國立政治大學111學年度下學期

金融科技概論

期末報告（二）

利用時間序列模型與深度學習演算法預測台灣出生人口數

指導教授：廖四郎 教授

學生：111753145 資科碩一 劉育佑

中華民國 112 年 6 月

摘要

本篇報告旨在利用時間序列模型（AR、MA、ARMA等）與深度學習演算法預測台灣的出生人口數。出生人口數是一個重要的社會指標，對於政府制定人口政策和社會福利計劃具有重要意義。然而，由於出生人口數受到多種因素的影響，如經濟環境、社會變遷、醫療條件等，其變化具有一定的不確定性和複雜性，除此之外，台灣的少子化問題非常嚴重，透過此報告也能夠對於未來的出生人口數有提前的準備。

為了解決這一挑戰，我們首先利用了基於時間序列模型來預測出生人口數，在獲得較不好的實驗結果後，再深度學習演算法的方法來預測台灣的出生人口數。我們利用中華民國政府官方的出生人口數來建立預測模型。

長短期記憶（LSTM）模型是一種適用於處理時間序列數據的強大模型。通過對歷史資料的學習，LSTM模型能夠捕捉到時間序列的長期依賴性和隱藏的模式，並據此進行準確的預測。在實驗中，我們將資料集分為訓練集和測試集，使用訓練集來訓練LSTM模型，並使用測試集來評估模型的預測準確性。

根據我們的實驗結果，時間序列模型的表現不盡理想，而LSTM模型卻能夠準確預測台灣的出生人口數。在這篇報告中，除了理解時間序列模型的知識之外，同時也理解了時間序列模型在人口預測領域的應用提供了實證和啟示。

關鍵詞：時間序列模型、深度學習、長短期記憶、預測、台灣、出生人口數

目錄

一、動機

過去幾十年來，台灣面臨著人口結構的變化和少子化的挑戰。隨著生育率的下降和老年人口比例的上升，人口結構的改變對社會、經濟和福利體系帶來了重大影響。因此，對台灣出生人口數進行準確的預測對於政府制定相應的人口政策和社會福利計劃至關重要。

過去的研究主要使用傳統的統計模型來預測人口數量，但這些模型通常無法充分考慮到時間序列中複雜的關係和隱含的模式。近年來，深度學習演算法的快速發展為我們提供了一種新的解決方案，這些演算法可以自動學習從未接觸過的時間序列模型。

利用時間序列模型和深度學習演算法，我們可以更好地理解和預測台灣的出生人口數。這種方法能夠捕捉到時間序列中的長期依賴性和隱藏的模式，並根據過去的趨勢和相關因素進行準確的預測。

此外，透過對出生人口數進行深入的研究和預測，我們可以更好地理解少子化現象的根本原因和影響因素。這將有助於政府和相關機構制定針對性的政策和措施，以應對少子化帶來的社會和經濟挑戰。

因此，本篇報告旨在運用時間序列模型和深度學習演算法，以預測台灣的出生人口數。通過這項研究，我們將能夠學習和應用新的技術方法，同時也能夠深入了解少子化現象的趨勢和影響，為相關政策的制定和社會福利的規劃提供更多的知識和洞察力。

二、時間序列模型簡介

時間序列模型是一種用於預測和分析時間序列數據的統計模型。時間序列數據是按照時間順序收集的數據點序列，例如每日股價、每月銷售量或每年的總體經濟指標。時間序列模型可以幫助我們理解和解釋數據中存在的趨勢、季節性和其他隨時間變化的模式。

以下是一些常見的時間序列模型：

1. 自回歸模型（Autoregressive Model，AR）：

AR模型是基於過去時間點的觀測值來預測未來值的模型。它假設當前觀測值與前一時間點的觀測值之間存在線性關係。AR模型的預測依賴於過去時間點的殘差項，也稱為自回歸項。

