元智大學工業工程與管理學系 畢業專題
以機器學習演算法建立時間序列資料預測模型

指導老師：鄭春生 教授　　學生：吳俊毅 蔡鎮安 吳昀澤

研究背景與動機

時間序列 (time series) 是一種依照時間順序記錄的數據集合。它是由一系列按照連續時間點或是時間間隔收集的數據組成，按照均勻間隔的序列，例如每天、每月或每年等。時間序列可用於描述和分析，資料隨時間變化的圖像、趨勢、季節性和周期，在現代社會中廣泛應用各種產業，像是物價指數、金融市場、供應鏈管理等領域，都需要對時間序列進行預測與分析並且時間序列能幫助我們了解數據。

本專題之動機來自於時間序列預測的重要性，在過去傳統的統計模型一直是時間序列預測的主要工具。機器學習之技術發展發展迅速，新的預測方法例如卷積神經網路 (convolutional neural networks, CNN) 和長短期記憶網路 (long short-term memory, LSTM) 因為其優異的預測表現，在近期受到了大眾的討論與關注，所以本實驗將會利用這兩種預測方法發展出來的卷積長短期記憶神經網路 (convolutional long short-term memory neural network, CNN-LSTM) 模型，來測試與傳統統計模型季節性差分整合移動平均自我迴歸模型 (seasonal autoregressive integrated moving average model, SARIMA)，進行比較評估何種模型在資料預測上表現更佳。

研究方法

・卷積神經網路 (convolutional neural networks, CNN)

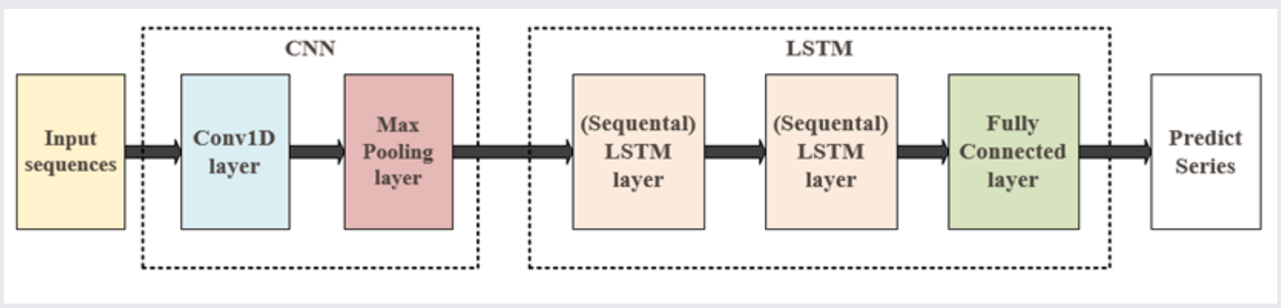
CNN是一種深度神經網路，廣泛應用於處理圖像處理，常用於機器視覺領域。CNN 是一種層次化的模型，能在分析時間序列資料使用原始資料，由多個層次的神經元組成，每一層都對輸入資料進行不同的轉換和處理，最終產生輸出結果。

・長短期記憶 (long short-term memory，LSTM)

LSTM是一種特殊的RNN，它由一組具有特徵的單元集合組成，利用這些特徵來記憶數據序列，集合中的單元用於捕獲並存儲數據流。此外，集合中的單元構成先前的模塊與當前的模塊的內部互連，從而將來自多個過去時間瞬間的信息傳送給當前的模塊。

・卷積長短期記憶神經網路 (convolutional long short-term memory neural network, CNN-LSTM)

CNN-LSTM是結合CNN 和長短期記憶網路LSTM 的混合模型，CNN-LSTM模型的運作分為數據準備、CNN層、LSTM層、輸出層與編譯能夠處理影像及圖片辨識以及時間預測分析。CNN-LSTM模型利用了CNN提取輸入子序列的空間特徵的能力。然後將這些特徵被傳遞給LSTM，而利用LSTM學習理解數據的序列性質，最終輸出由密集層生成，提供數據的預測分析結果，另外也解決了LSTM較長的運算時間。



CNN-LSTM架構圖 (資料來源: 本研究所整理)

