第一部分

使用 Random Forest 和 XGBoost 預測 2020/1 至 2022/12 的月營收金額（共 36 個月）

Q1、分析 2020/1 至 2022/12 月營收金額的預測結果（e.g. 討論各模型的預測分數差異和變數重要性、最佳和最差模型分別為何）並標明判斷預測結果好壞的衡量指標。

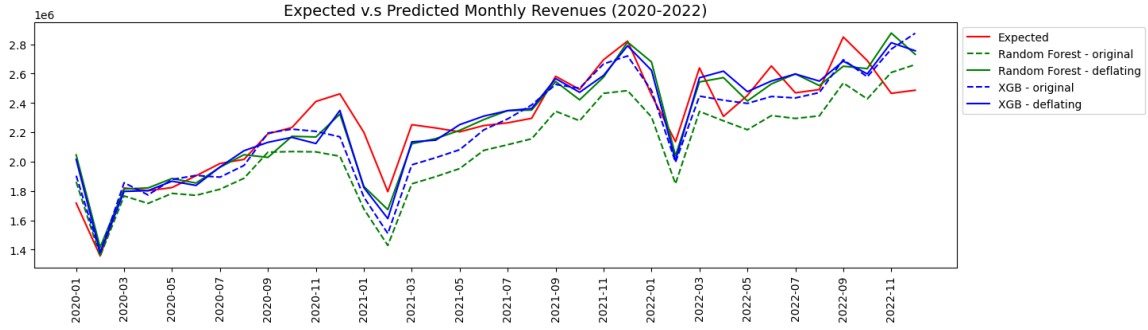
1. 各模型預測能力比較:
    a. XGBoost 表現較 Random Forest 好:
        i. 由【圖 1】預測金額圖形來看，在有進行資料平減以及沒有進行資料平減的兩情況下，XGBoost 模型預測出來的金額都離 Expected 較接近。
        ii. 由【圖 2】預測分數 RMSE 圖形來看，無論有沒有進行資料平減，藍線大部分皆落在綠線以下，也就是 XGB 相比於 Random Forest 有較低的均方根誤差，代表預測值和實際值之間的距離小、模型預測能力好。
        iii. 由【圖 3】預測分數 MAE 圖形來看，無論有沒有進行資料平減，藍線大部分皆落在綠線以下，也就是 XGB 相比於 Random Forest 有較低的平均絕對誤差，代表預測值和實際值之差地絕對值小、模型預測能力好。
    b. 資料經過平減後，模型預測能力變好:
        i. 由【圖 4】預測金額圖形來看，無論是 Random Forest Model 還是 XGBoost Model，經過資料平減的模型(實線)相較於使用原始資料的模型(虛線)，離 Expected 較接近，因此可知資料經過平減後，模型的預測能力提高。
        ii. 由【圖 5】預測分數 RMSE 圖形來看，無論使用 Random Forest Model 還是 XGBoost Model，實線大部分落在虛線以下，也就是平減後可以有較低的均方根誤差，代表預測值和實際值之間的距離小、模型預測能力好。
        iii. 由【圖 6】預測分數 MAE 圖形來看，無論使用 Random Forest Model 還是 XGBoost Model，實線大部分落在虛線以下，也就

是平減後有較低的平均絕對誤差，代表預測值和實際值之差地絕對值小、模型預測能力好。

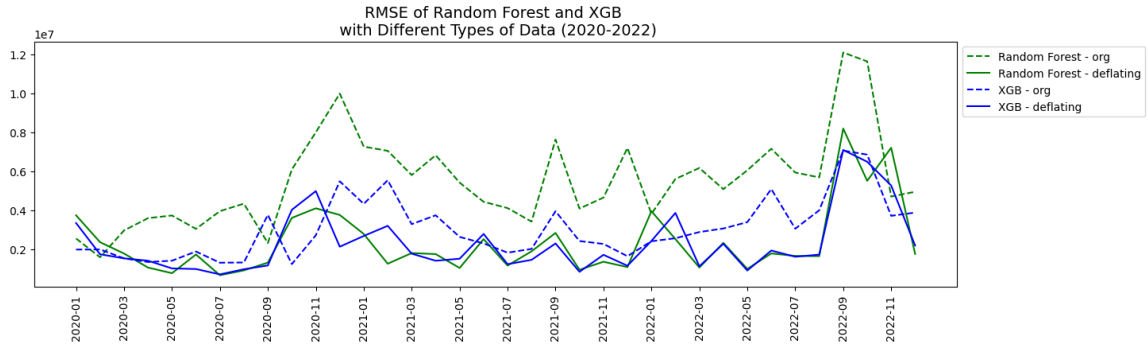
推測原因如下:

- 降低離群值的影響: 進行標準化可以減少離群值或異常值對模型的影響，避免對模型訓練產生干擾，進而提高模型預測能力。
- 減少特徵之間的差異性: 進行標準化可以把不同數值範圍和單位統一到相同標準，減少特徵間的差異性，避免某些特徵權重過大或過小，影響模型學習。