2. 移動平均模型（Moving Average Model，MA）：

MA模型基於過去時間點的殘差項來預測當前值。它假設當前觀測值與前一時間點的殘差項之間存在線性關係。MA模型用於捕捉時間序列中的隨機波動。

3. 自回歸移動平均模型（Autoregressive Moving Average Model，ARMA）：

ARMA模型結合了AR模型和MA模型的特性。它同時考慮了過去觀測值和殘差項的影響，以預測未來值。ARMA模型能夠處理時間序列中的趨勢和隨機波動。

4. 差分自回歸移動平均模型（Autoregressive Integrated Moving Average Model，ARIMA）：

ARIMA模型在ARMA模型的基礎上引入了差分運算，用於處理非平穩時間序列數據。差分操作可以將非平穩序列轉換為平穩序列，使其更適合應用ARMA模型進行建模和預測。

這些時間序列模型可以根據數據的特點和預測的需求進行選擇和調整。例如，ARIMA模型常用於預測具有趨勢和季節性的時間序列數據，而AR模型則適用於捕捉時間序列中的自相關結構。這些模型的參數估計和預測可以使用最大概似估計、最小二乘法或其他統計方法進行。

在模型參數的部分，主要可透過ACF、PACF或者AIC、BIC來進行判斷：ACF自相關函數（Autocorrelation Function）和PACF 偏自相關函數（Partial Autocorrelation Function，PACF）是用於分析時間序列數據中自相關結構的工具。

首先自相關函數（ACF）可衡量時間序列觀測值與其在不同時間點的自身觀測值之間的相關性。ACF通常用於檢測和驗證時間序列數據是否存在自相關結構。ACF的值範圍在-1到

1之間，越接近1表示正相關，越接近-1表示負相關，接近0則表示無相關。ACF圖表是描述ACF值隨著時間延遲的變化情況。

而偏自相關函數PACF則衡量了時間序列觀測值與其在特定時間點的自身觀測值之間的相關性，排除了其他時間點的干擾。PACF的計算是基於移除了其他時間點的影響，專注於兩個時間點之間的相關性。PACF圖表提供了直接衡量時間序列中特定時間延遲與當前觀測值之間關係的信息。

ACF和PACF圖表常用於時間序列模型的診斷和參數估計過程中。觀察ACF和PACF圖表可以幫助我們識別時間序列中的自相關結構和季節性。例如，在ACF圖表中，如果自相關係數在某個時間延遲處截尾，這可能暗示著AR模型的適用性。而在PACF圖表中，如果偏自相關係數在某個時間延遲處截尾，這可能暗示著MA模型的適用性。

AIC（Akaike Information Criterion）和BIC（Bayesian Information Criterion）則是兩種常用的模型選擇準則，用於在統計建模中評估和比較不同模型的適合程度。

首先 AIC（Akaike Information Criterion）是由日本統計學家赤池弘次（Hirotugu Akaike）於1973年提出的。AIC基於最大概似估計和信息理論的概念，用於比較不同模型對數據的擬合程度。AIC越小表示模型對數據的擬合越好。AIC考慮了模型的拟合优度和复杂度，透過平衡模型拟合的好壞和模型中使用的參數个数，提供了一種衡量模型的相對優劣的指標。

而BIC（Bayesian Information Criterion）則是由斯洛文尼亞統計學家Hirotsugu Akaike於1978年引入的。與AIC相似，BIC也是一種模型選擇準則，用於評估和比較不同模型的適合程度。BIC在考慮模型拟合优度和模型复杂度的基礎上引入了一個更強的懲罰項，根據貝葉斯信息理論原則，更嚴格地對模型的复杂度進行控制。BIC越小表示模型對數據的擬合越好，且模型越簡單。

AIC和BIC的主要差異在於對模型複雜度的處理。AIC傾向於選擇更複雜的模型，對於擬合程度的提升更加敏感，而BIC則更加偏向於選擇更簡單的模型，對模型的複雜度給予更大的懲罰。

在實際應用中，選擇使用AIC還是BIC取決於具體的研究目的和假設。如果關注的是擬合程度，可以傾向使用AIC；如果更注重模型的簡單性和解釋性，可以傾向使用BIC。無論選擇哪一種準則，它們都提供了一個有用的方式來比較和評估不同模型之間的優劣。

