機器學習與資料分析在會計領域的應用 HW2　　B08303006 財金四 林姝延

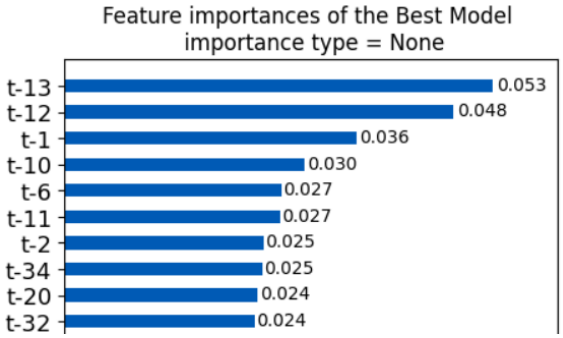
第一部分

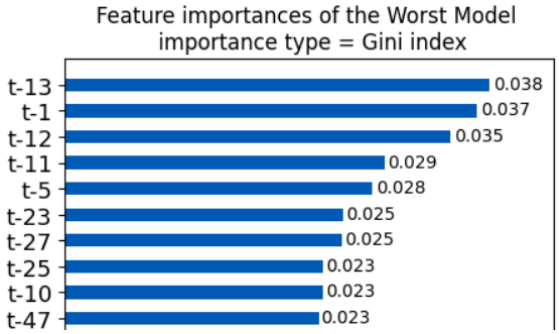
Q1:分析 2020/1 至 2022/12 月營收增減的預測結果（討論最佳模型的變數重要性、分類結果是否有異於其他模型）

1. **最佳模型變數重要性:**
    a. t-1 並非最重要的變數: 如下【圖 1】，前一個月營收數字確實會對 t 期營收有一定程度影響，因 t-1 若營收高，表示公司在該時點於市場中表現好，但前一期的營收很可能受到一些隨機的因素影響，例如放假、促銷等行銷活動這些特殊短期活動通常只會稍微影響下一個月的營收，但不一定會是最重要的影響參數。
    b. 前一年或幾年前同期的營收資料相對重要: 如下【圖 1】，t-13、t-12、t-10、t-11、t-34 等期間變數重要性較高，原因可能是營收存在季節性的波動，而前一年或前幾年同期的營收資料能夠消除季節性波動對預測的影響並提供更全面的參考。

【圖 1、最佳模型前十重要變數】　　【圖 2、最差模型前十重要變數】

XGBoost_Deflating_2021-1　　Random Forest_Original_2020-2



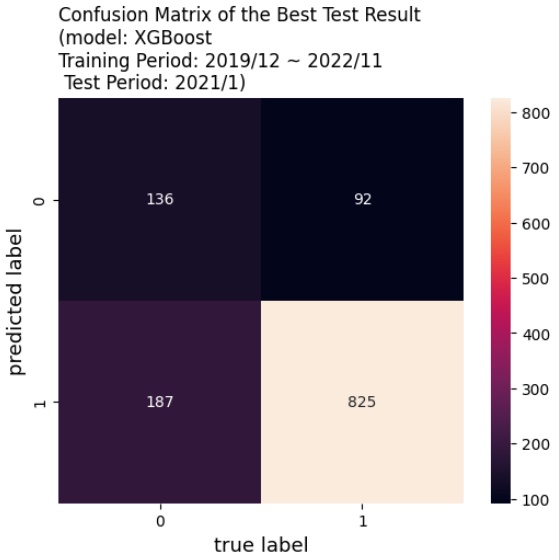


2. **模型分類結果比較:**
    a. 如下【圖 3】，最佳模型的誤判結果中，高估與低估的比數差異不大，因此模型較不會偏頗判斷，從兩類別分別觀察準確率，真實為達預期月營收的資料準確率為 825/(92+825)，大約為 90%，真實為未達預期月營收的資料準確率為 136/(136+187)，大約為 42%，得知個別來看，模型學習「達預期月營收資料」的特徵還是較多，模型在「達預期營收資料」方面學習得比較充分。

b. 如下【圖 4】，最差模型的模型預測結果當中，有高達 813 筆，即 813/921 大約 88%的比例是誤將實為高於預期的月營收，低估為不如預期的月營收，而高估的比率相對較低，因此模型相對容易過於悲觀，缺乏挖掘較佳月營收資訊的能力。