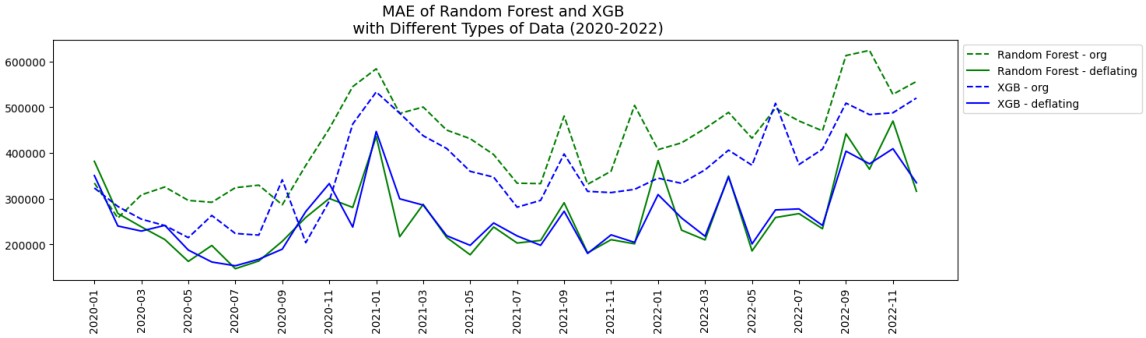
【圖 1、各模型預測營收金額】



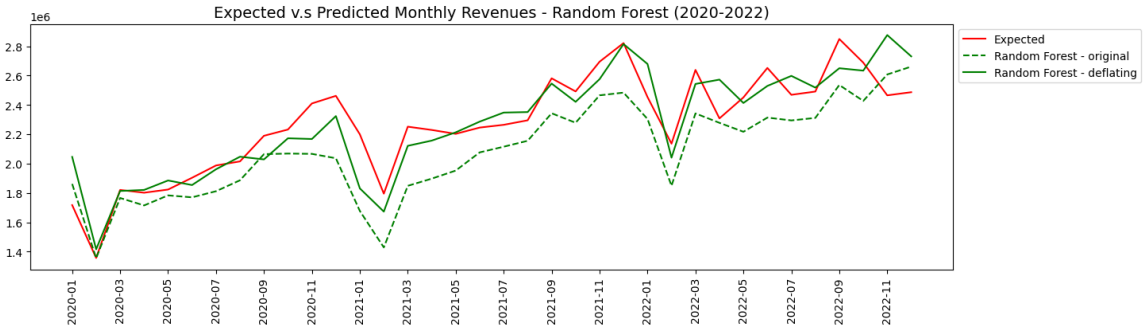
【圖 2、各模型 RMSE】

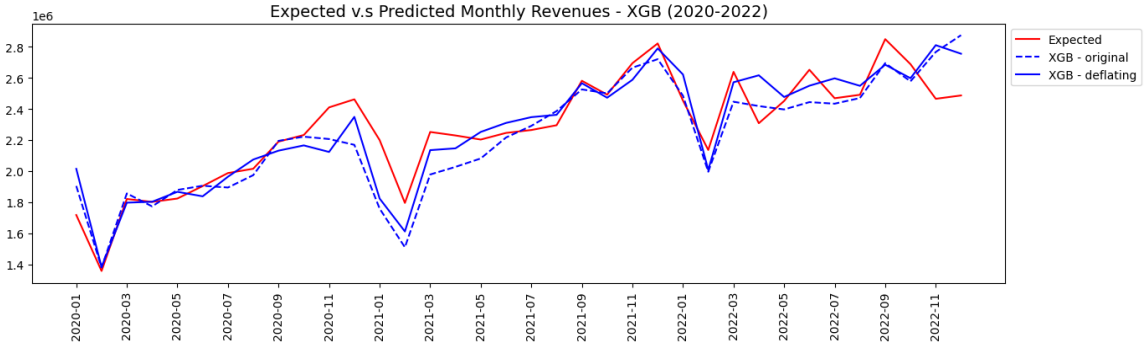


【圖 3、各模型 MAE】

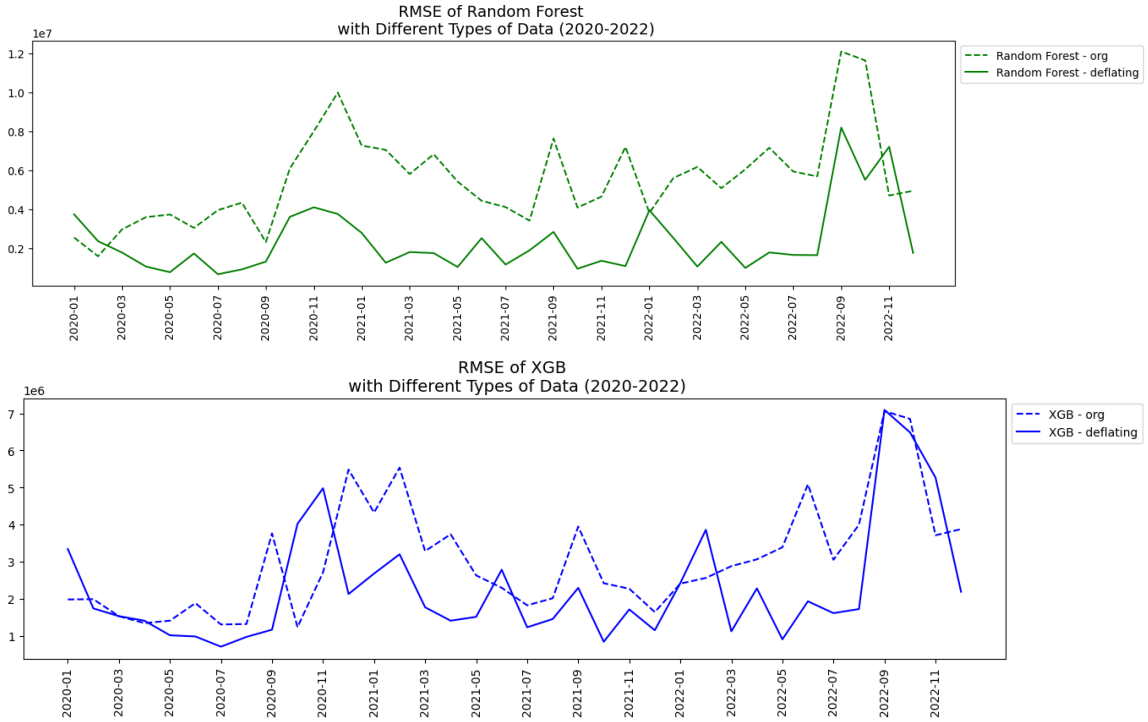


【圖 4、各模型預測營收金額】

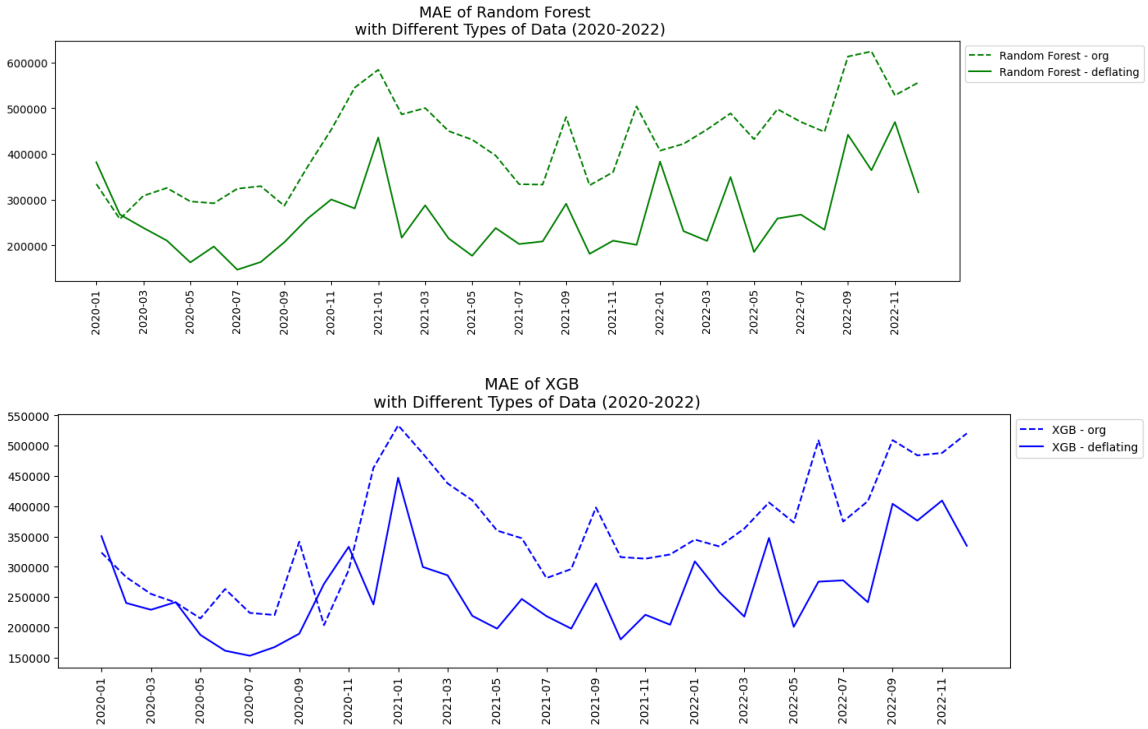




【圖 5、各模型 RMSE】



【圖 6、各模型 MAE】



2. 各模型變數重要性:

