Notes: DNN、CNN、Transformer 模型各跑 2018、2020 12 個月。

第一部分

使用 DNN/CNN/Transformer 預測 2020/1 至 2020/12 的月營收金額（共 12 個月）

Q1、分析營收金額預測結果 (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) 並標明判斷結果好壞的衡量指標。

1. 各模型預測能力比較:
    a. CNN 表現相較 DNN、Transformer 差:
        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，DNN 和 Transformer 模型預測出來的金額都離 Expected 較接近，而 CNN 則偏離較遠。
        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及紅線以上，也就是 CNN 相比於 DNN 和 Transformer 有較高的均方根誤差，代表預測值和實際值之間的距離大、模型預測能力較差。
        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及紅線以上，也就是 CNN 相比於 DNN 和 Transformer 有較高的平均絕對誤差，代表預測值和實際值之差地絕對值大、模型預測能力差。

        推測原因如下:

        - 可能是因為每個模型都有各自最佳的使用場景，或許 CNN 更擅長進行圖像或視覺相關的處理，因此在數據或是自然語言方面表現相較 DNN 和 Transformer 較差。

    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， DNN 和 Transformer Model 在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 DNN、CNN 還是 Transformer，實線大部分落在虛線以下，也就是平減後可

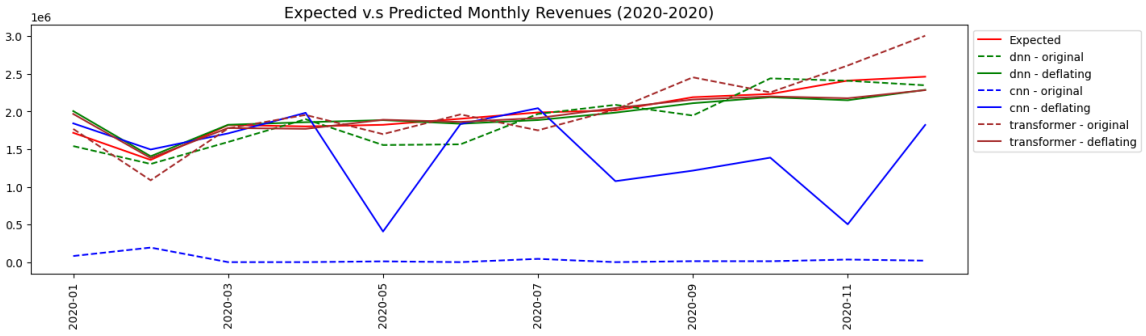
以有較低的均方根誤差，代表預測值和實際值之間的距離小、模型預測能力好。

iii. 由【圖 6】預測分數 MAE 圖形來看，無論使用 DNN、CNN 還是 Transformer，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。

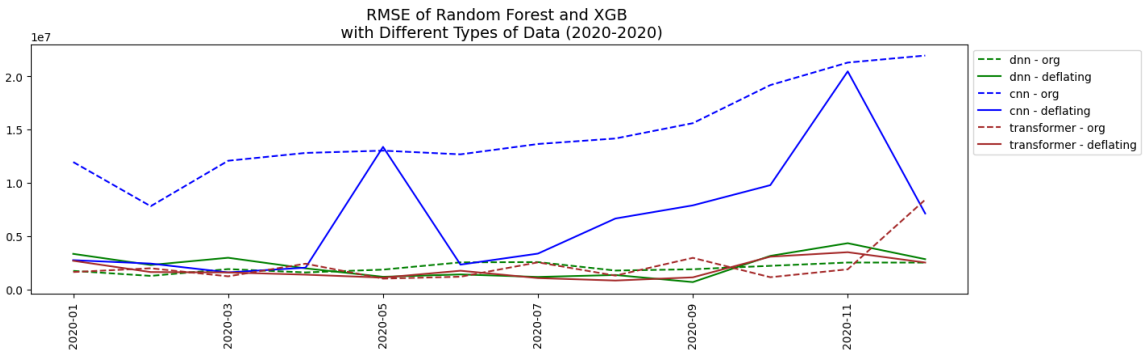
推測原因如下:

- 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。

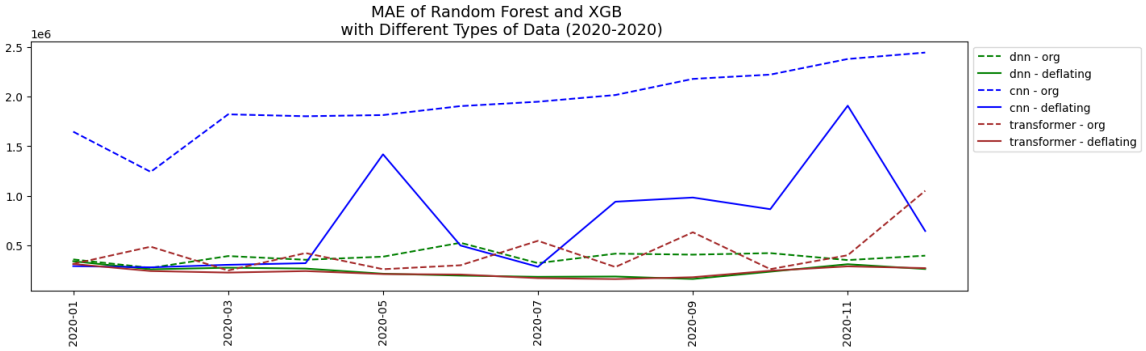
【圖 1、各模型預測營收金額】



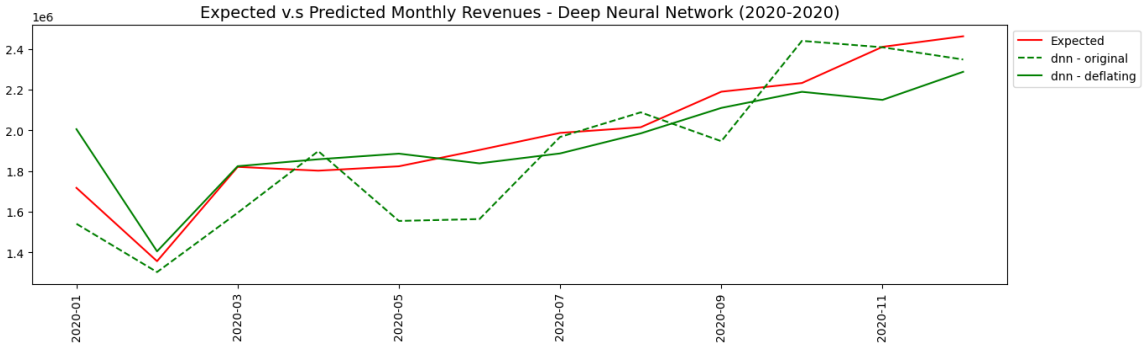
【圖 2、各模型 RMSE】

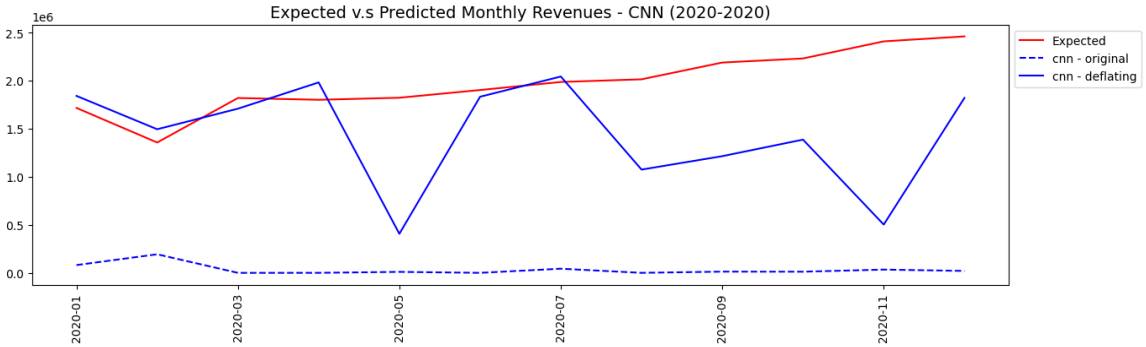


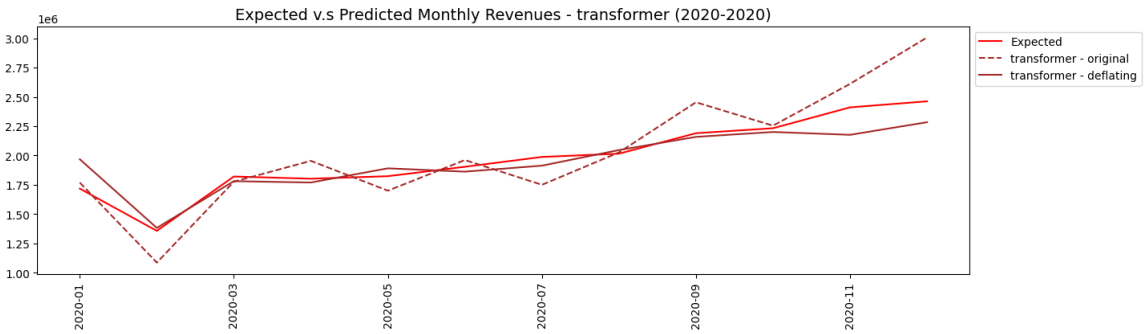
【圖 3、各模型 MAE】



【圖 4、各模型預測營收金額】

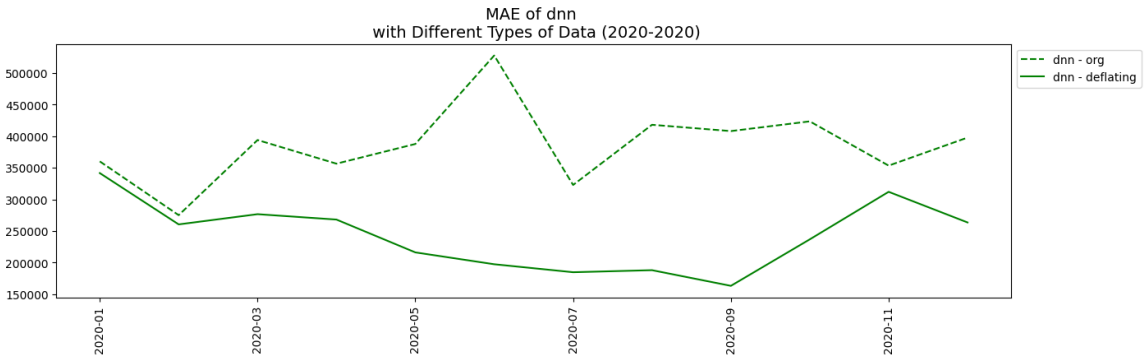


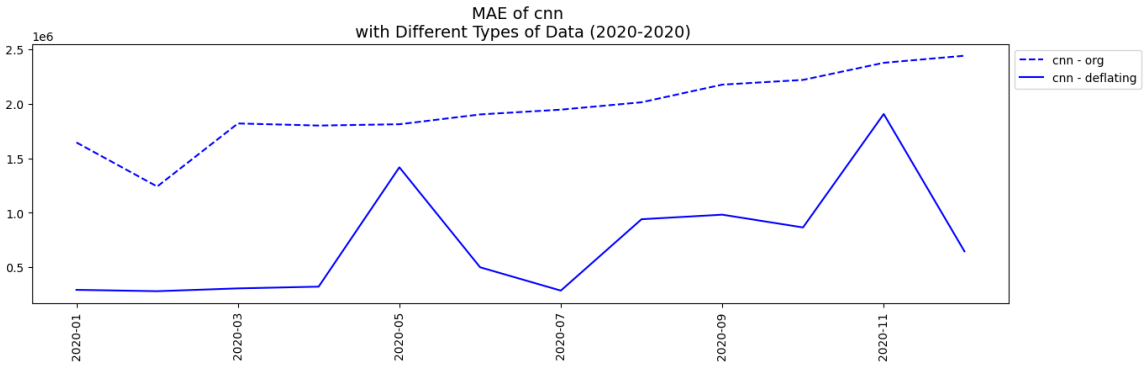


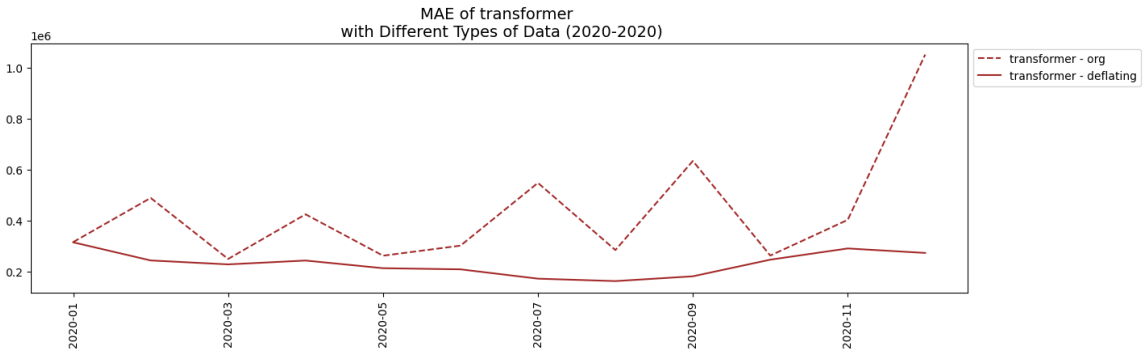


【圖 5、各模型 RMSE】

【圖 6、各模型 MAE】





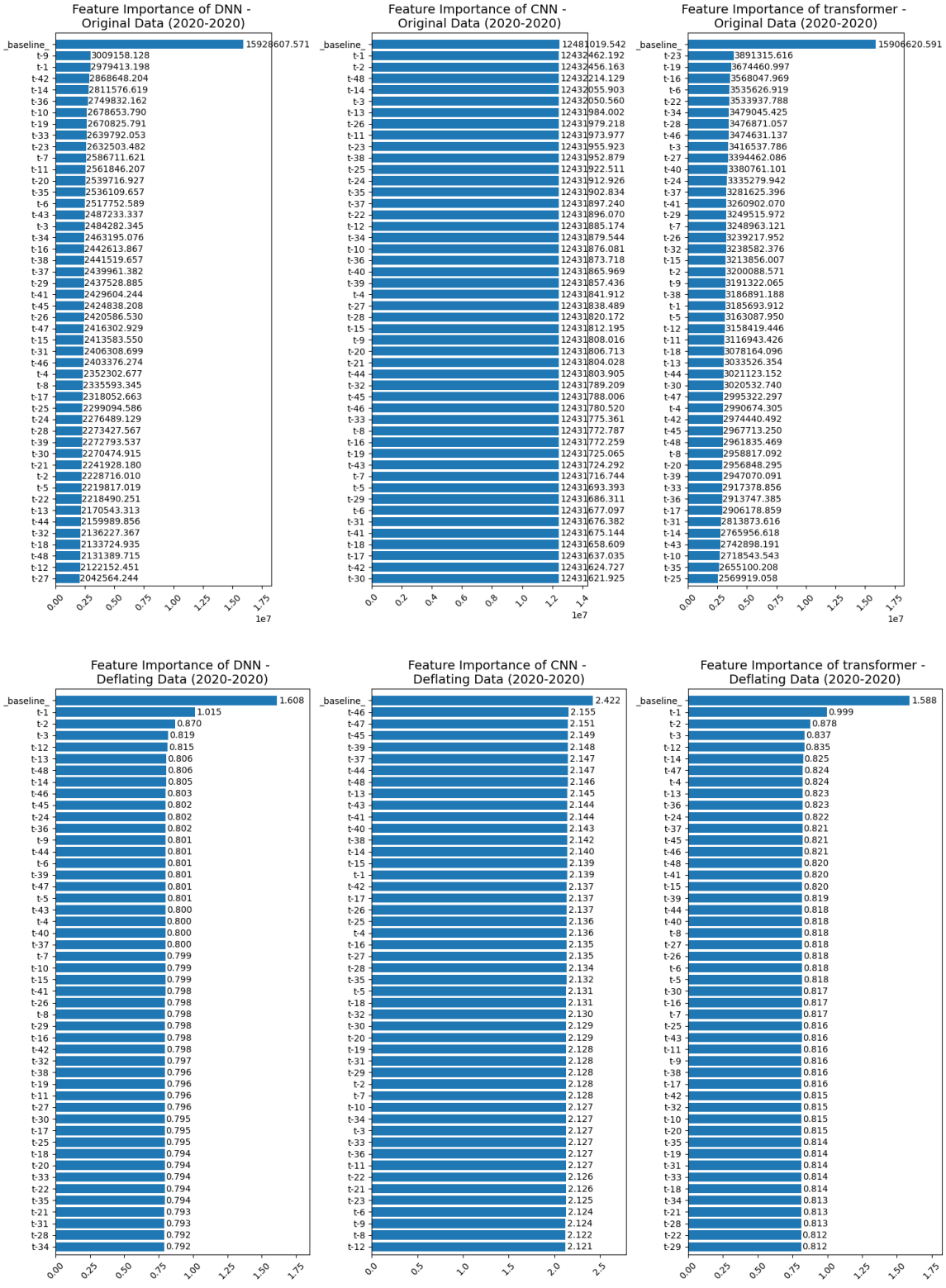


2. 各模型變數重要性:

   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、t - 13、t - 14: 前一年或前一年附近月份的數據反映季節性變化，因此對預測也重要。
   c. t - 36、t - 48: 前三年獲四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】


Feature Importance of DNN - Original Data (2020-2020)
baseline 15928607.571
t-9 3009158.128
t-1 2979413.198
t-42 2868648.204
t-14 2811576.619
t-36 2749832.162
t-10 2678653.790
t-19 2670825.791
t-33 2639792.053
t-23 2632503.482
t-7 2586711.621
t-11 2561846.207
t-20 2539716.927
t-35 2536109.657
t-6 2517752.589
t-43 2487233.337
t-3 2484282.345
t-34 2463195.076
t-16 2442613.867
t-38 2441519.657
t-37 2439961.382
t-29 2437528.885
t-41 2429604.244
t-45 2424838.208
t-26 2420586.530
t-47 2416302.929
t-15 2413583.550
t-31 2406308.699
t-46 2403376.274
t-4 2352302.677
t-8 2335593.345
t-17 2318052.663
t-25 2299094.586
t-24 2276489.129
t-28 2273427.567
t-39 2272793.537
t-30 2270474.915
t-21 2241928.180
t-2 2228716.010
t-5 2219817.019
t-22 2218490.251
t-13 2170543.313
t-44 2159989.856
t-32 2136227.367
t-18 2133724.935
t-48 2131389.715
t-12 2122152.451
t-27 2042564.244
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75
1e7
Feature Importance of CNN - Original Data (2020-2020)
baseline 12481019.542
t-1 12432462.192
t-2 12432456.163
t-48 12432214.129
t-14 12432055.903
t-3 12432050.560
t-13 12431984.002
t-26 12431979.218
t-11 12431973.977
t-23 12431955.923
t-38 12431952.879
t-25 12431922.511
t-24 12431912.926
t-35 12431902.834
t-37 12431897.240
t-22 12431896.070
t-12 12431885.174
t-34 12431879.544
t-10 12431876.081
t-36 12431873.718
t-40 12431865.969
t-39 12431857.436
t-4 12431841.912
t-27 12431838.489
t-28 12431820.172
t-15 12431812.195
t-9 12431808.016
t-20 12431806.713
t-21 12431804.028
t-44 12431803.905
t-32 12431789.209
t-45 12431788.006
t-46 12431780.520
t-33 12431775.361
t-8 12431772.787
t-16 12431772.259
t-19 12431725.065
t-43 12431724.292
t-7 12431716.744
t-5 12431693.393
t-29 12431686.311
t-6 12431677.097
t-31 12431676.382
t-41 12431675.144
t-18 12431658.609
t-17 12431637.035
t-42 12431624.727
t-30 12431621.925
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
1e7
Feature Importance of transformer - Original Data (2020-2020)
baseline 15906620.591
t-23 3891315.616