在接下來的實驗中，我們會利用程式做出ACF和PACF圖表並計算AIC與BIC參數，來選擇適用的模型。

三、LSTM演算法簡介

LSTM（Long Short-Term Memory）是一種循環神經網絡（Recurrent Neural Network，RNN）的變體，專門設計用於處理時間序列數據。

傳統的RNN在處理長期依賴性時常遇到「梯度消失」或「梯度爆炸」的問題，即在反向傳播過程中，梯度在時間步長中的傳播逐漸衰減或增長，導致長期記憶和學習能力的限制。LSTM被提出來克服這些問題。

LSTM引入了稱為「記憶單元」（memory cell）的結構，通過一系列稱為「門控」（gates）的機制來控制信息的流動和遺忘。LSTM的記憶單元由三個主要的門控組成：

1. 遺忘門（Forget Gate）：控制之前記憶的保留程度，決定過去記憶對當前時刻的影響。
2. 輸入門（Input Gate）：決定當前時刻新信息的重要程度，以及是否更新記憶單元中的內容。
3. 輸出門（Output Gate）：控制輸出的選擇性，決定當前時刻記憶單元中的信息是否對外部可見。

這些門控結構有效地調節了記憶單元的內部狀態，使得LSTM能夠有效地捕捉長期依賴性和學習長期模式。LSTM模型可以進行長期記憶和短期記憶的控制，使其在處理時間序列數據時具有優越的能力。

LSTM在各種領域中取得了廣泛的應用，包括語言處理、語音識別、機器翻譯、時間序列預測等。它具有強大的建模能力和靈活性，能夠處理具有複雜結構和長期相依性的序列數據。

本篇報告中，由於時間序列模型的預測結果不盡理想，因此後續我們採用較為進階、同樣也能針對時間序列相關數據做預測與模擬的LSTM模型來進行預測，發現效果較好。後續也會針對造成如此現象的原因去提出我們的看法。

四、實驗設計

本次報告中，我們所採用的數據來源為中華民國統計資訊網（https://statdb.dgbas.gov.tw/pxweb/Dialog/statfile9L.asp）之中所提供的出生人口數數據。在此網站上提供了週期單位為年的資料以及週期單位為月的資料。其中週期為年的單位資料從1993年開始，共30筆資料，而週期為月的資料從2000年1月開始，直到2023年5月，共281筆資料。起初我們利用年資料來進行預測，但後來發現資料筆數太少，可能會造成預測結果失準，故後來採用月資料來進行預測。

程式實作中，我們所執行的動作大致分為幾步：

準備資料：將每年的出生人數數據整理成適合AR模型的格式，在此利用Pandas讀取CSV檔

拆分資料：將數據集劃分為訓練集和測試集。前者用於估計模型參數，後者用於評估模型預測性能。

選擇階數：根據數據的特性和領域知識，選擇適當的AR階數（p）。

估計參數：對訓練集的數據進行擬合，估計AR模型的參數。

模型預測：使用估計的參數對未來值進行預測。

模型診斷：驗證模型的合理性、預測值與測試值做比較。

完整程式碼如下：

```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt

# 準備數據
data = pd.read_csv('ETHUSD.csv')
series = data['Open']

# 繪製ACF圖
sm.graphics.tsa.plot_acf(series, lags=20)  # 設定lags參數來指定要繪製的滯後階數
plt.xlabel('Lag')
plt.ylabel('ACF')
plt.title('Autocorrelation Function (ACF)')
plt.show()

# 繪製PACF圖
sm.graphics.tsa.plot_pacf(series, lags=14)  # 設定lags參數來指定要繪製的滯後階數
plt.xlabel('Lag')
plt.ylabel('PACF')
plt.title('Partial Autocorrelation Function (PACF)')
plt.show()
```