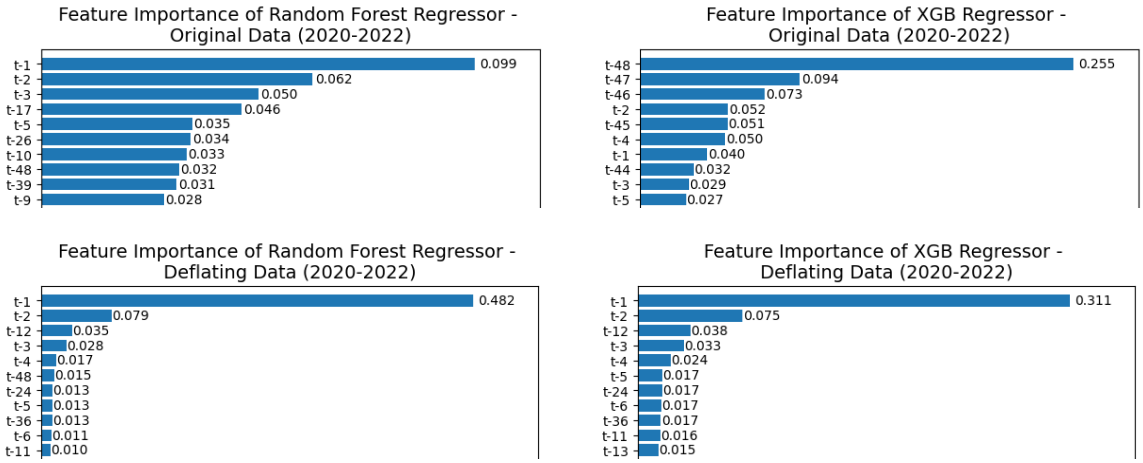
   由【圖 7】中可以看出以下為較重要的變數，並推測原因如下:

   a. t - 1、t - 2、t - 3: 前 1~3 個月反應了最近的趨勢，通常影響當月營收較大，因此模型預測時變數重要性高。

   b. t - 12: 前一年的數據反映季節性變化，因此對預測也重要。

   c. t - 48: 前兩年的數據反應較長期的趨勢變化，因此也可能影響模型預測。

【圖 7、各模型變數重要性】



3. 最佳 & 最差模型

   *以 **RMSE** 作為模型好壞衡量標準，最佳及最差模型特性分析如下:*

   a. 如【圖 8】，Random Forest 及 XGBoost 兩者最佳模型皆在 2020-07

      i. 如【圖 10】，t-1 及 t-2 為最重要的變數: 因為月營收屬於時間序列資料，相鄰的月份之間通常存在某種關係，尤其受到前一至兩個月的趨勢影響，所以前一個月以及前兩個月的營收資料通常對月營收預測最為重要。

   b. 如【圖 9】，Random Forest 及 XGBoost 兩者最差模型皆在 2022-09

      i. 如【圖 10】，最差模型的 t-12 相較於最佳模型而言較不重要，而 t-3、t-4 則較為重要: 這意味著一年前的數據相比於三個月或四個月前的數據更為重要，可能因為營收具有季節性，也就是在每年的某個月營收可能都會較高或較低，因此在進行預測時，去年同期營收對於模型準確度貢獻較大。

【圖 8、以 RMSE 衡量的最佳模型】

	modelName	dataType	scoreType	min_month	min_score
2	Random Forest	def	RMSE	2020-07	659573.0

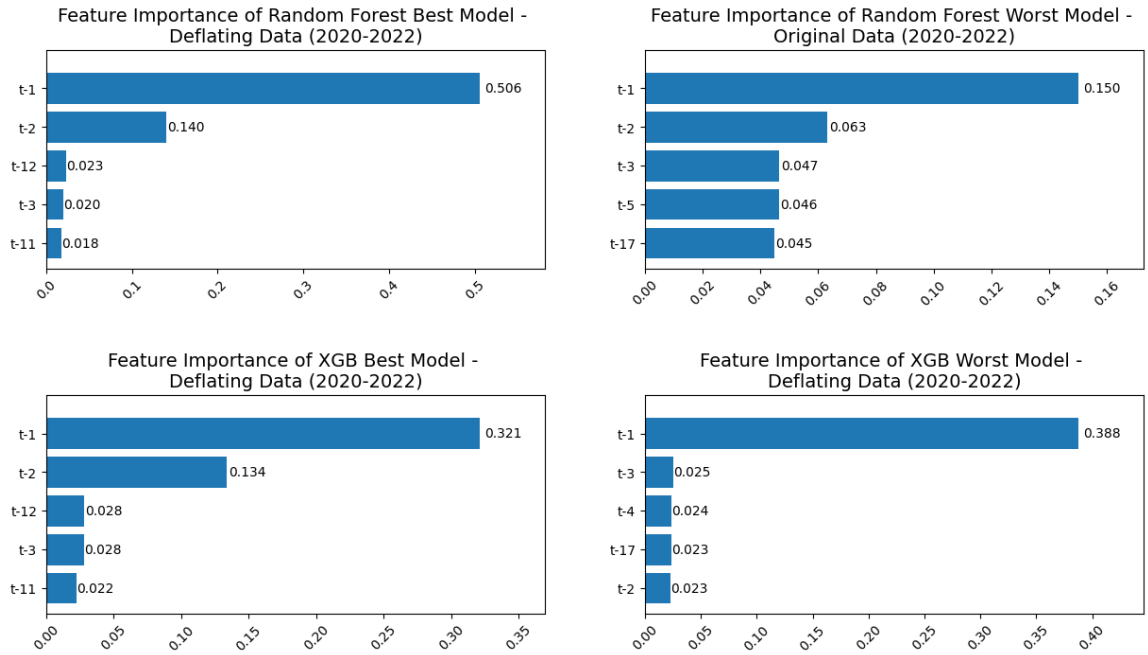
	modelName	dataType	scoreType	min_month	min_score
2	XGB	def	RMSE	2020-07	703651.0

【圖 9、以 RMSE 衡量的最差模型】

	modelName	dataType	scoreType	max_month	max_score
0	Random Forest	org	RMSE	2022-09	12096525.0

	modelName	dataType	scoreType	max_month	max_score
2	XGB	def	RMSE	2022-09	7093180.0

【圖 10、以 RMSE 衡量的最佳&最差模型變數重要性比較】



以 MAE 作為模型好壞衡量標準，最佳及最差模型特性分析如下:

c. 如【圖 11】，Random Forest 及 XGBoost 兩者最佳模型皆在 2020-07
   i. 如【圖 13】，t-1 及 t-2 為最重要的變數: 因為月營收屬於時間序列資料，相鄰的月份之間通常存在某種關係，尤其受到前一至兩個月的趨勢影響，所以前一個月以及前兩個月的營收資料通常對月營收預測最為重要。

d. 如【圖 12】，Random Forest 及 XGBoost 最差模型分別在 2022-10、2021-01
   i. 如【圖 13】，最差模型的 t-12 相較於最佳模型而言較不重要，而 t-44、t-48 則較為重要: 這意味著一年前的數據相比於兩年前的數據更為重要，可能因為營收具有季節性，也就是在每年的某個月營收可能都會較高或較低，而較長期的趨勢則影響較小，因此在進行預測時，去年同期營收對於模型準確度貢獻較大。

【圖 11、以 MAE 衡量的最佳模型】

	modelName	dataType	scoreType	min_month	min_score
3	Random Forest	def	MAE	2020-07	146709.0