t-19 3674460.997
t-16 3568047.969
t-6 3535626.919
t-22 3533937.788
t-34 3479045.425
t-28 3476871.057
t-46 3474631.137
t-3 3416537.786
t-27 3394462.086
t-40 3380761.101
t-24 3335279.942
t-37 3281625.396
t-41 3260902.070
t-29 3249515.972
t-7 3248963.121
t-26 3239217.952
t-32 3238582.376
t-15 3213856.007
t-2 3200088.571
t-9 3191322.065
t-38 3186891.188
t-1 3185693.912
t-5 3163087.950
t-12 3158419.446
t-11 3116943.426
t-18 3078164.096
t-13 3033526.354
t-44 3021123.152
t-30 3020532.740
t-47 2995322.297
t-4 2990674.305
t-42 2974440.492
t-45 2967713.250
t-48 2961835.469
t-8 2958817.092
t-20 2956848.295
t-39 2947070.091
t-33 2917378.856
t-36 2913747.385
t-17 2906178.859
t-31 2813873.616
t-14 2765956.618
t-43 2742898.191
t-10 2718543.543
t-35 2655100.208
t-25 2569919.058
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75
1e7
Feature Importance of DNN - Deflating Data (2020-2020)
baseline 1.608
t-1 1.015
t-2 0.870
t-3 0.819
t-12 0.815
t-13 0.806
t-48 0.806
t-14 0.805
t-46 0.803
t-45 0.802
t-24 0.802
t-36 0.802
t-9 0.801
t-44 0.801
t-6 0.801
t-39 0.801
t-47 0.801
t-5 0.801
t-43 0.800
t-4 0.800
t-40 0.800
t-37 0.800
t-7 0.799
t-10 0.799
t-15 0.799
t-41 0.798
t-26 0.798
t-8 0.798
t-29 0.798
t-16 0.798
t-42 0.798
t-32 0.797
t-38 0.796
t-19 0.796
t-11 0.796
t-27 0.796
t-30 0.795
t-17 0.795
t-25 0.795
t-18 0.794
t-20 0.794
t-33 0.794
t-22 0.794
t-35 0.794
t-21 0.793
t-31 0.793
t-28 0.792
t-34 0.792
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75
Feature Importance of CNN - Deflating Data (2020-2020)
baseline 2.422
t-46 2.155
t-47 2.151
t-45 2.149
t-39 2.148
t-37 2.147
t-44 2.147
t-48 2.146
t-13 2.145
t-43 2.144
t-41 2.144
t-40 2.143
t-38 2.142
t-14 2.140
t-15 2.139
t-1 2.139
t-42 2.137
t-17 2.137
t-26 2.137
t-25 2.136
t-4 2.136
t-16 2.135
t-27 2.135
t-28 2.134
t-35 2.132
t-5 2.131
t-18 2.131
t-32 2.130
t-30 2.129
t-20 2.129
t-19 2.128
t-31 2.128
t-29 2.128
t-2 2.128
t-7 2.128
t-10 2.127
t-34 2.127
t-3 2.127
t-33 2.127
t-36 2.127
t-11 2.127
t-22 2.126
t-21 2.126
t-23 2.125
t-6 2.124
t-9 2.124
t-8 2.122
t-12 2.121
0.0 0.5 1.0 1.5 2.0 2.5
Feature Importance of transformer - Deflating Data (2020-2020)
baseline 1.588
t-1 0.999
t-2 0.878
t-3 0.837
t-12 0.835
t-14 0.825
t-47 0.824
t-4 0.824
t-13 0.823
t-36 0.823
t-24 0.822
t-37 0.821
t-45 0.821
t-46 0.821
t-48 0.820
t-41 0.820
t-15 0.820
t-39 0.819
t-44 0.818
t-40 0.818
t-8 0.818
t-27 0.818
t-26 0.818
t-6 0.818
t-5 0.818
t-30 0.817
t-16 0.817
t-7 0.817
t-25 0.816
t-43 0.816
t-11 0.816
t-9 0.816
t-38 0.816
t-17 0.816
t-42 0.815
t-32 0.815
t-10 0.815
t-20 0.815
t-35 0.814
t-19 0.814
t-31 0.814
t-33 0.814
t-18 0.814
t-34 0.813
t-21 0.813
t-28 0.813
t-22 0.812
t-29 0.812
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75