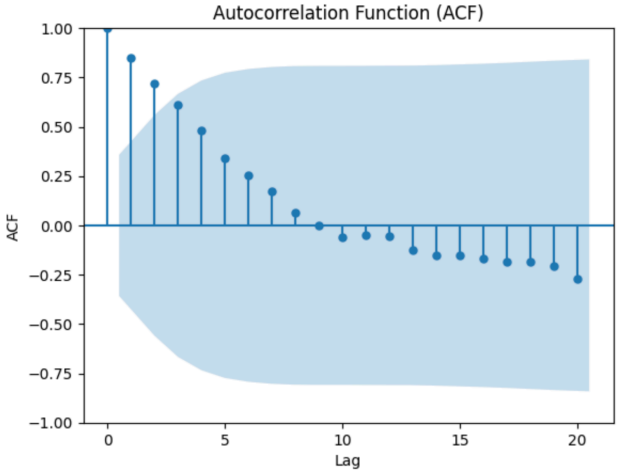
```
diff1 = data.diff(1)
diff1['Born'].plot()
data_new = diff1['Born'].iloc[1:]
result = adfuller(data_new)

    # 提取ADF檢定結果中的關鍵信息
adf_statistic = result[0]
p_value = result[1]
critical_values = result[4]

    # 進行恆定性判斷
if p_value < 0.05 and adf_statistic < critical_values['5%']:
    print("序列是恆定序列")

else:
    print("序列不是恆定序列")
```

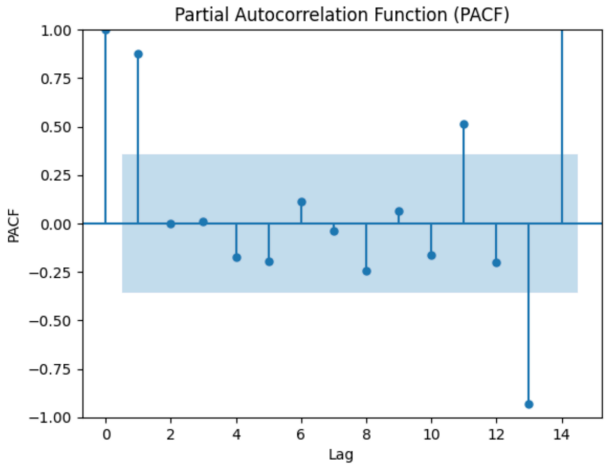
```python
import pandas as pd
import numpy as np
import matplotlib.pyp
from statsmodels.tsa.ar_model import AutoReg

data = pd.read_csv('DATA.csv')
data['Year'] = pd.to_datetime(data['Year'], format='%Y')
data.set_index('Year', inplace=True)

train_data = data.iloc[:-3]  # 使用除了最後三年的資料作為訓練集
train_data = train_data['Born']
test_data = data.iloc[-3:]  # 使用最後三年的資料作為測試集
model = AutoReg(train_data, lags=2)  # 設定lags參數來指定滯後階數
model_fit = model.fit()
predictions = model_fit.predict(start=len(train_data), end=len(train_data)+len(test_data)-1)
plt.plot(data.index, data['Born'], label='Actual')
plt.plot(test_data.index, predictions, label='Predicted')
plt.xlabel('Year')
plt.ylabel('Born')
plt.title('AR Model - Born Prediction')
plt.legend()
plt.show()
print(model_fit.aic,model_fit.bic,model_fit.hqic)
```