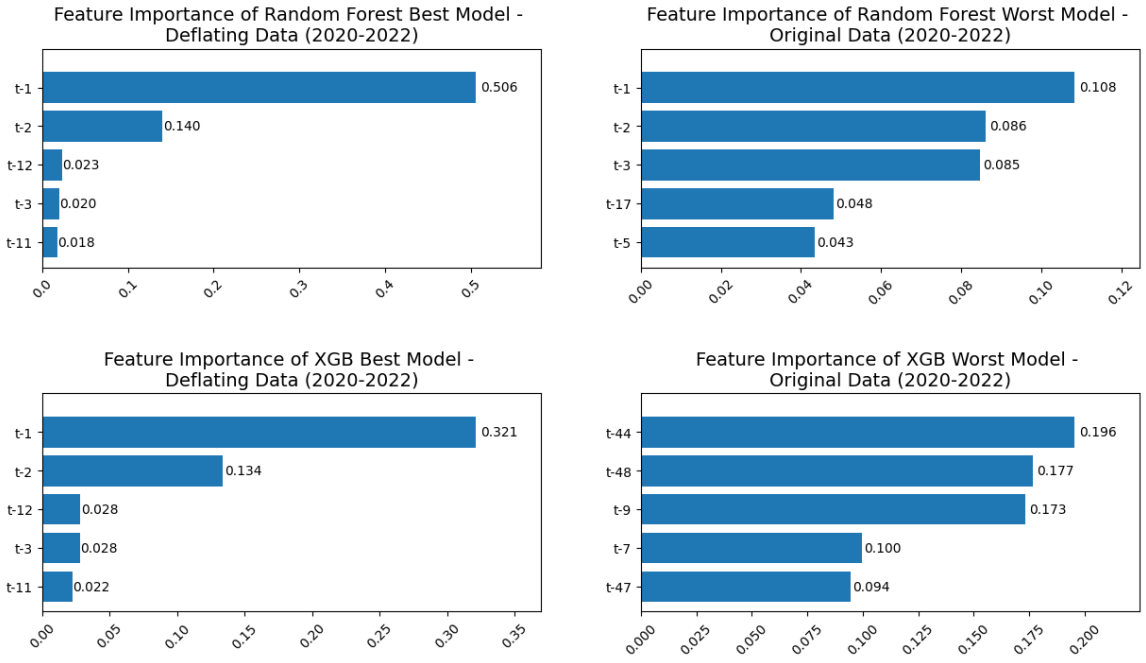
	modelName	dataType	scoreType	min_month	min_score
3	XGB	def	MAE	2020-07	153066.0

【圖 12、以 MAE 衡量的最差模型】

	modelName	dataType	scoreType	max_month	max_score
1	Random Forest	org	MAE	2022-10	624279.0

	modelName	dataType	scoreType	max_month	max_score
1	XGB	org	MAE	2021-01	533372.0

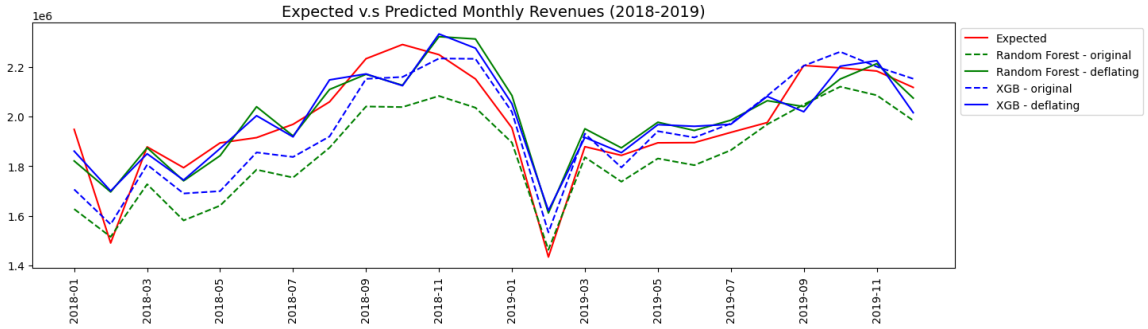
【圖 13、以 MAE 衡量的最佳&最差模型重要性比較】



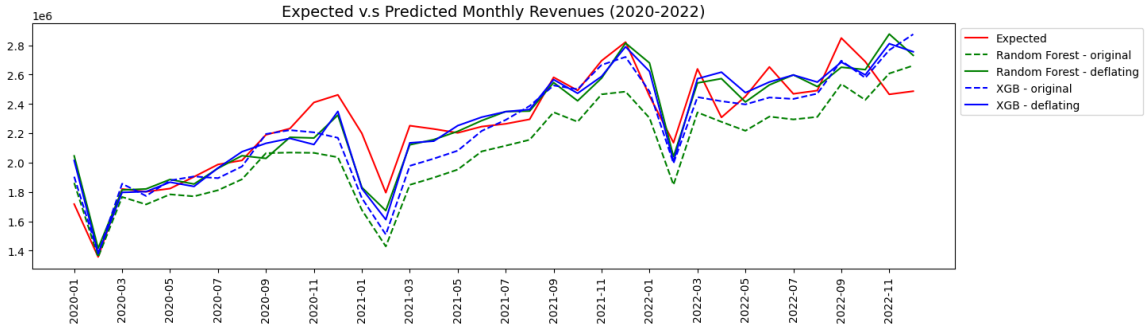
Q2、和 201801 - 201912 的預測結果做比較

1. 相同:
    a. XGB 相較 Random Forest 準確度較高: 由【圖 14】~【圖 19】預測金額圖、RMSE、MAE 圖片比較可知，無論資料是否經過平減，XGBoost 模型表現大致比 Random Forest 還好，因為 XGBoost 模型進行了正則化避免 overfitting，提度提升技術也能幫助模型更加優化。
    b. 進行資料平減後模型預測能力較高: 由【圖 14】~【圖 19】預測金額圖、RMSE、MAE 圖片比較可知，無論使用 XGBoost 或是 Random Forest 模型進行預測，進行平減後的資料大只表現較好，因為資料經過平減後，可以減少異常值對模型的影響、提高模型學習效率和穩定性，以及準確性。