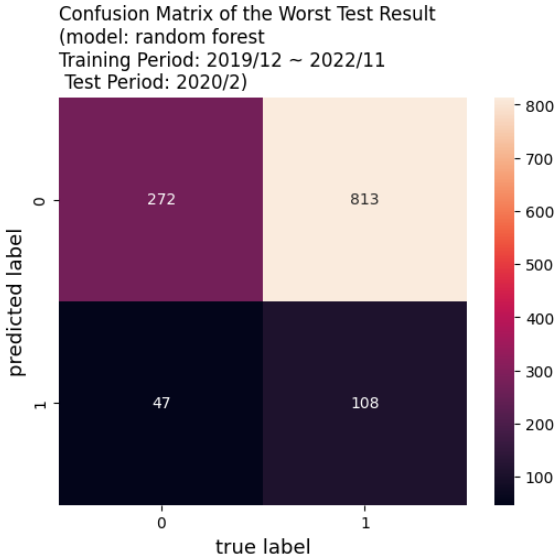
【圖 3、最佳模型前十重要變數】

XGBoost_Deflating_2021-1



【圖 4、最差模型前十重要變數】

Random Forest_Original_2020-2



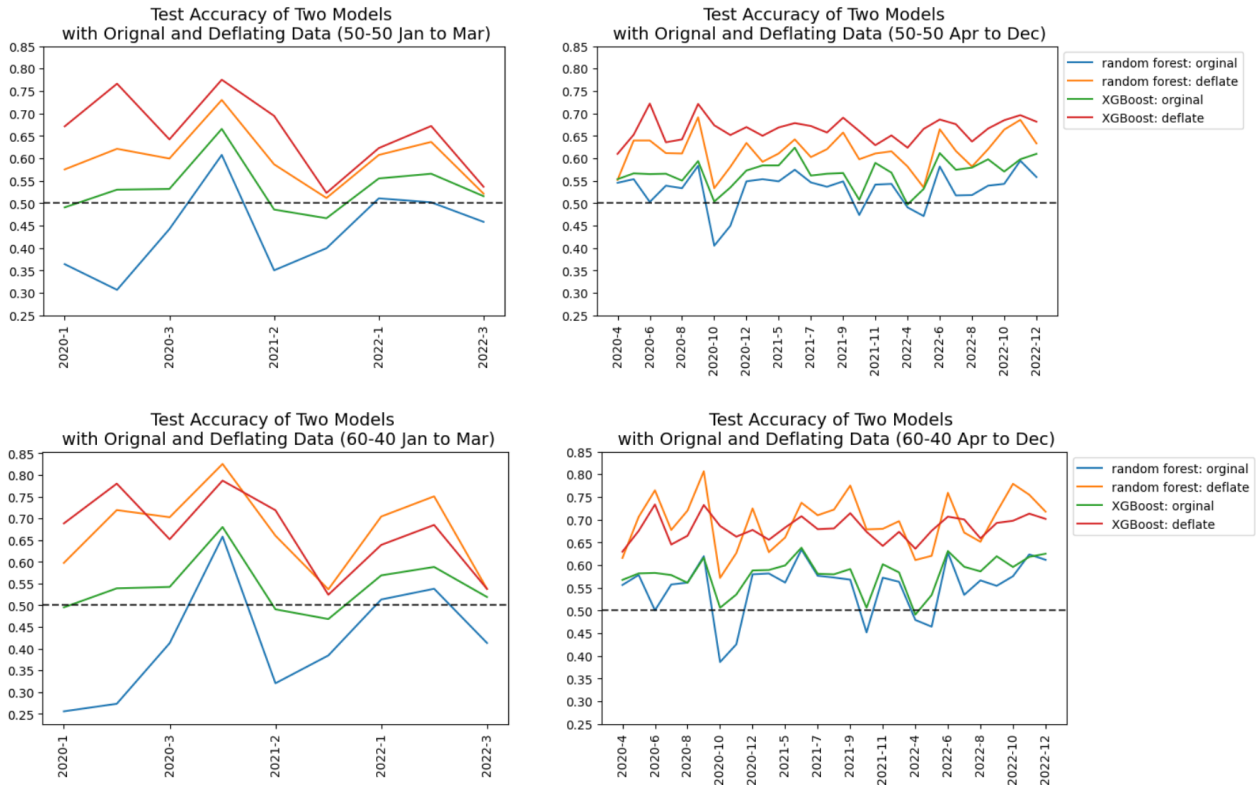
3. 其他 insights:

a. 1~3 月的模型準確度差異大:
如下【圖 5】，沒有進行平減化的模型，準確率較低，1-3 月尤其明顯。 原因可能是營收和財報數據為時間序列資料，比較容易受到前一個月營收表現影響，而年假通常落在 1 月或 2 月，營收波動會較大，因此往後一個月的 2 月和 3 月營收預測難度會較高。此外 1~3 月營收通常會受到前一年同期的營收和過去歷史營收資料的影響，如果模型中沒有將時間序列特徵考慮進去，較可能出現預測偏差，使得有無進行資料平減對模型準確率影響大。