3. 最佳 & 最差模型

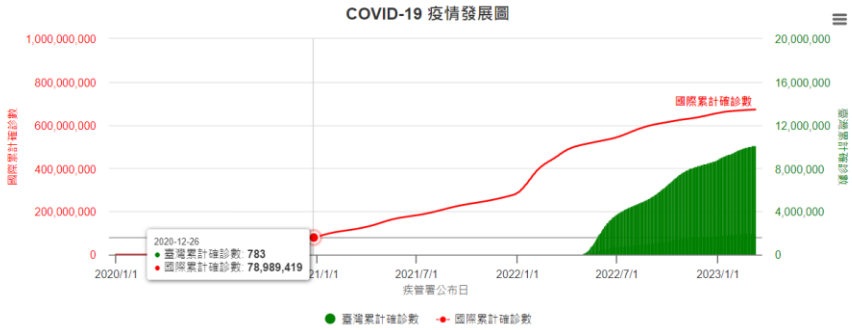
*以 **RMSE** 作為模型好壞衡量標準，最佳及最差模型特性分析如下:*

a. 如【圖 8】，排除表現較差的 CNN 模型，DNN 及 Transformer 兩者最佳模型在 2020-08 及 2020-09

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】，CNN、DNN 及 Transformer 最差模型皆在 2020 年底

i. 推測可能的原因是 2020 年疫情流行，到年底呈現升溫趨勢，全球經濟的衰退或是消費模式變化導致模型較難進行準確的預測。



【圖 8、以 RMSE 衡量的最佳模型】

modelName	dataType	scoreType	min_month	min_score
dnn	def	RMSE	2020-09	680977.0

modelName	dataType	scoreType	min_month	min_score
cnn	def	RMSE	2020-03	1591284.0

modelName	dataType	scoreType	min_month	min_score
transformer	def	RMSE	2020-08	827583.0

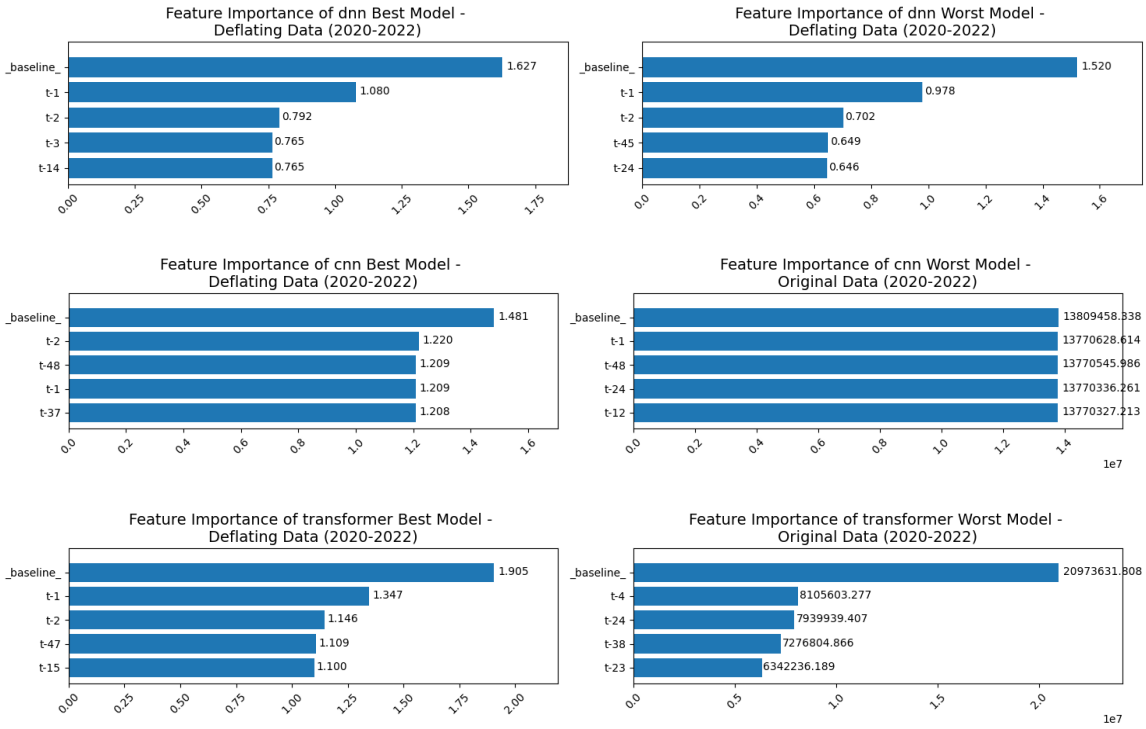
【圖 9、以 RMSE 衡量的最差模型】

	modelName	dataType	scoreType	max_month	max_score
2	dnn	def	RMSE	2020-11	4333622.0

modelName	dataType	scoreType	max_month	max_score
cnn	org	RMSE	2020-12	21960777.0

modelName	dataType	scoreType	max_month	max_score
transformer	org	RMSE	2020-12	8421225.0

【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】，排除表現較差的 CNN 模型，DNN 及 Transformer 兩者最佳模型同樣在 2020-08 及 2020-09
    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】，排除表現相較好的 DNN 最差模型，CNN、Transformer 的最差模型皆在 2020-12
    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-24、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
dnn	def	MAE	2020-09	163233.0

modelName	dataType	scoreType	min_month	min_score
cnn	def	MAE	2020-02	280394.0

modelName	dataType	scoreType	min_month	min_score
transformer	def	MAE	2020-08	162106.0