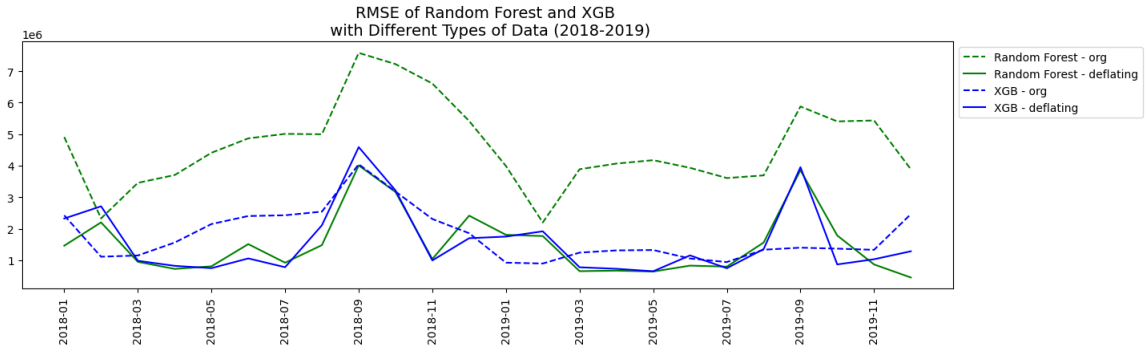
【圖 14、201801-201912 預測金額】



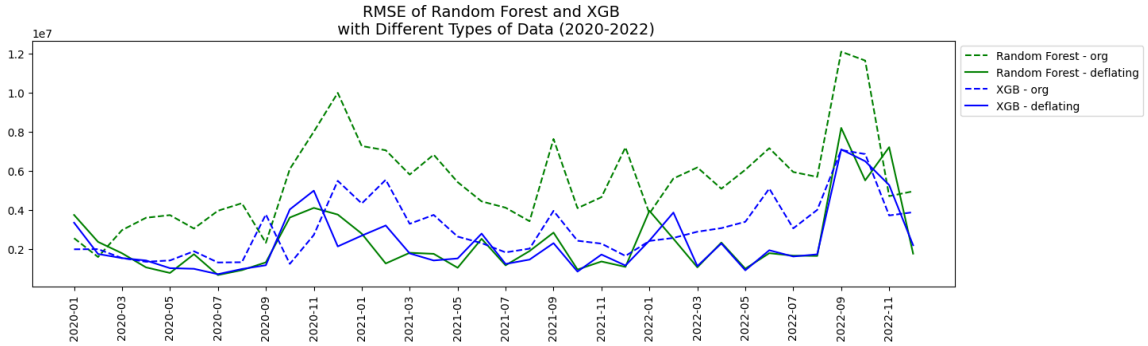
【圖 15、202001-202212 預測金額】



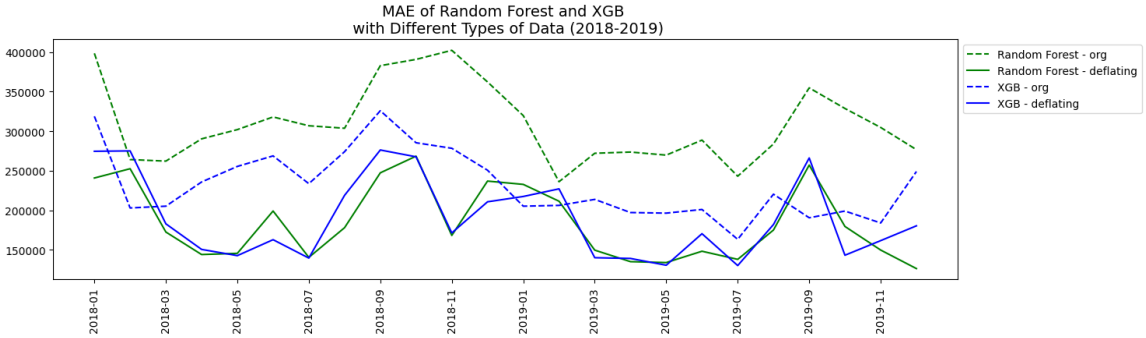
【圖 16、201801-201912 RMSE】



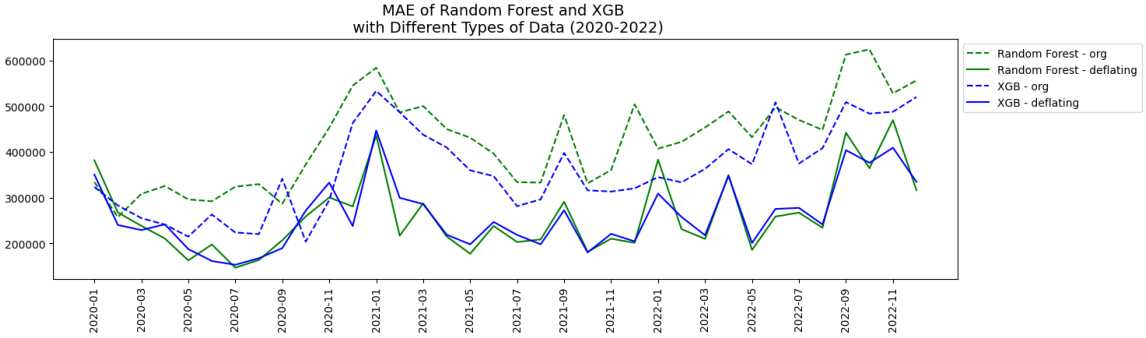
【圖 17、202001-202212 RMSE】



【圖 18、201801-201912 MAE】

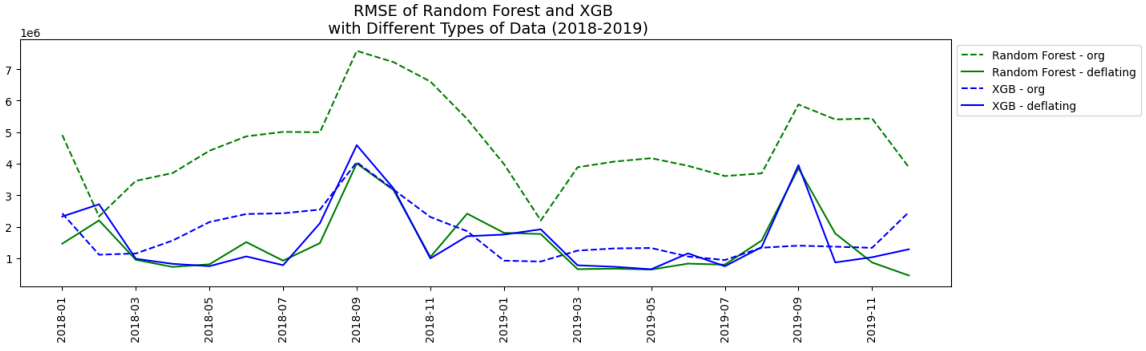


【圖 19、202001-202212 MAE】

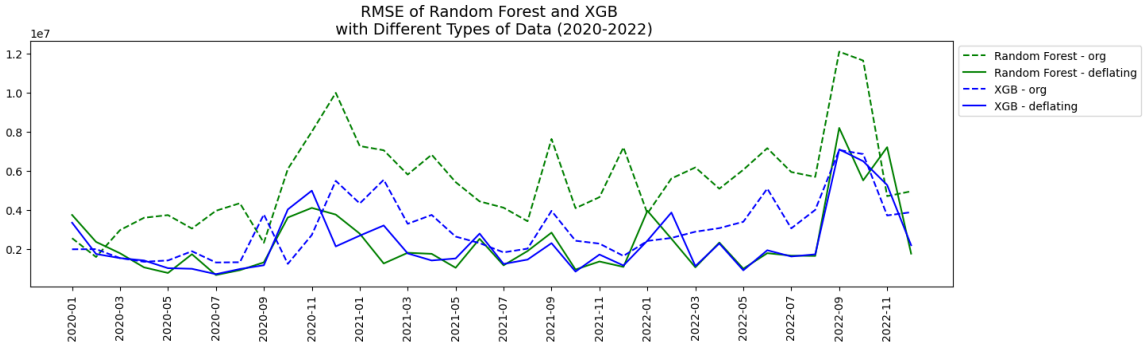


2. 相異:
    a. 近期預測相比於以前預測能力變差: 由【圖 20】~【圖 23】RMSE 以及 MAE 圖片比較可發現 202001-202212 的預測表現相比於 201801-201912 來得差，推測是因為近幾年受到疫情等因素影響，導致使得經濟環境或市場變化較大，模型無法完全捕捉這些較大的變化，因此預測能力比較低。

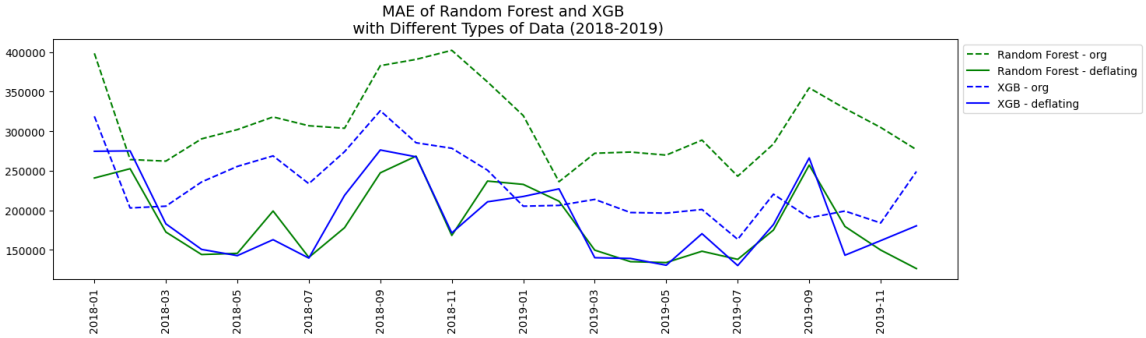
【圖 20、201801-201912 RMSE】



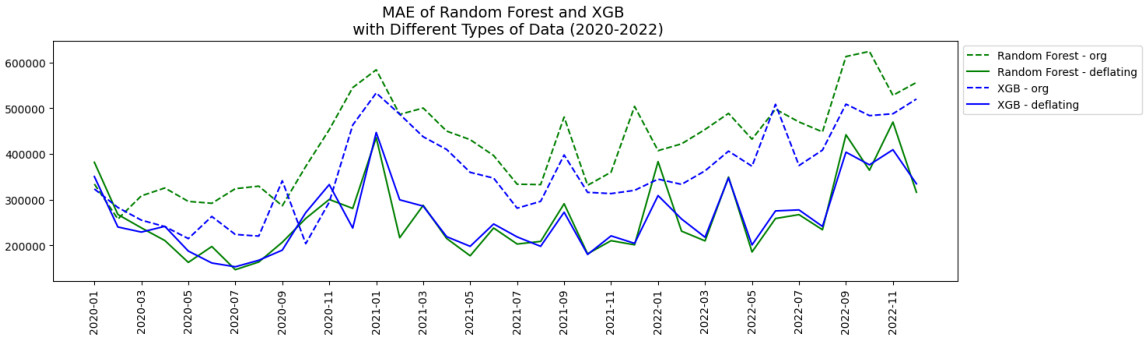
【圖 21、202001-202212 RMSE】



【圖 22、201801-201912 MAE】



【圖 23、202001-202212 MAE】



第二部分

鎖定產業進行預測，並分析預測結果

Q1: 定義所挑選的產業、說明資料集前處理理的方式

1. 產業選擇:

選擇「半導體」(TSE 產業別 = 24)。

2. 選擇依據:

選擇資料筆數足夠大的產業，確保模型準確率及可比較性。

Q2: 分析 2020/1 至 2022/12 月營收金額的預測結果（須標明使用之模型和衡量指標

1. 資料集處理:

複製已經修改時間資料格式資料處理的 dataframe，命名為 org_data_semi，將所有非「半導體」分類的公司資料全部刪除後， 共有 129 間公司。

2. 預測模型及衡量指標:

使用 Random forest 及 XGBoost 模型，並以 RMSE、MAE 作為模型好壞的衡量指標。

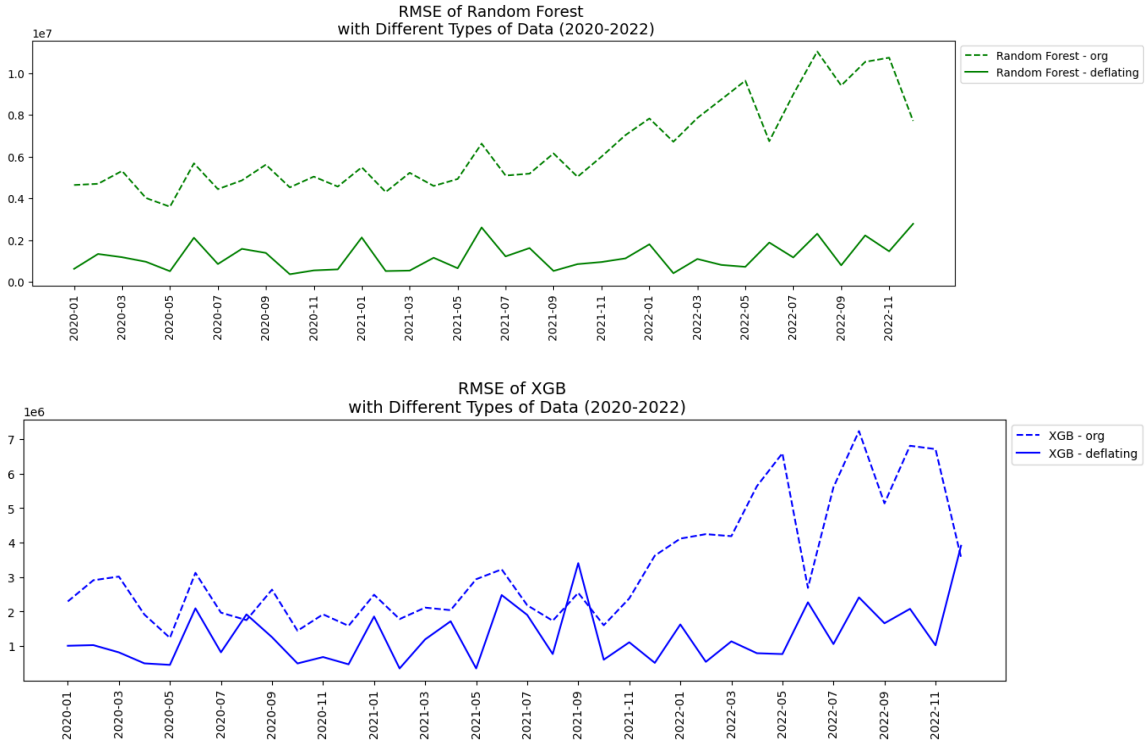
3. 資料平減:

使用標準化方式進行平減，對 X 做標準化，並以 X 的平均數及標準差對 y 做標準化，期望可以提高模型準確性和穩定性。

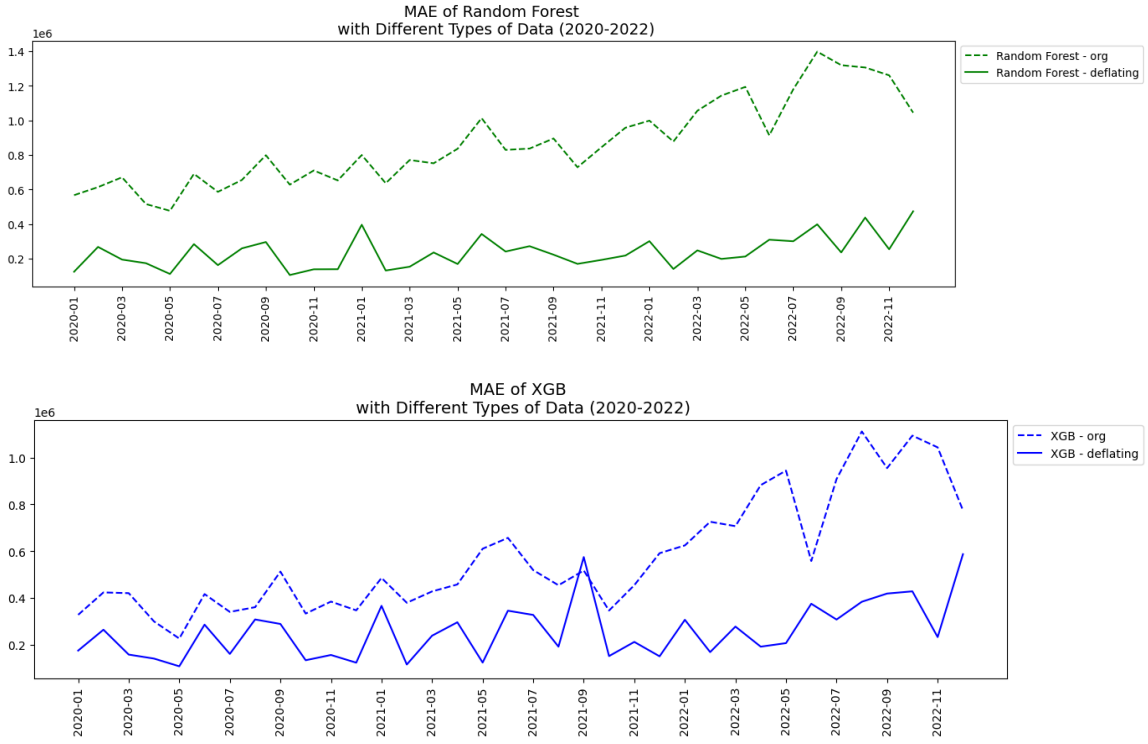
4.模型準確率預測結果:

4.1 由【圖 24】、【圖 25】可得資料平減後表現較佳，且平減與否的表現差異比全產業營收預測來得大:
從 RMSE、MAE 圖形中可以發現經過資料平減後的模型同樣表現大致較好。