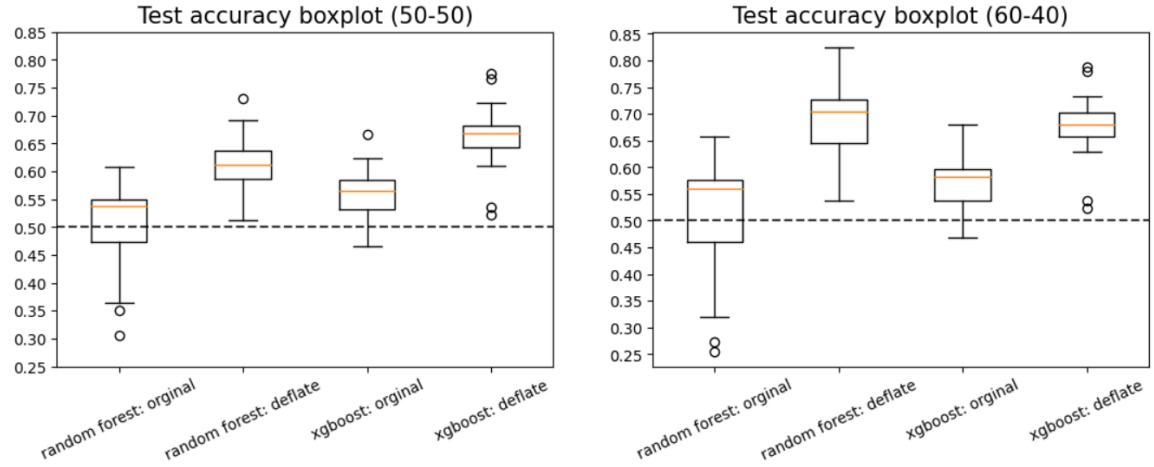
b. 如下【圖 6】，60-40 衡量條件的準確率略高於 50-50:
當衡量條件的閾值降低時(由 60% 降為 50%)，模型會傾向把更多樣本歸類為 1(認定營收上升)，而忽略掉一些實際可能營收下降的樣本，進而產生預測偏差，影響模型準確率。

c. 如下【圖 7】，最差模型時間點解讀:
最差模型為使用原始資料、Random Forest 進行預測的 2020 年 2 月的模型，除了沒有進行平減化、使用隨機森林模型而非 XGBoost 模型造成準確率較低以外，2020 年 2 月剛好是疫情爆發初期，各產業市場極度不穩定，因此提高預測難度。