研究績效指標

為了評估傳統統計預測方法SARIMA與機器學習演算法CNN-LSTM的預測結果，我們使用以下績效指標做為評估衡量之依據。

平均絕對誤差 (mean absolute error, MAE)：計算觀察值與其實際值和預測值之間的絕對誤差，如公式所示MAE越小，表示模型的預測越準確。

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

均方誤差 (mean square error, MSE)：計算實際觀測值和模型預測值之間的平方誤差的平均值，如公式所示表示模型的預測越準確。

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

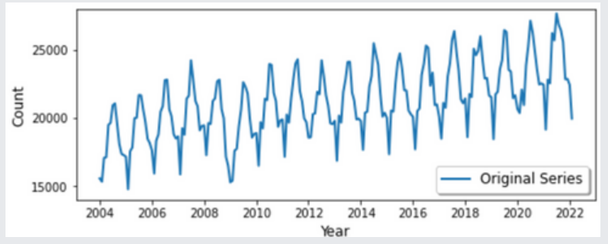
平均絕對百分比誤差 (mean absolute percentage error, MAPE)：計算實際觀測值的相對誤差量化為百分比，如公式所示相較MAE與MSE能更直觀地衡量預測的準確度。

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}(|y_i - \hat{y}_i|)/y_i$$

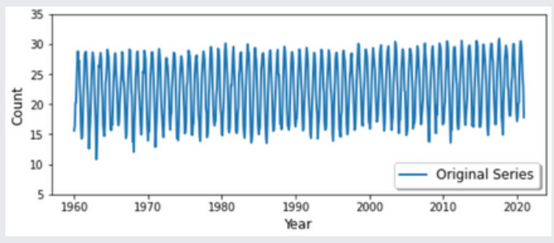
研究資料

本研究使用的資料取自於內政部政府資料開放平台與Kaggle平台，共三組資料，分別為2004年至2022年每月份電力供給量、1960年至2022年臺北市月均氣溫以及1956年至1995年澳洲啤酒月銷量；資料總筆數分別為217筆、721筆與465筆。

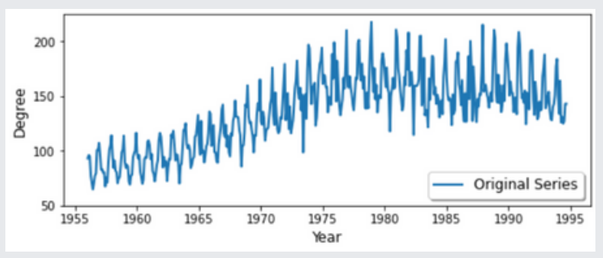
資料集	(訓練用筆數, 預測用筆數)	(history, future)
電力供給量	(173, 44)	(12, 1)
臺北市月均溫	(576, 145)	(12, 1)
澳洲啤酒月銷量	(372, 93)	(12, 1)