(module) ar_model

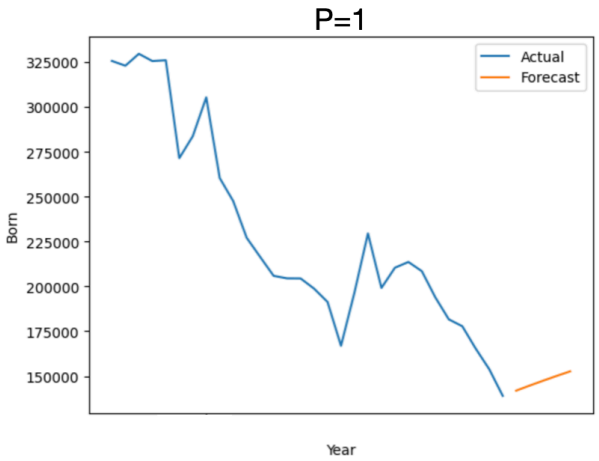
```python
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.tsa.arima.model import ARIMA

# 讀取人口數據集
data = pd.read_csv('DATA.csv')

# 將時間列設置為索引並轉換為時間序列
data['Year'] = pd.to_datetime(data['Year'])
data.set_index('Year', inplace=True)

# 繪製人口數時間序列
data['Born'].plot()
plt.xlabel('Year')
plt.ylabel('Born')
plt.show()

# 擬合ARMA模型
model = ARIMA(data['Born'], order=(2, 0, 0))
model_fit = model.fit()
# 預測未來人口數
forecast = model_fit.forecast(steps=5)

# 繪製原始數據和預測結果
plt.plot(data['Born'], label='Actual')
plt.plot(forecast, label='Forecast')
plt.xlabel('Year')
plt.ylabel('Born')
```