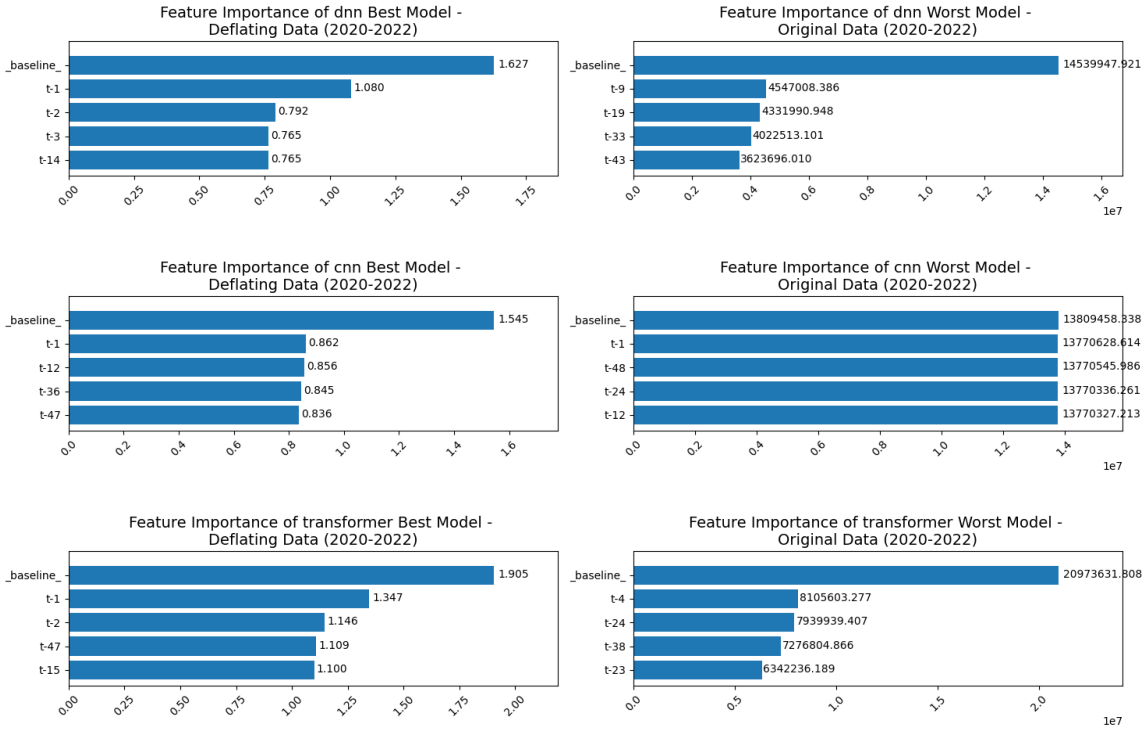
【圖 12、以 MAE 衡量的最差模型】

modelName	dataType	scoreType	max_month	max_score
dnn	org	MAE	2020-06	527774.0

modelName	dataType	scoreType	max_month	max_score
cnn	org	MAE	2020-12	2442582.0

modelName	dataType	scoreType	max_month	max_score
transformer	org	MAE	2020-12	1050891.0

【圖 13、以 MAE 衡量的最佳&最差模型重要性比較】



Q2、和 2018 的預測結果做比較，並與作業三 RandomForest 和 XGBoost 的結果做比較。

Q2-1: 2018 和 2020 預測結果比較:

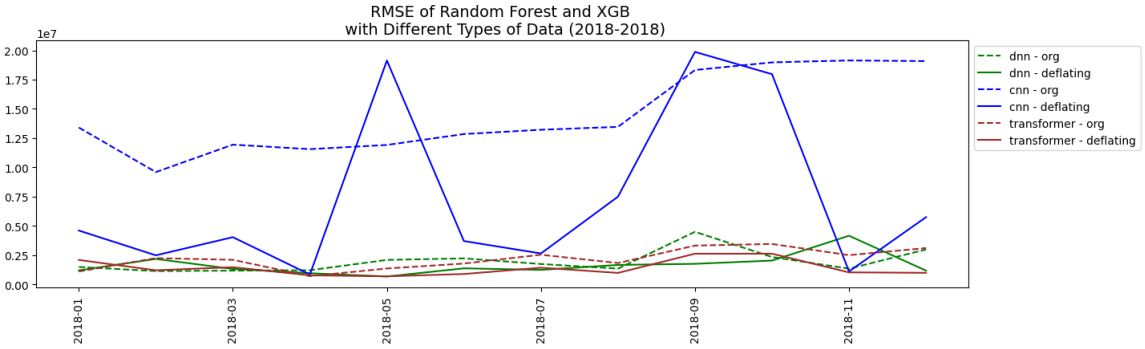
1. 相同:
    a. DNN、Transformer 模型相較 CNN 準確度較高: 由【圖 14】~【圖 17】2018 及 2020 年 RMSE、MAE 圖片比較可知，無論預測年度為何或是

資料是否經過平減，DNN、Transformer 模型表現大致比 CNN 還好，可能因為 CNN 較適用於視覺化的處理，所相較例外兩個模型較難準確進行數據預測。

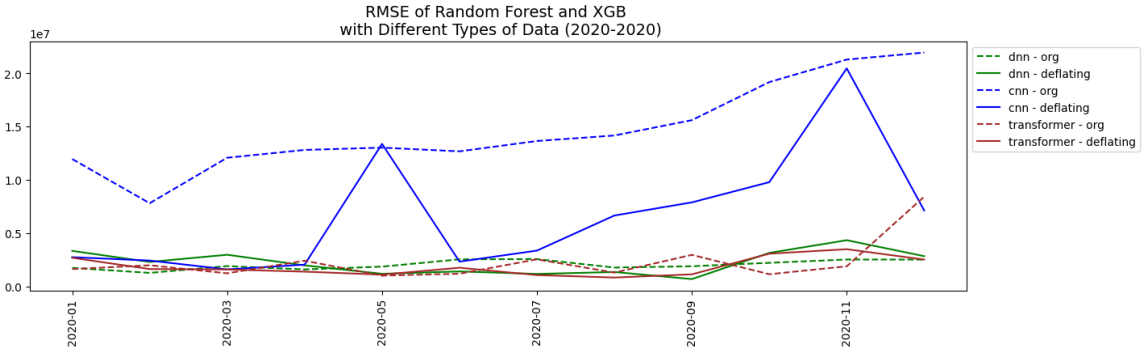
b. 進行資料平減後模型預測能力較高: 由【圖 14】~【圖 17】2018 及 2020 年 RMSE、MAE 圖片比較可知，無論使用 DNN、CNN 或是 Transformer 模型進行預測，進行平減後的資料大致表現較好，因為資料經過平減後，可以減少異常值對模型的影響、提高模型學習效率和穩定性，以及準確性。

c. t-1、t-2 為最佳模型重要變數: 由【圖 18】~【圖 18】2018 及 2020 年 RMSE、MAE 衡量標準下各模型最佳/最差模型最佳變數圖片 可知，t-1、t-2 為最佳模型預測時的重要變數，推測因為月營收屬於時間序列資料，而過去資料能夠提供數據趨勢，且當月資料尤其受到前 1~2 個月影響最深，因此 t-1、t-2 為預測時的重要變數。