電力供給量資料



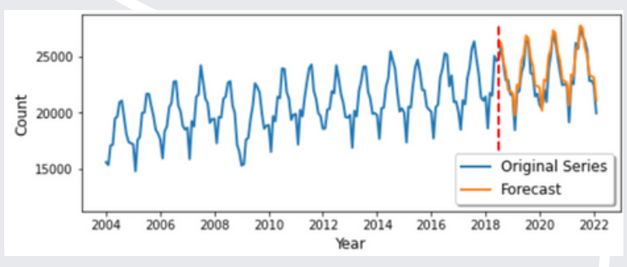
臺北市月均溫資料



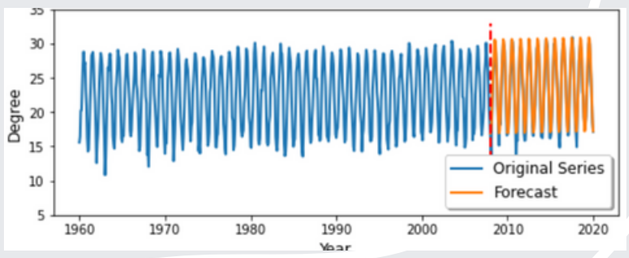
澳洲啤酒月銷量資料

預測結果

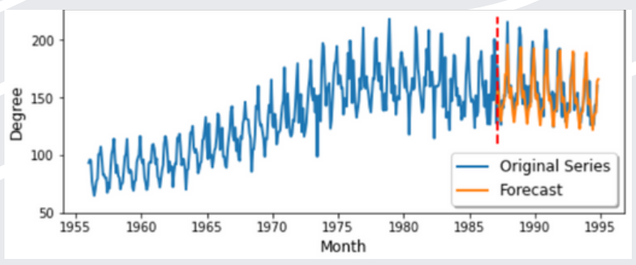
SARIMA之預測結果



SARIMA電力供給量預測結果

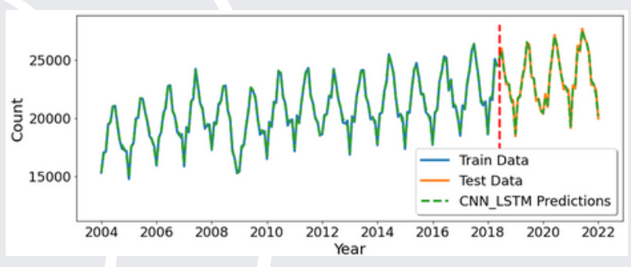


SARIMA臺北市月均溫預測結果

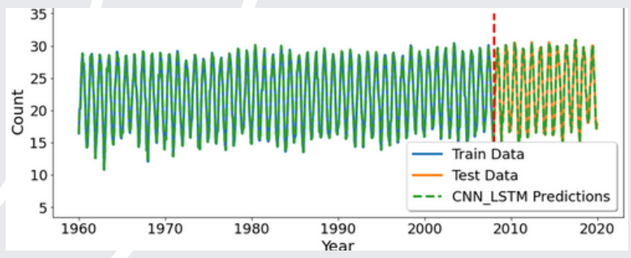


SARIMA澳洲啤酒月銷量預測結果

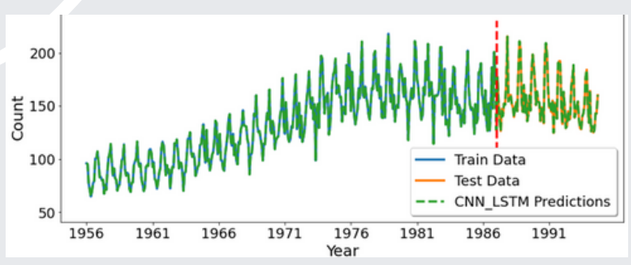
CNN-LSTM之預測結果



CNN-LSTM電力供給量預測結果



CNN-LSTM臺北市月均溫預測結果



CNN-LSTM澳洲啤酒月銷量預測結果

預測績效指標結果

Model	MSE	MAE	MAPE
電力供給預測 (SARIMA)	700491.9396	718.6157	0.0317
電力供給預測 (CNN-LSTM)	2008.0319	36.1658	0.0016
月均溫預測 (SARIMA)	8.0757	2.3726	0.1102
月均溫預測 (CNN-LSTM)	0.0010	0.0232	0.0011
啤酒月銷量預測 (SARIMA)	575.5812	19.2328	0.1193
啤酒月銷量預測 (CNN-LSTM)	0.0481	0.1574	0.0010

預測績效指標結果

將三份資料分別以兩種模型進行預測後，可以觀察到三份資料使用CNN-LSTM預測的誤差皆比SARIMA低，以本次實驗而言，機器學習之預測績效要高於傳統統計模型。

本研究經過三次實驗後，發現CNN-LSTM之預測績效皆優於SARIMA模型。SARIMA建構較容易且預測過程快速，然而需針對不同資料進行前處理以選擇合適模型，且預測非穩定之時間序列會有較大的預測誤差。CNN-LSTM模型針對資料進行訓練，隨迭代數增加，預測績效越準確，並且適用於預測非穩定之時間序列。

我們認為選擇CNN-LSTM作為時間序列模型會是較佳的選擇。未來的研究可考慮對更大筆數或更具隨機性質的資料進行預測，透過更多評估比較進以加強我們的結論。時間序列在金融、供應鏈管理等產業都有廣泛應用，未來利用機器學習可進行更準確之估計與預測，勢必對產業發展提供相當大的幫助。