其中，ACF、PACF執行結果如下：

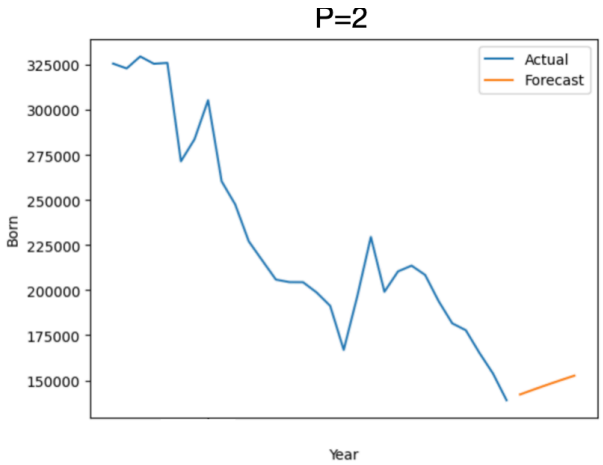




AR Model預測結果如下：

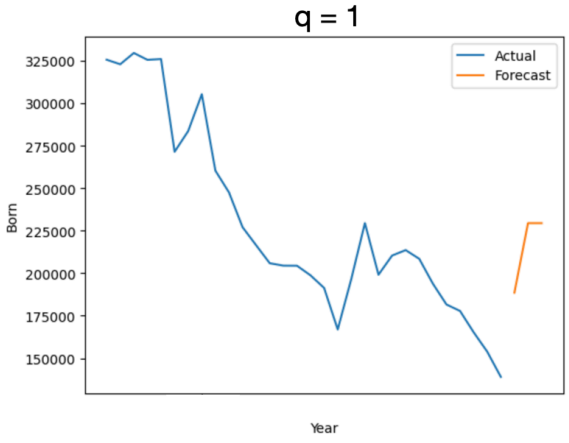


686.5661877085007 690.7697798534872 687.91095295271

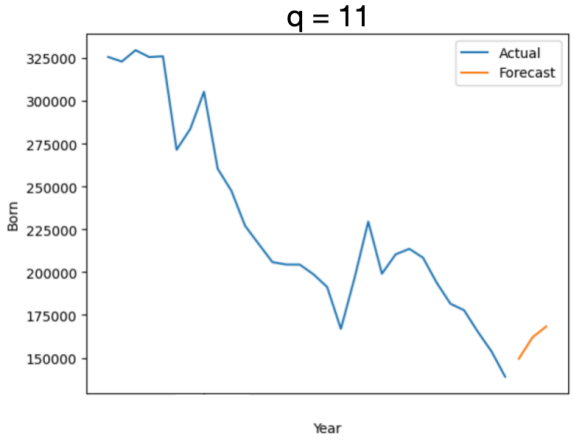


688.5399295698886 694.1447190965372 690.332949895501

MA Model 預測如下：

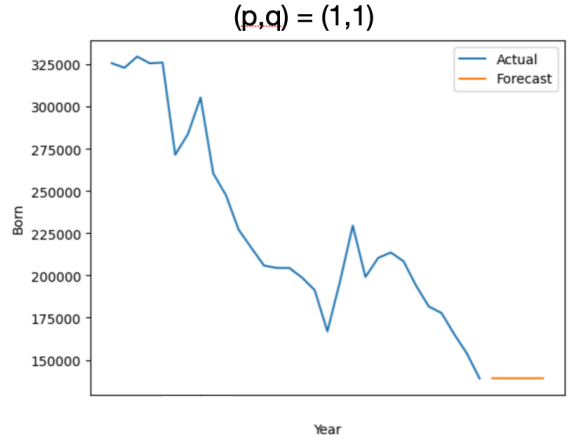


716.876410757263 721.0800029022495 718.2211760014723

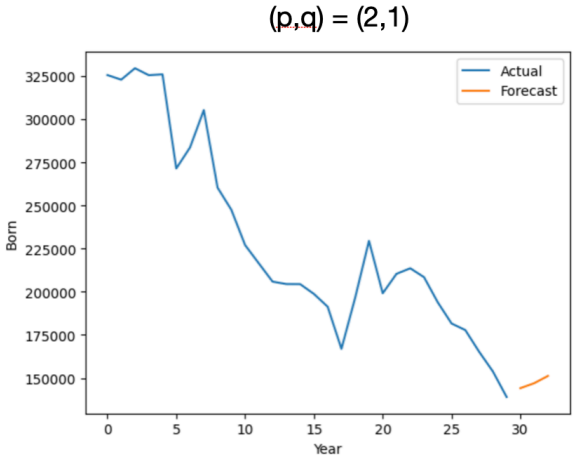


734.3598804601615 752.5754464217696 740.1871965184016

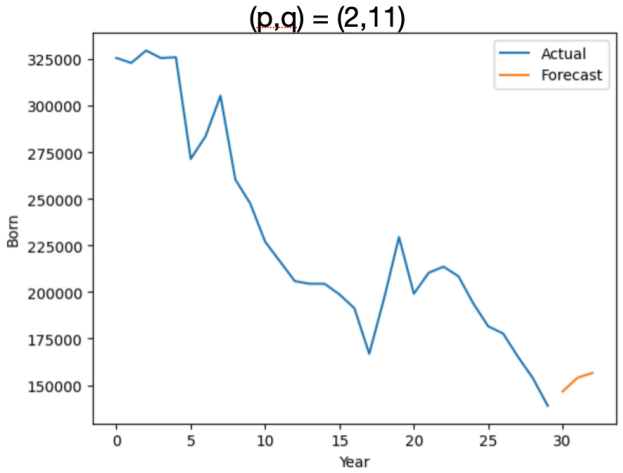
ARMA Model預測結果如下：



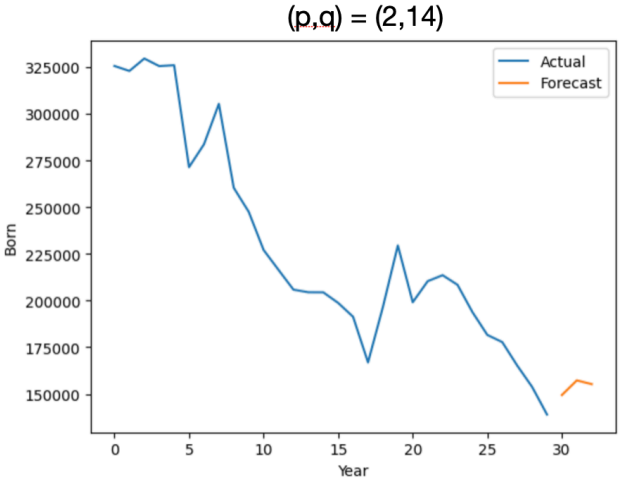
659.4318412274694 662.1664328874423 660.288281219122