【圖 24、半導體產業 RMSE】



【圖 25、半導體產業 MAE】

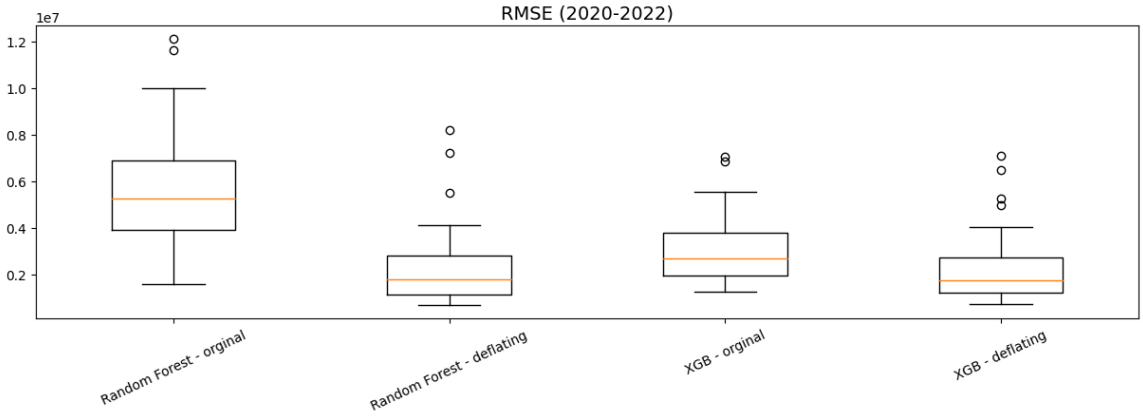


4.2 經資料平減的模型中，單一產業分析模型表現效能較好:

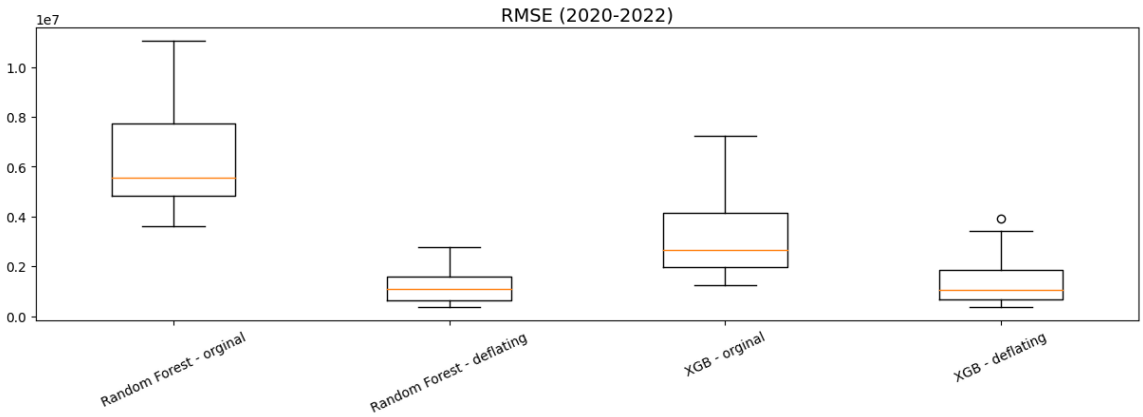
由【圖 26】~【圖 29】RMSE 及 MAE 圖形可看出，相比於全產業的營收預測，把產業限縮到半導體單一產業，平減模型的預測表現較好，推測可能有以下幾個原因:

a. 全產業預測較為複雜: 相比於單一產業，全產業模型同時捕捉各產業的數據波動，需考慮多個因素與營收之間以及多個產業間的複雜關係，使模型的複雜度較高、預測較不確定。
b. 數據多樣性: 相比於單一產業模型只需要處理特定產業數據，全產業的數據來自不同領域，具有較大的差異性，因此模型準確解釋數據並進行預測的難度較高。

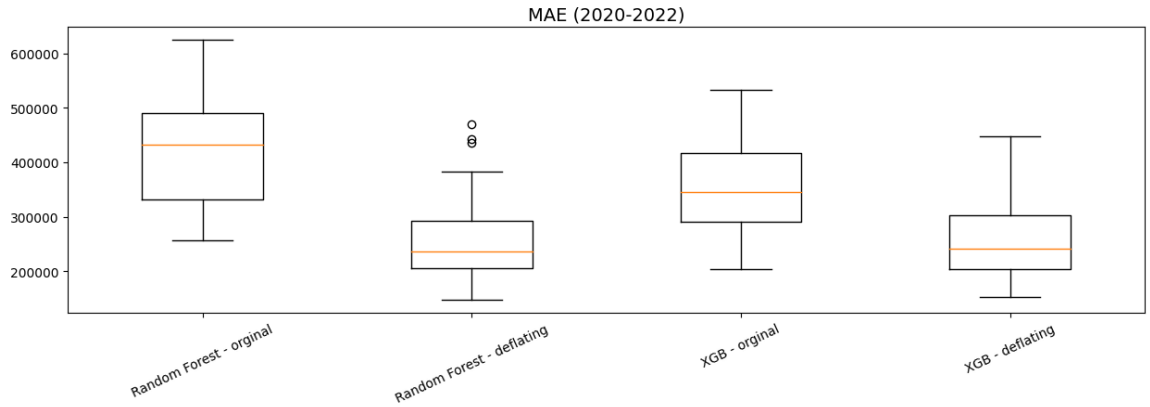
【圖 26、全產業各模型 RMSE】



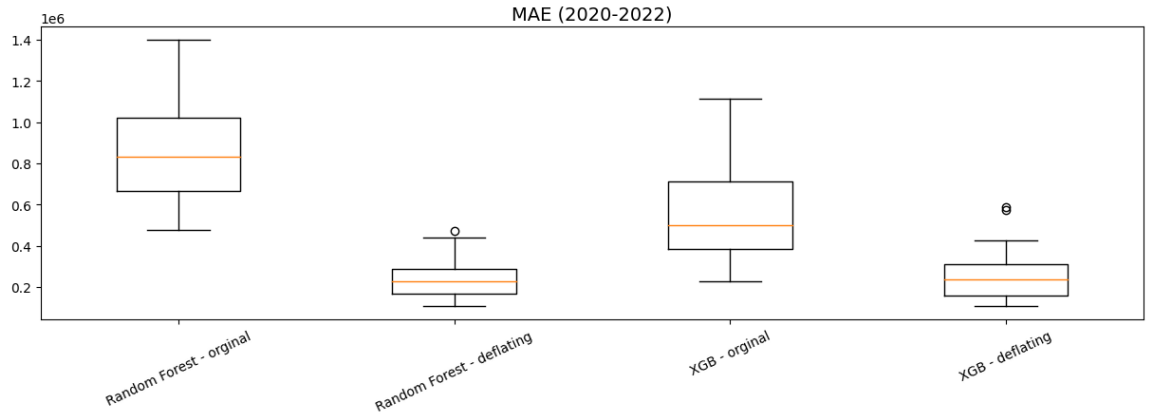
【圖 27、半導體產業各模型 RMSE】



【圖 28、全產業各模型 MAE】



【圖 29、半導體產業各模型 MAE】



最佳&最差模型特性及重要變數:

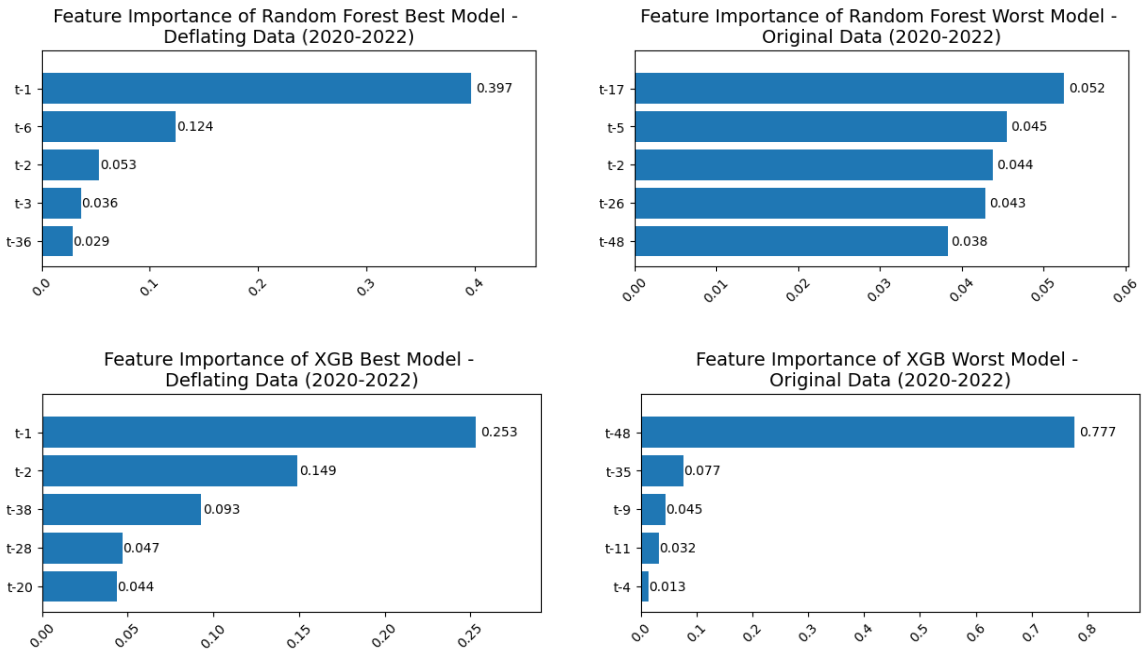
1.RMSE 作為衡量標準的最佳模型中， t-1、t-2 為重要變數:

由【圖 30】可以看出進行半導體產業營收預測時，最佳模型的前三重要變數包含 t-1、t-2 期資料，相比於全產業營收預測，近期歷史資料重要性更高，推測可能為以下原因:

- 時間序列相關性與季節變化: 半導體業的營收可能受到市場需求、供應鏈狀況、技術發展的因素影響，而這些因素通常有時間相依性，所以過去 1、2 個月的營收數據對未來的預測有較高參考價值。

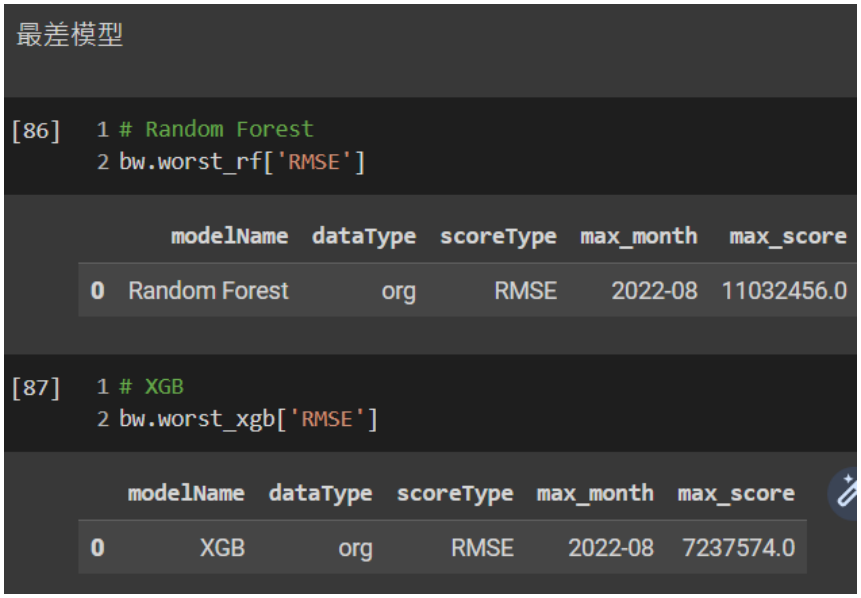
- 滯後效應: 半導體產業的生產和銷售周期相比其他產業稍微較長，所以有時需要數月時間才能完全反應營收數據，因此過去幾個月的營收數據對於當期的預測通常較為重要。

【圖 30、RMSE 標準下最佳&最差模型重要變數】



3. 由【圖 31】、【圖 32】可知，使用 RMSE、MAE 作為衡量標準下，Random Forest 及 XGBoost 的最差模型皆為 2022-08，上網搜尋半導體產業相關事蹟以及自行猜測，推測可能為以下原因導致模型較不容易進行月營收預測:
    - 美國總統拜登在 2022 年 8 月初聯合美光科技（Micron Technology）、英特爾（Intel）、洛克希德馬丁（Lockheed Martin）、超微半導體（AMD）等高層，簽署《晶片和科學法案》，強化美國半導體產業競爭力。
    - 半導體產業可能於 2022 年 8 月左右有重大技術革新，或是特殊行銷活動，導致該期模型準確率較低。

【圖 31、RMSE 標準下 Random Forest & XGBoost 最差模型】



最差模型

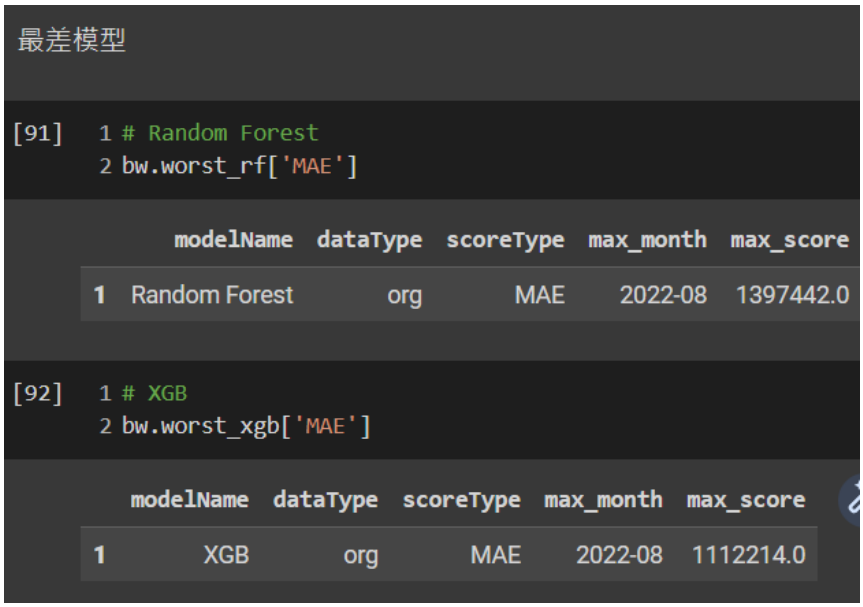
```
[86]  1 # Random Forest
      2 bw.worst_rf['RMSE']
```

	modelName	dataType	scoreType	max_month	max_score
0	Random Forest	org	RMSE	2022-08	11032456.0

```
[87]  1 # XGB
      2 bw.worst_xgb['RMSE']
```

	modelName	dataType	scoreType	max_month	max_score
0	XGB	org	RMSE	2022-08	7237574.0

【圖 32、MAE 標準下 Random Forest & XGBoost 最差模型】



最差模型

```
[91]  1 # Random Forest
      2 bw.worst_rf['MAE']
```

	modelName	dataType	scoreType	max_month	max_score
1	Random Forest	org	MAE	2022-08	1397442.0

```
[92]  1 # XGB
      2 bw.worst_xgb['MAE']
```

	modelName	dataType	scoreType	max_month	max_score
1	XGB	org	MAE	2022-08	1112214.0