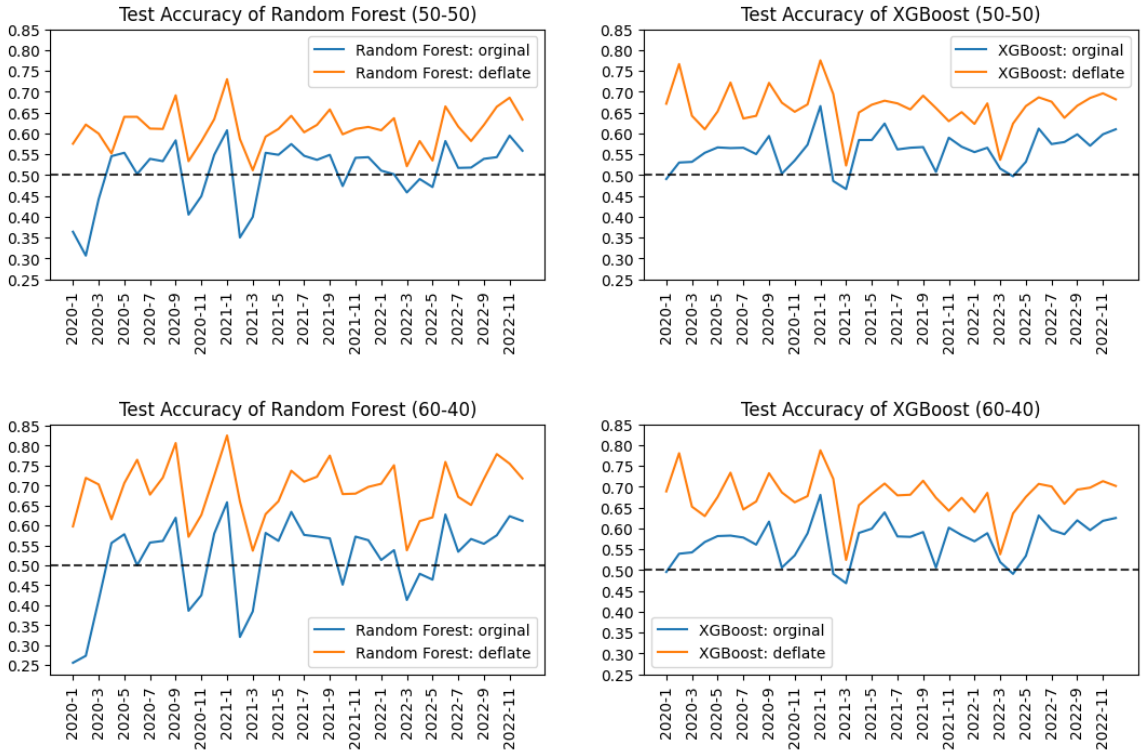
【圖 5、各模型於 1~3 月、4~12 月準確率比較】



【圖 6、50-50、60-40 衡量條件下的各模型準確率比較】



【圖 7、各模型於各衡量標準下各期的準確率】

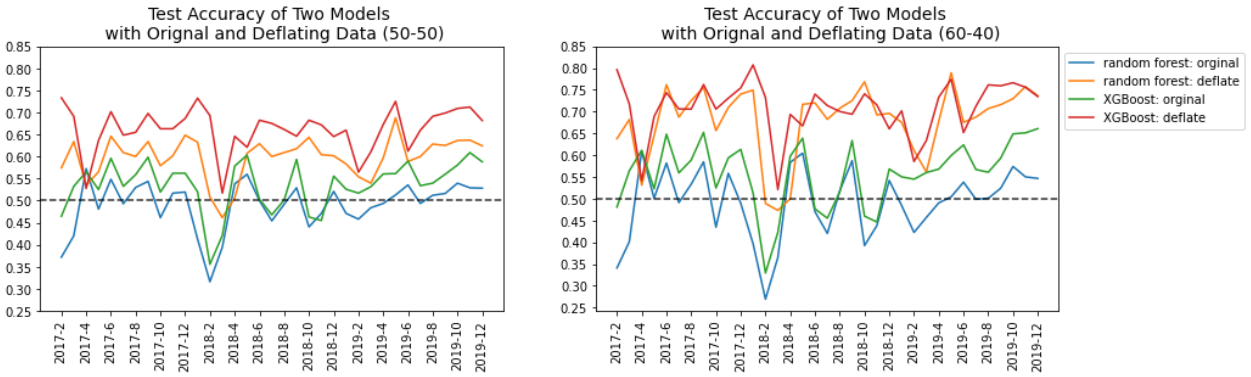


Q2: 和 2017/2 至 2019/12 的預測結果做比較

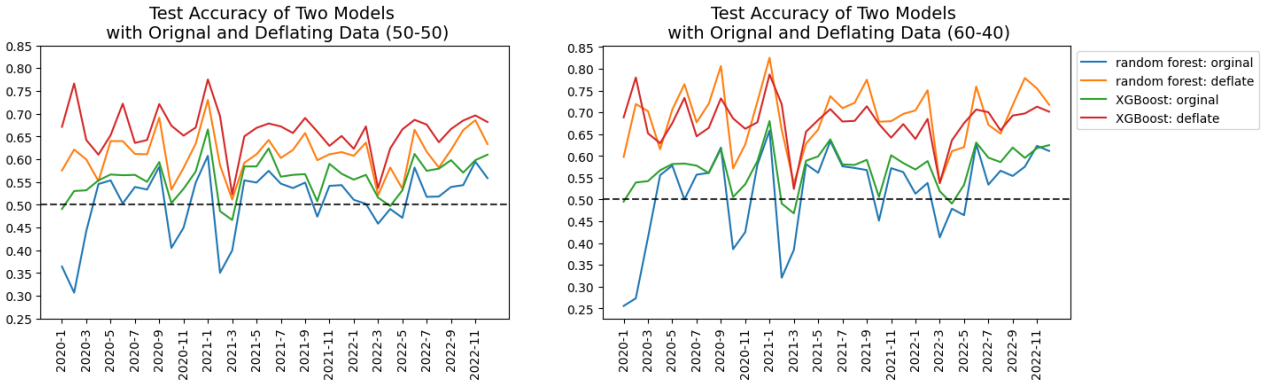
1. **2020/1~2022/12 和 2017/2~2019/12 預測相似處:**
    a. 如下【圖 8】、【圖 9】，4 種模型準確率排序大致皆為 XGBoost_deflate 優於 random forest_deflate 優於 XGBoost_original 優於 random forest_original:
    可能原因如下:
        - 經過資料平減可以減少模型偏差，提高模型準確率。
        - XGBoost 相比於 random forest 多了正則化，可以防止過度擬合，更好的控制模型複雜度來提高預測準確率。
    b. 如下【圖 10】~【圖 13】，T-1 變數皆在最佳模型中排名相較最差模型靠後:
    如上所述，前一個月營收數字確實會對 t 期營收有一定程度影響，因 t-1 若營收高，表示公司在該時點於市場中表現好，但前一期的營收