688.0117214892912 695.017708397602 690.2529968963066

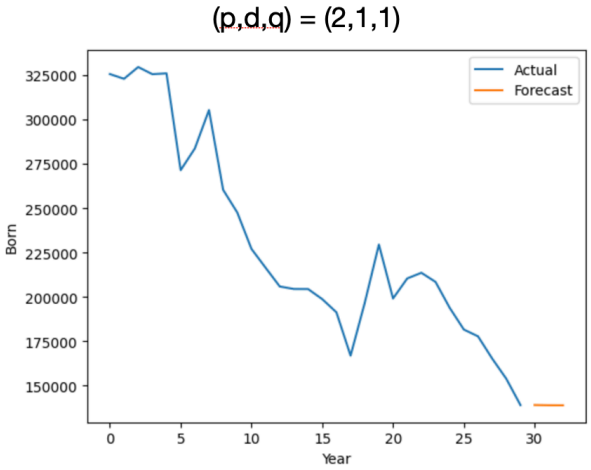


738.202430172306 759.2203908972383 744.9262563933522

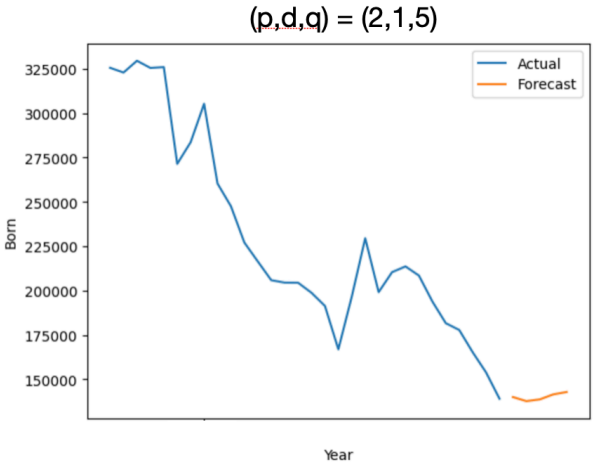


742.1941634425255 767.4157163124443 750.262754907781

ARIMA Model 預測結果如下：



663.4103492655028 668.8795325854486 665.1232292488082

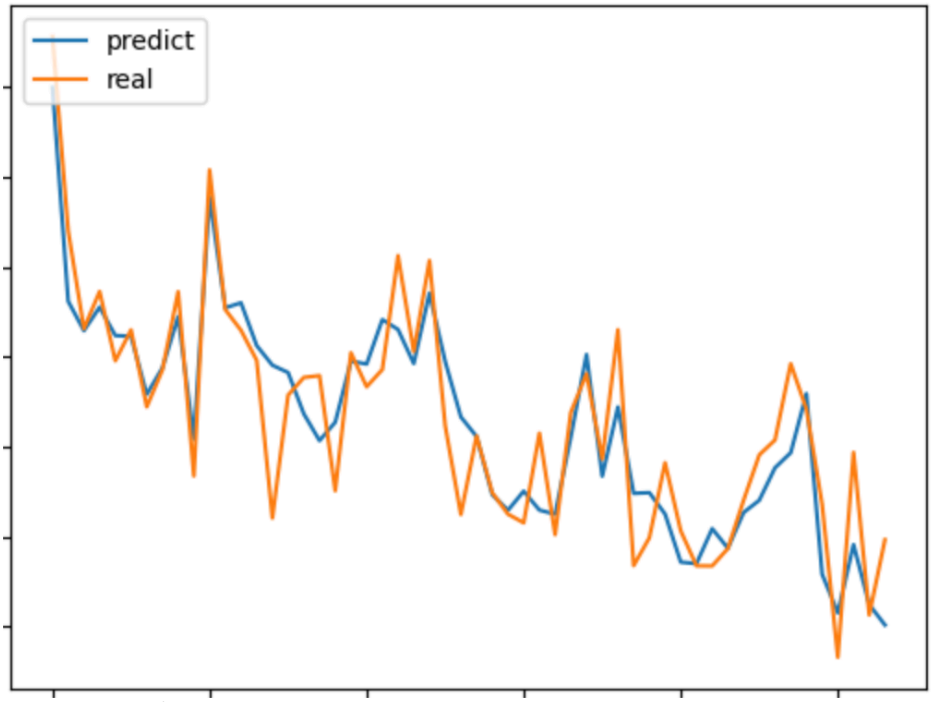


686.9391789985824 697.8775456384742 690.3649389651932

我們可以發現預測結果其實蠻不理想的，我們分析的原因可能有以下幾個：（１）影響出生人數的要素很多，並非單一因素影響，若要解決此方面問題，可以嘗試多變數模型（如：VAR、VMA、VARIMA...等等）並考慮結婚/離婚率、GDP...等其他數據（２）預測的初期

