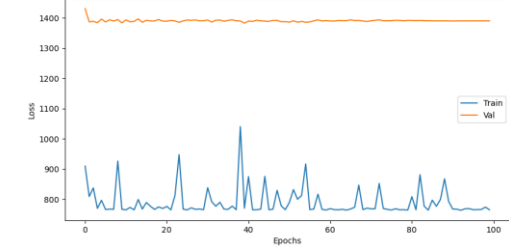
將"Volume"當作 input feature 時，會發現 Valid_Loss 沒有明顯下降，訓練不起來，我認為 Volume 的 scale 太大，與開盤價、收盤價等其他 feature 比較起來，反而影響性能。

- (ii) Explore and report on the best combination of input features that yields the best MSE. Briefly describe the reasons of your attempts and analyze the final, optimal input combination (window sizes=20, steps=5)

Combinations	Input features	MSE
0 (Original)	High, Low, Open, Close	4.16
1	Open, High, Low	2.3037
2	High, Low	9.3484
3	High	7.5659

由於是要預測股價高點，我認為和開盤價與收盤價較無關聯，實驗中透過拿掉收盤價和開盤價，來觀察 MSE 變化。在 window sizes=20, steps=5 下，最好的 input feature 為 Open, High, Low。

3. Analyze the performance of the model with and without normalized inputs in Lab 4.
4. You can use experimental results or external references (which must be cited) to support your conclusions on whether normalization improves the model's performance.

	未標準化	標準化
Input Features Setting	<pre>features = df[['Open', 'High', 'Low', 'Close']] labels = df['High'].shift(-1) # Next day's high price</pre>	<pre>features = norm_data[['Open', 'High', 'Low', 'Close']] labels = df['High'].shift(-1) # Next day's high price</pre>
Loss Picture		
MSE	1.0968	1382.6162

使用標準化的目的是希望能提高效能，並且消除一些規模不同所帶來的影響。然而從實驗結果可以得到，經過標準化後的效果明顯較差，沒有學到應該要學的特徵。由於開盤價、收盤價、高低點，都在同一個規模下，並沒有資料偏斜等問題，因此原來的數據已經有高度可解釋性，因此沒有標準化的必要。

Ref/ Data Normalization in Machine Learning

<https://www.almabetter.com/bytes/tutorials/data-science/normalization-in-machine-learning>

4. Why should the window size be less than the step size in Lab 4? Do you think this is correct? If you use external sources, please include references to support your response.

如果 step size 比 window size 還要大，會導致資料不重疊。這代表每個窗口所使用的資料都是不同的，沒有重複使用的部分，這可能會影響模型的訓練效果。通常情況下，我們希望設置 step size 小於 window size，以確保資料有重疊，這樣可以使模型能夠更好地學習時間序列資料中的模式和趨勢。

5. Describe one method for data augmentation specifically applicable to time-series data. Cite references to support your findings. (Approximately 100 words.)

對於時序數據，一些常見的數據增強技術可能包括：

- 時間平移：將整個時序資料向前或向後平移一個固定的時間步長，以模擬在不同時間點收集到的資料。
- 噪聲注入：在時序資料中加入隨機噪聲，以模擬在現實環境中的干擾。
- 剪切和縮放：對時序資料的一部分進行剪切或縮放操作，以模擬不同時

間尺度的資料。

Ref/ CSDN: 時間序列資料如何進行資料增強？

6. Discuss how to handle window size during inference in different model architectures:

(i) Convolution-based models

CNN 通常用於圖像識別和處理任務，但它也可以用於處理時間序列數據，在 CNN 模型中，窗口大小通常由卷積核的大小決定。在推斷期間，可以將整個序列作為輸入，並在每個時間步應用卷積操作，然後使用池化層對卷積結果進行彙聚。這樣就可以覆蓋整個序列並生成預測。

(ii) Recurrent-based models

RNN 是一種神經網路演算法，其核心思想是利用前一時刻的輸出作為後一時刻的輸入，從而考慮了序列數據前後時刻之間的關聯性。通常採用滑動窗口的方式處理輸入序列。具體來說，我們從序列的開頭開始，每次移動一個時間步，將當前時間步及其之前的窗口作為模型的輸入。

(iii) Transformer-based models

透過位置編碼建立時間序列中不同的關係後，在每個時間步 Transformer 模型使用自注意力機制來計算每個位置與序列中其他位置的注意力分數。使得模型能夠捕捉到不同時間點之間的依賴關係，而無需固定大小的時間窗口。

Ref/ 稀土掘金：RNN 实现时间序列预测、CNN 实现时间序列预测、ChatGPT