很可能受到一些隨機的因素影響，例如放假、促銷等行銷活動這些特殊短期活動通常只會稍微影響下一個月的營收，但不一定會是最重要的影響參數。

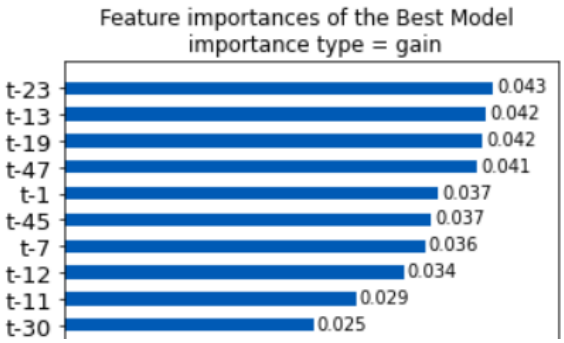
【圖 8、2017/2~2019/12 各模型準確率排序比較】



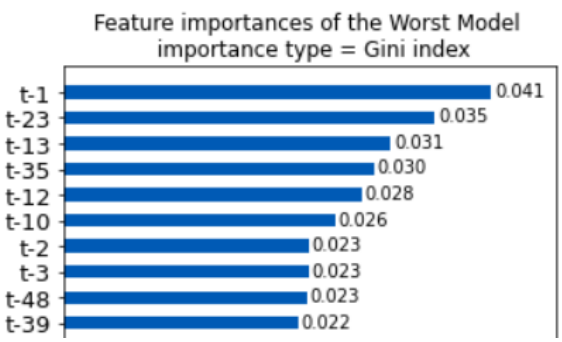
【圖 9、2020/1~2022/12 各模型準確率排序比較】



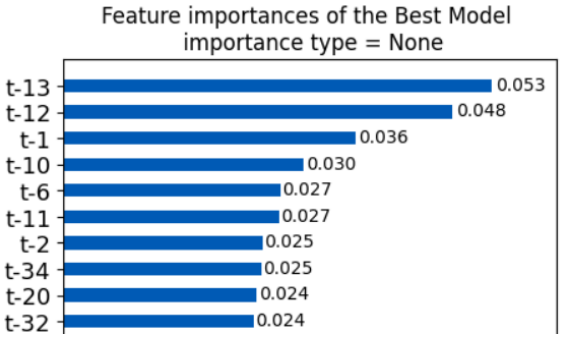
【圖 10、2017/2~2019/12 最佳模型前十重要變數】



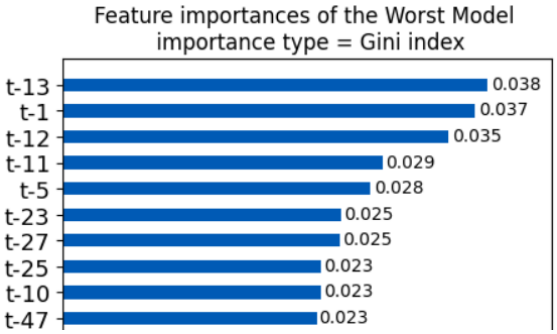
【圖 11、2017/2~2019/12 最差模型前十重要變數】



【圖 12、2020/1~2022/12 最佳模型前十重要變數】



【圖 13、2020/1~2022/12 最差模型前十重要變數】



2. 2020/1~2022/12 和 2017/2~2019/12 預測相異處:

a. 如下【圖 14】，2018 年初模型準確度明顯較低:
可能因為中美貿易戰造成以下影響，導致營收預測較不準確:

- 市場需求變化波動大: 中美貿易戰可能會導致某些產品的進出口受到限制或增加關稅，特別是台灣公司在中國生產的產品可能也要面對更高的成本，進而影響到市場的需求，導致產品的銷售量和價格發生變化而影響到營收。
- 貨幣匯率波動大: 中美貿易戰可能會導致人民幣匯率貶值，進而影響到台灣出口產品的價格和成本。這可能會對台灣公司的營收和利潤率產生影響。
- 供應鏈調整: 如果中美貿易戰導致全球供應鏈發生變化，台灣公司可能需要調整其生產和供應鏈策略。如果台灣公司的供應鏈涉及到中國和美國，這可能會對其營收波動產生影響。
- 政策變動: 中美貿易戰可能會導致政策風險的增加，例如貿易政策、稅收政策和投資限制等。這可能會對台灣公司的營收預測和投資決策產生不確定性而影響營收。

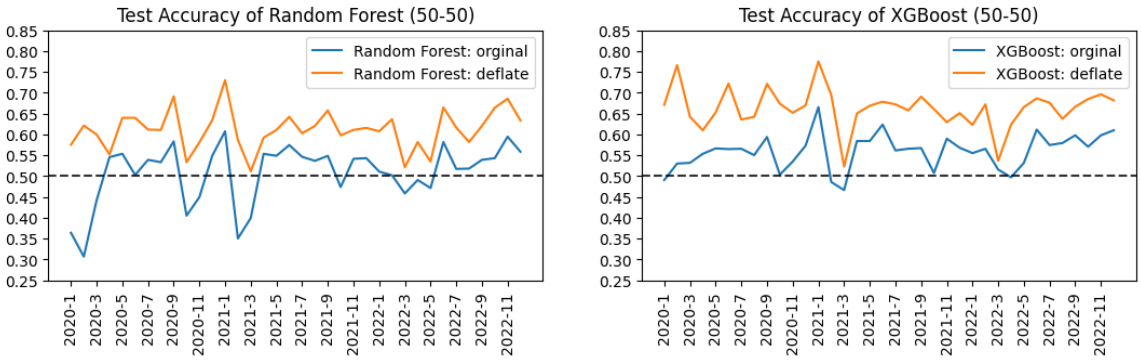
b. 如下【圖 15】，2021 年 2 月模型準確度明顯較低:
如前所述，2020 年 2 月剛好是疫情爆發初期，各產業市場極度不穩定，因此提高預測難度。