我們輸入的資料為年資料，其筆數僅30筆，算是有點少，因此可以嘗試換成每月出生人數資料，筆數較多。（３）AR、MA模型較適合處理恆定序列，而出生人數有明顯趨勢，並非恆定序列，若要解決此方面問題，可進行資料的拆分使其成為恆定序列。（４）使間序列模型可能稍微簡易，可嘗試利用更複雜、更高階的機器學習或深度學習來進行模擬與預測，可能會得到更好的結果。

最終，我們為了使預測結果更加準確，我們採用了LSTM模型來進行預測，最終執行預測結果可視化後如下所示：



可以發現預測效果好非常多，表明了當初因為對於時間序列模型認識的不夠貿然選用時間序列模型來預測人口數是一個非常不明智的抉擇。

五、結論

本篇報告旨在利用時間序列模型預測台灣的出生人口數，並在獲得非常不理想的結果後，改利用LSTM模型進行預測，發現LSTM模型在預測上表現非常準確。

最初，我們採用傳統的時間序列模型進行預測，但經過分析和評估後，我們發現這些模型無法捕捉到複雜的時間序列模式和長期相依性。預測結果不理想，無法準確預測台灣的出生人口數。然而，我們後來轉而採用LSTM模型，預測結果發生了顯著的改善。LSTM模型具有捕捉長期依賴性和隱藏模式的能力，能夠有效地學習和預測時間序列數據中的模式。通過對過去數據的學習，LSTM模型能夠提取關鍵的特徵並進行準確的預測。

在本篇報告中，我們使用LSTM模型對台灣的出生人口數進行了預測，並與實際數據進行比較。結果顯示，LSTM模型的預測結果相較之下準確不少，與實際數據非常接近。這證實了LSTM在時間序列預測中的優越性，特別是在處理具有複雜結構和長期相依性的數據時。

本篇報告展示了LSTM模型在時間序列預測領域的應用潛力，為未來相關研究提供了啟示。然而，我們也要認識到LSTM模型的使用可能需要更多的數據和計算資源，以及對參數調整和模型設計的進一步研究。儘管如此，本篇報告的結果為利用LSTM模型進行人口預測提供了實證支持，並在相關領域的實踐中具有重要的參考價值。

參考資料：

[1] 中華民國統計資訊網 https://statdb.dgbas.gov.tw/pxweb/Dialog/statfile9L.asp

[2] 遞歸神經網路（RNN）和長短期記憶模型（LSTM）的運作原理 https://brohrer.mcknote.com/zh-Hant/how_machine_learning_works/how_rnns_lstm_work.html

[3] LSTM 如何做多步預測 https://ithelp.ithome.com.tw/questions/10199599

[4] https://zh.wikipedia.org/zh-tw/ARIMA%E6%A8%A1%E5%9E%8B

[5] 凌晨狂上上下下23次-時間序列分析終局之戰，ARIMA差分整合移動平均自迴歸模型 https://ithelp.ithome.com.tw/articles/10252815

[6] 時間序列探索(二)：ARIMA家族簡介 https://medium.com/@cindy050244_52136/%E6%99%82%E9%96%93%E5%BA%8F%E5%88%97%E6%8E%A2%E7%B4%A2-%E4%BA%8C-arima%E5%AE%B6%E6%97%8F%E7%B0%A1%E4%BB%8B-8d533f0b18d6

[7] https://docs.aws.amazon.com/zh_tw/forecast/latest/dg/aws-forecast-recipe-arima.html