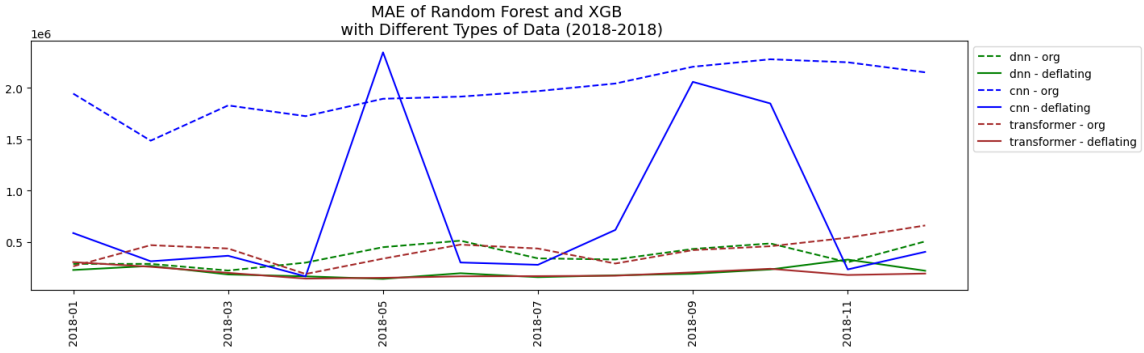
【圖 14、2018 RMSE】



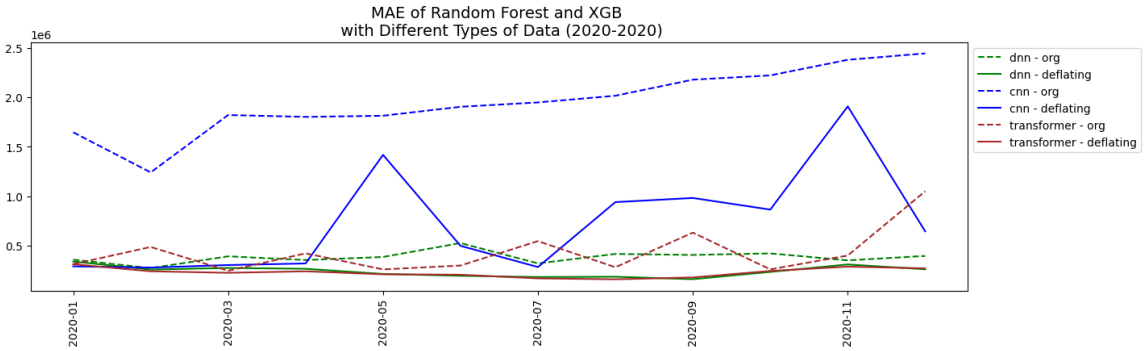
【圖 15、2020 RMSE】



【圖 16、2018 MAE】

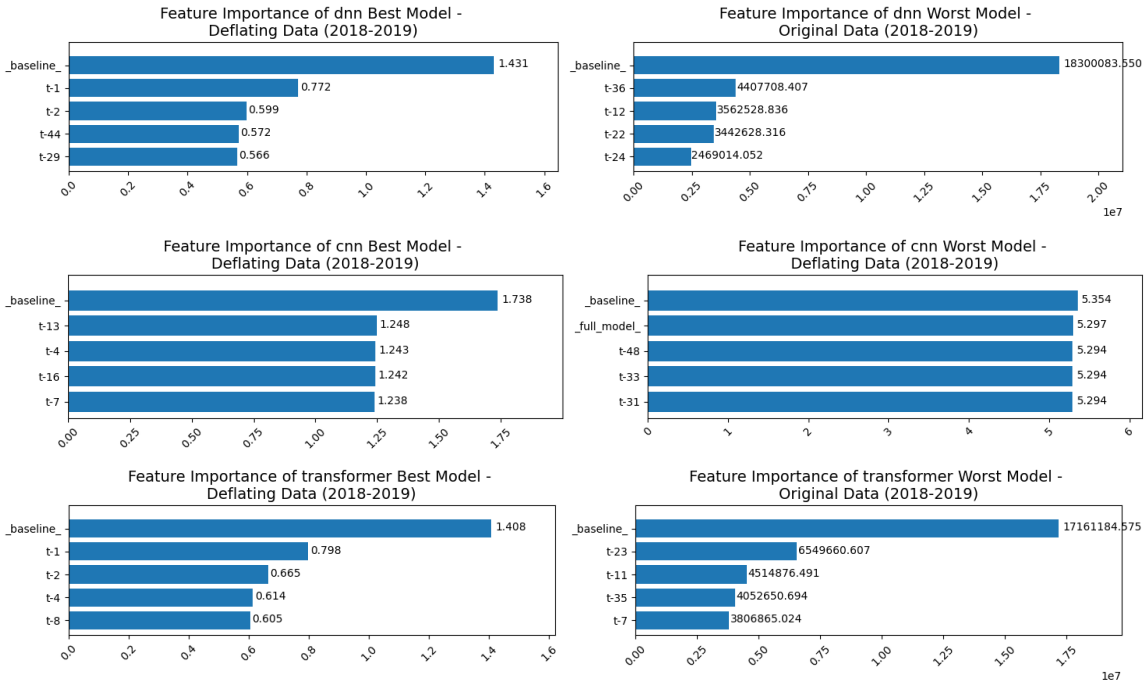


【圖 17、2020 MAE】

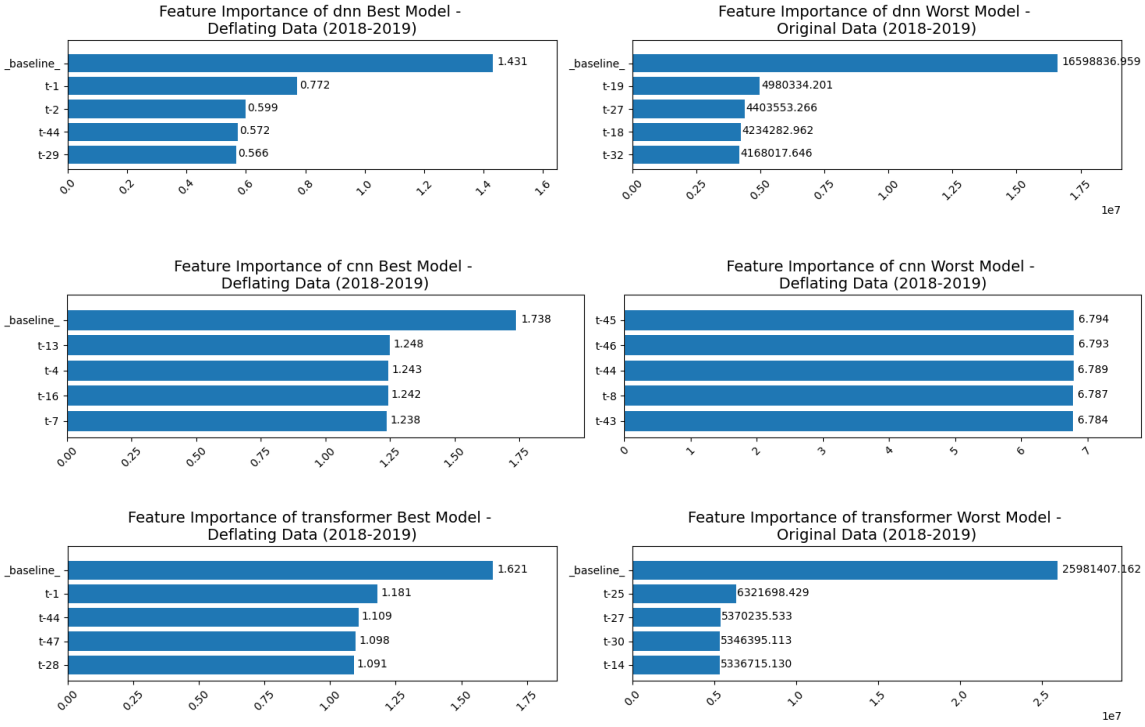


【圖 18、2018 最佳/最差模型重要變數】

RMSE

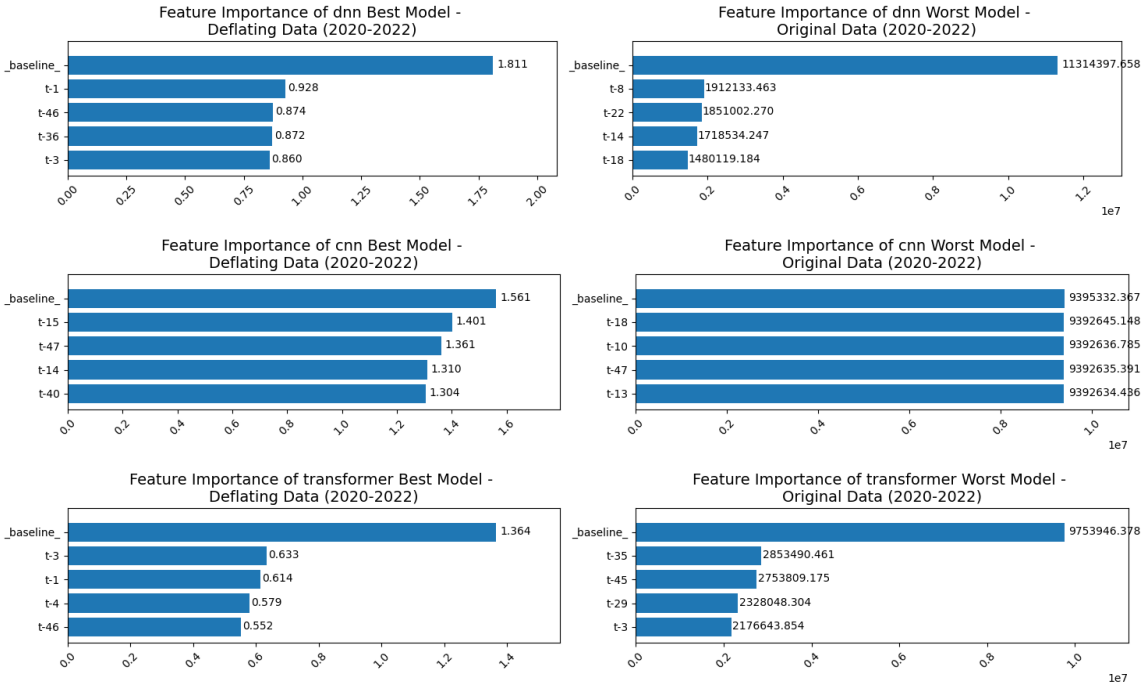


MAE