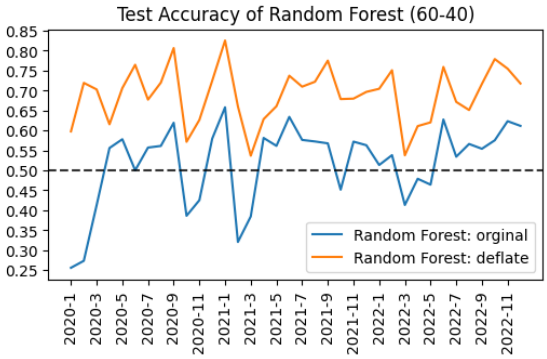
c. 如下【圖 16】、【圖 17】，2020/01 - 2022/12 Random forest_original 的模型精準度優於 2017/02 - 2019/12 的預測:
推測以下為可能原因:

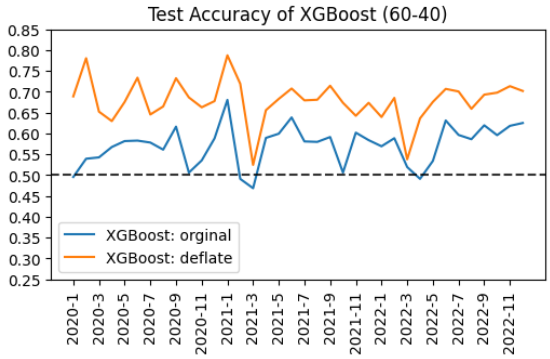
- 資料品質不同：2017~2019 可能受到各種因素影響，導致資料完整性、一致性或準確性低於 2020~2022 的資料，進而影響模型的準確率。
- 市場變化：2017~2019 可能市場波動較大，導致模型準確率較低。

【圖 14、2017/2~2019/12 各模型於各衡量標準下各期的準確率】

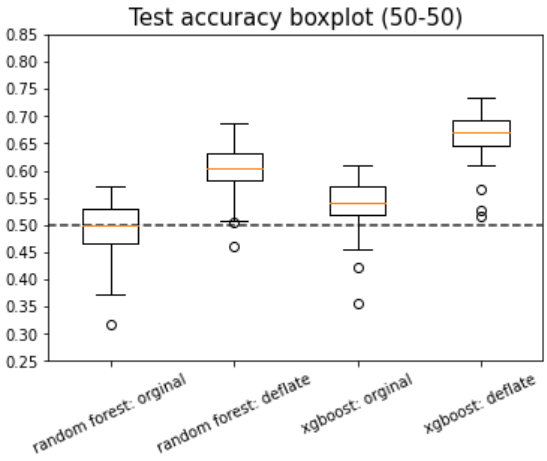
【圖 15、2020/1~2022/12 各模型於各衡量標準下各期的準確率】

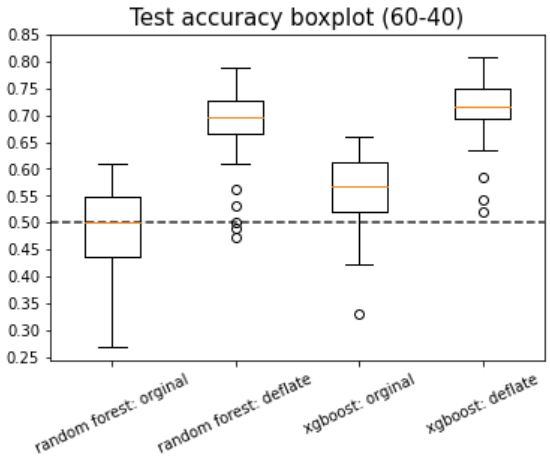




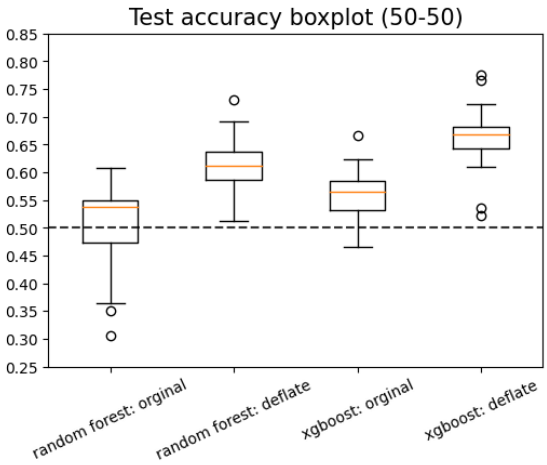


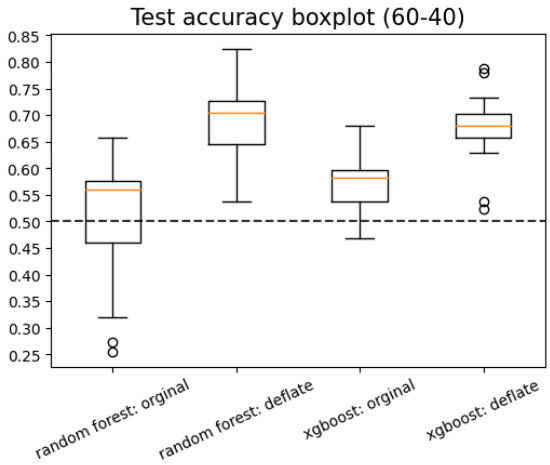
【圖 16、2017/2~2019/12 各模型準確率】





【圖 17、2020/1~2022/12 各模型準確率】





第二部分

Q1: 定義所挑選的產業 (以程式碼進行篩選)

1. 產業選擇:

選擇「半導體」(TSE 產業別 = 24)。

2. 選擇依據:

主要想要驗證半導體產業受季節性影響的程度是否明顯。就我對半導體產業的概略了解，智慧型手機、電腦等對半導體有需求的產品通常會是季節性的，譬如在假期或季節促銷的時候銷量最高。想藉此了解終端產品的需求有季節性是否也會導致上游原料受季節性影響，也就是看看 t-12、t-10、t-11 是否為最佳模型重要參數的前幾名。

Q2: 分析 2020/1 至 2022/12 月營收增減的預測結果（須標明使用之模型，並說明資料集前處理的方式）

1. 資料集處理:

複製已經經過保留月營收大於 0、修改時間資料格式資料處理的 dataframe，命名為 remain_semi，將所有非「半導體」分類的公司資料全部刪除後， 資料檔中共有 129 間公司。

2. 預測模型:

使用 Random forest 及 XGBoost 模型。

3. 資料平減:

使用 t-48 到 t-1 期的平均數和標準差進行平減，也就是把資料減去 t-48 到 t-1 的平均數後除以 t-48 到 t-1 的標準差，。經過確認，平減後的資料平均數比較接近 0、標準差接近 1。但或許因為資料筆數不足，因此平減效果沒有到非常的好。

4.模型準確率預測結果:

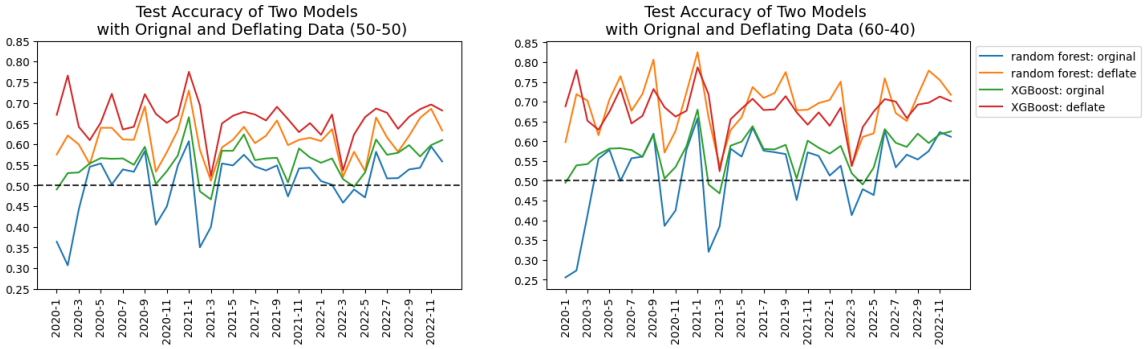
4-1 . 如下【圖 18】，半導體產業模型同樣在 2021 年 2 月表現較差，可能原因是受到疫情影響，大家對筆電、平板、手機等使用量提高、醫院對呼吸機或監控設備需求提高、電子物流業對無接觸運送機器的需求也提升，導致需要半導體的電子設備需求短期內需求突然暴增，因此模型較難進行營收預測。

4-2. 如下【圖 18】，半導體產業模型準確率排序同樣大致皆為 XGBoost_deflate 優於 random forest_deflate 優於 XGBoost_original 優於 random forest_original。

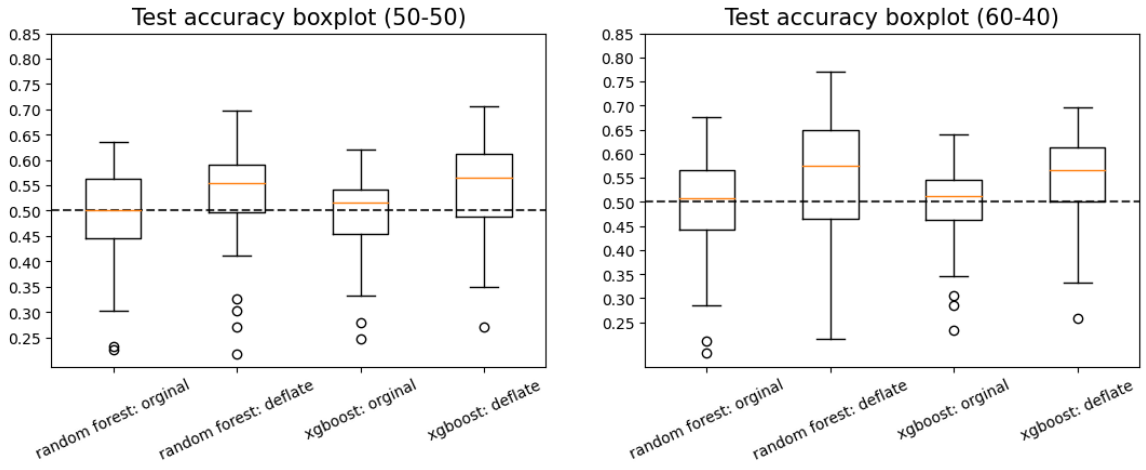
4-3. 如下【圖 19】，半導體產業同樣在沒有進行平減化的情況下，準確率明顯較低。