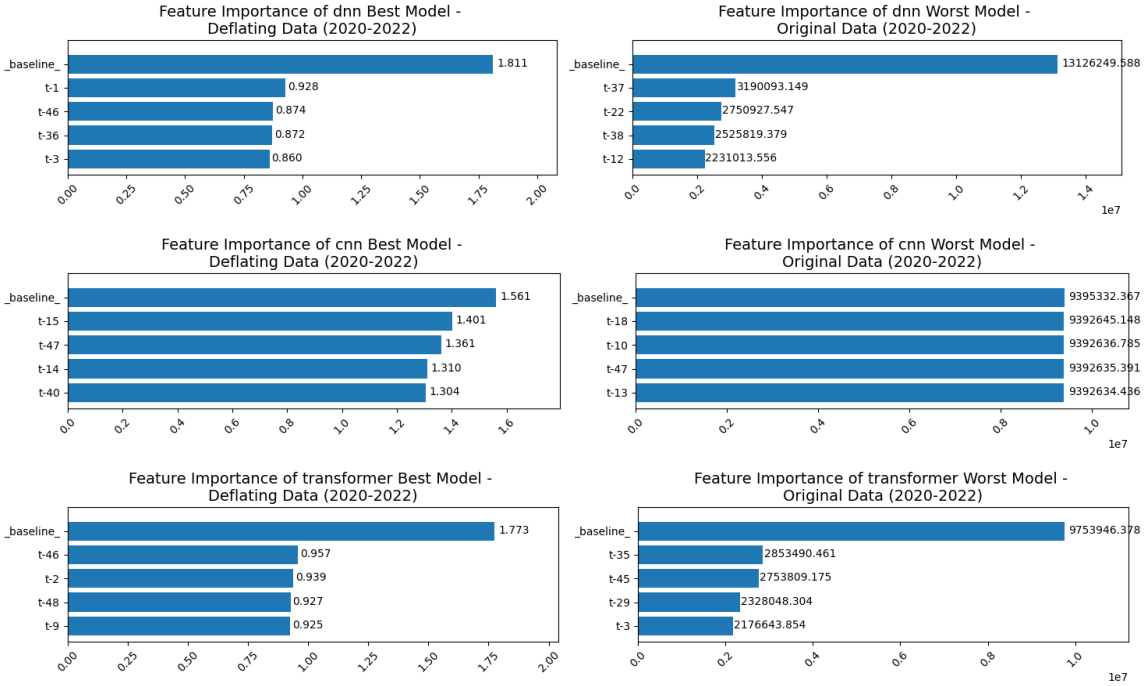


【圖 19、2020 最佳/最差模型重要變數】

RMSE

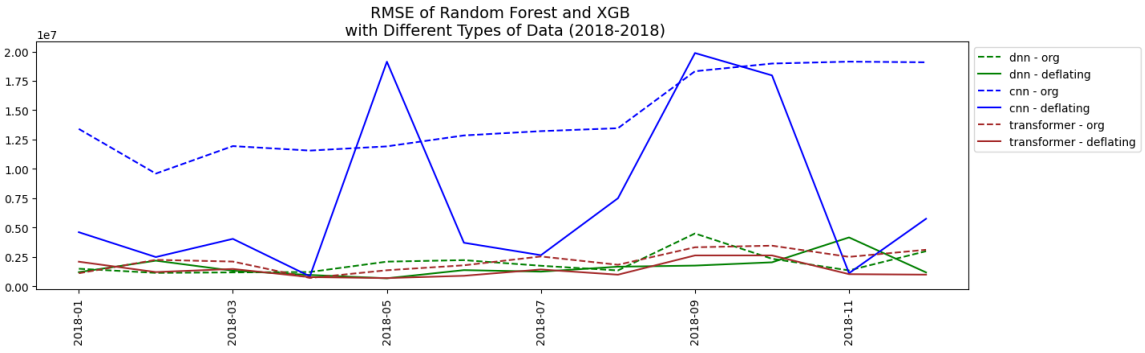


MAE

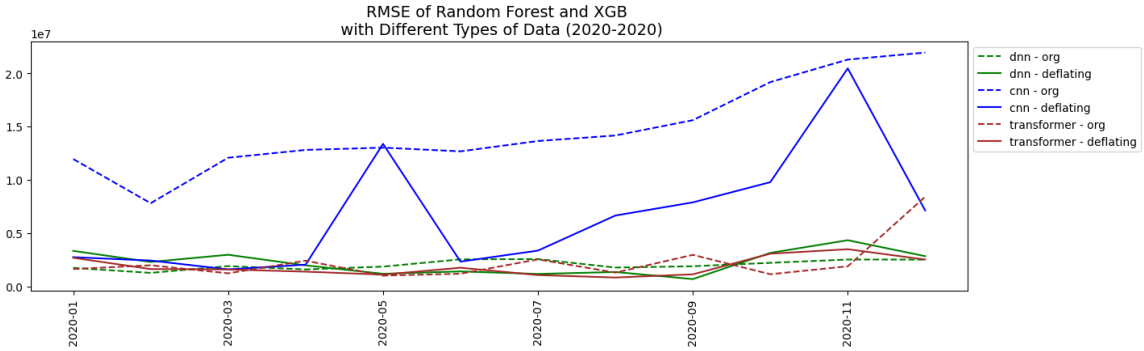


2. 相異:
   a. 近期預測相比於以前預測能力變差: 由【圖 20】~【圖 23】RMSE 以及 MAE 圖片比較可發現 202001-202212 的預測表現相比於 201801-201912 來得差，推測是因為近幾年受到疫情等因素影響，導致使得經濟環境或市場變化較大，模型無法完全捕捉這些較大的變化，因此預測能力比較低。