4-4. 如下【圖 20】，半導體推測半導體產業公司同樣有放年假，導致產量降低或營收延後認列。

【圖 18、半導體產業 2020/1~2022/12 各模型準確率】



【圖 19、半導體產業 2020/1~2022/12 各模型準確率】



【圖 20、半導體產業各模型於 1~3 月、4~12 月準確率比較】

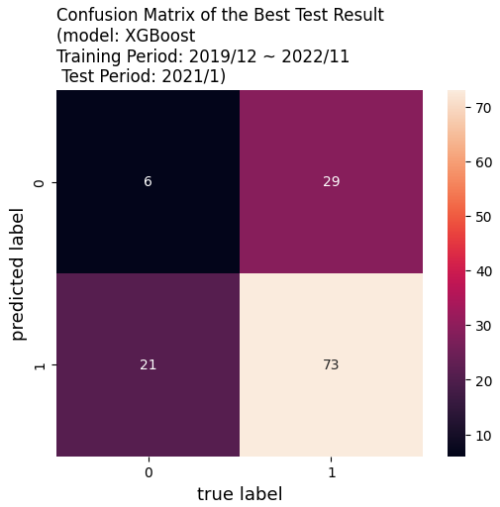
最佳&最差模型分類結果比較:

1. 如下【圖 21】，最佳模型的誤判結果中，高估與低估的比數差異不大，因此模型較不會偏頗判斷，從兩類別分別觀察準確率，真實為達預期月營收的資料準確率為 73/(29+73)，大約為 22%，真實為未達預期月營收的資料準確率為 6/(6+21)，大約為 42%，得知個別來看，模型學習「未達預期月營收資料」的特徵還是較多，模型在「未達預期營收資料」方面學習得比較充分。

2. 如下【圖 22】，最差模型的模型預測結果當中，有高達 95 筆，即 95/108 大約 88%的比例是誤將實為高於預期的月營收，低估為不如預期的月營收，而高估的比率相對較低，因此模型相對容易過於悲觀，缺乏挖掘較佳月營收資訊的能力。

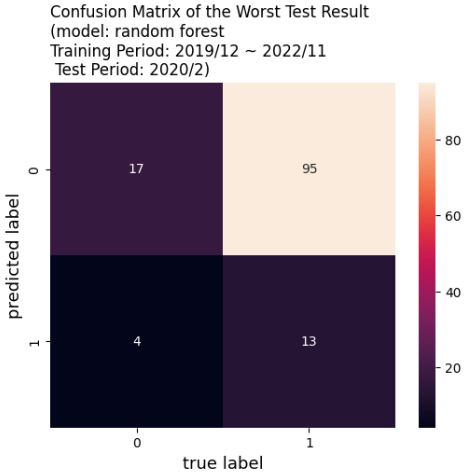
【圖 21、半導體產業最佳模型前十重要變數】

XGBoost_Deflating_2022-4



【圖 22、半導體產業最差模型前十重要變數】

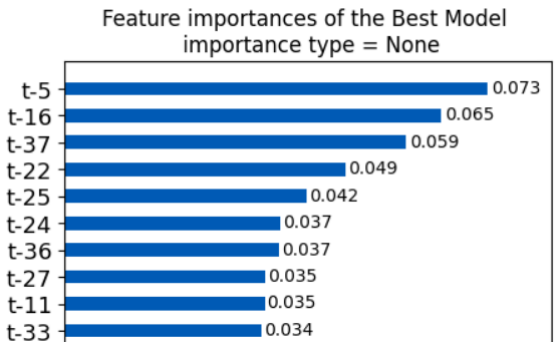
Random Forest_Original_2021-2



最佳模型&最差模型重要變數比較:

如下【圖 23】，半導體產業的最佳模型中， t-1 、t-11、t-12、t-13、t-23、t-24、t-25 皆非最重要變數，因此推測半導體產業營收比較不容易受到前一個月或是去年、前幾年同期的營收影響，推測可能是因為半導體產業容易受到整體經濟環境影響，而景氣的通常不一定規律，所以半導體產業的需求或營收比較不會是季節性的波動。不過因資料筆數較少，也有可能存在些微偏誤。

【圖 23、半導體產業 最佳模型前十重要變數】



【圖 24、半導體產業 最差模型前十重要變數】