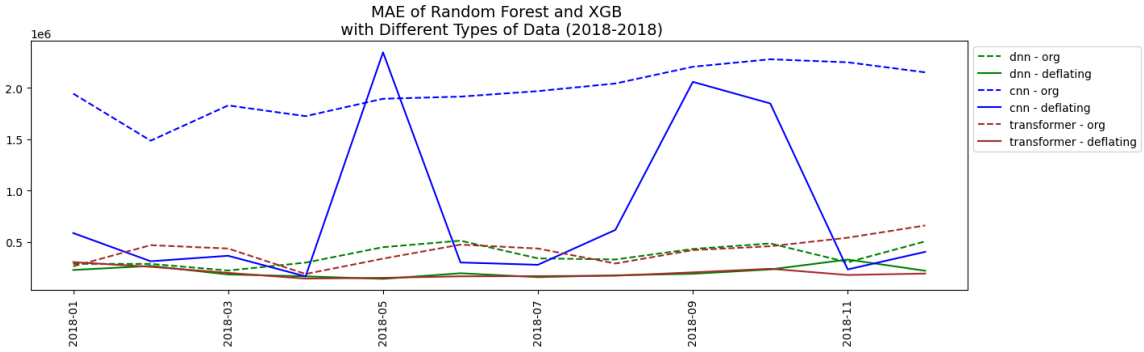
【圖 20、2018 RMSE】



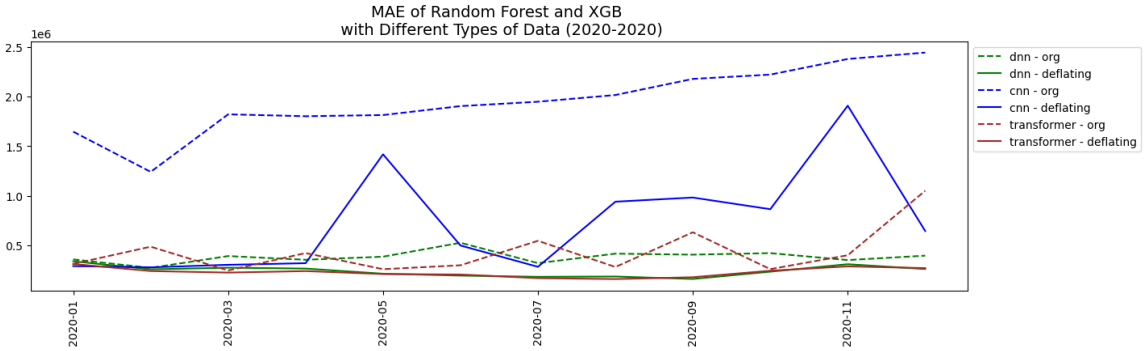
【圖 21、2020 RMSE】



【圖 22、2018 MAE】



【圖 23、2020 MAE】

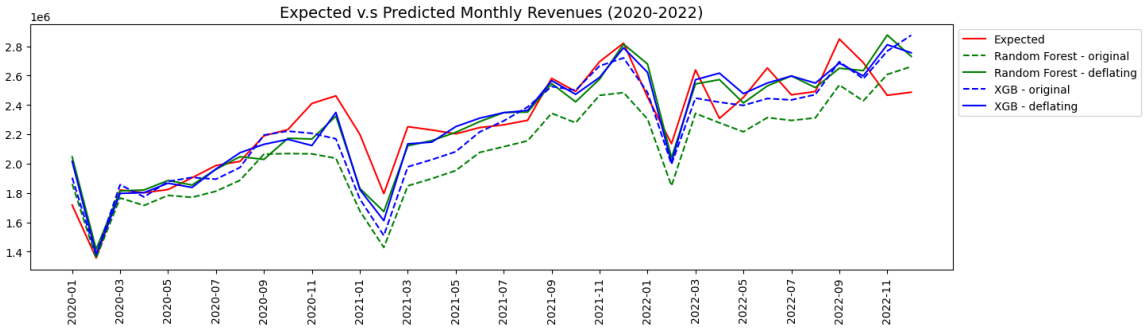


Q2-2 DNN、CNN、Transformer 和作業三 RandomForest 和 XGBoost 預測結果比較:

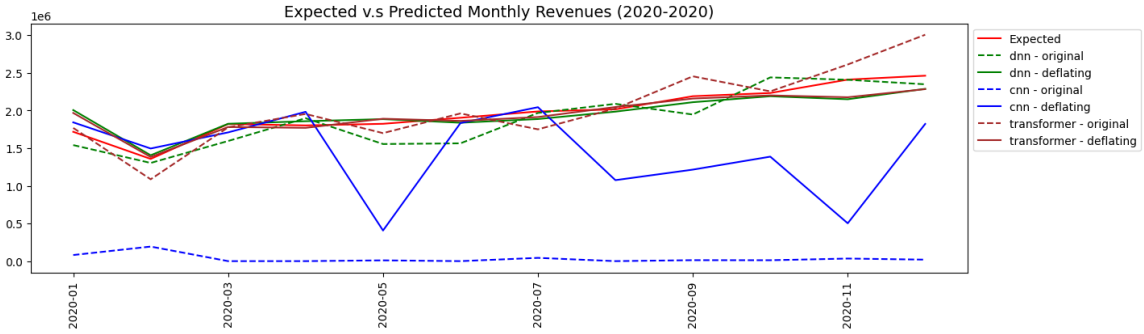
1. 相同:
    a. 進行資料平減後模型預測能力較高: 由【圖 24】~【圖 29】預測金額圖、RMSE、MAE 圖片比較可知，無論使用 XGBoost、 Random Forest 、DNN、CNN，或是 Transformer 模型進行預測，進行平減後的

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。

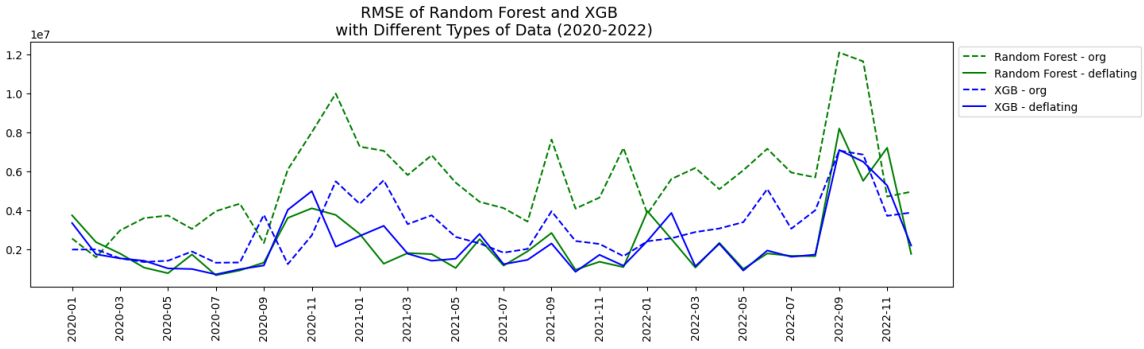
【圖 24、RandomForest、XGBoost 預測金額】



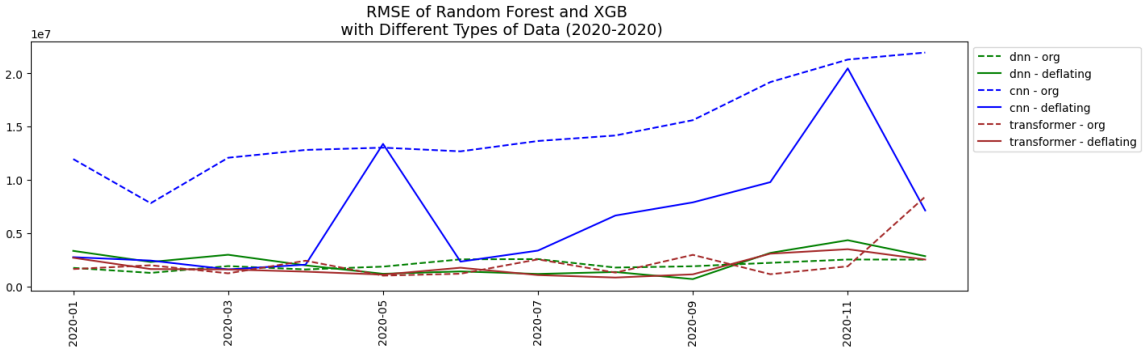
【圖 25、DNN、CNN、Transformer 預測金額】



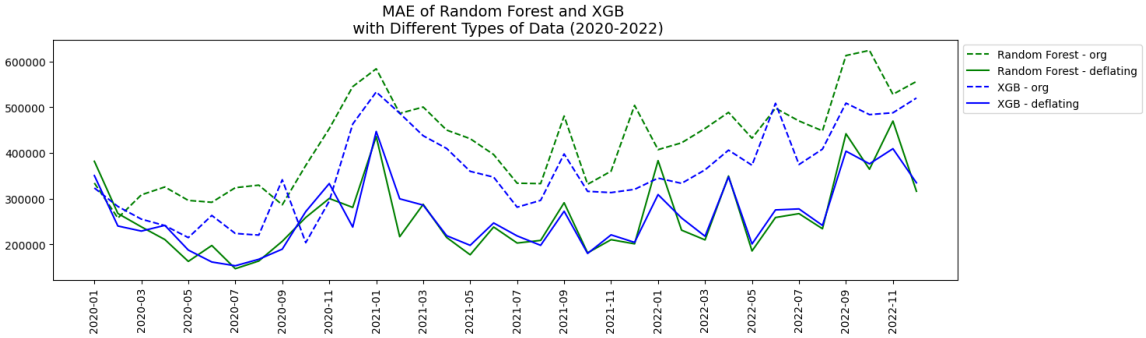
【圖 26、RandomForest、XGBoost RMSE】



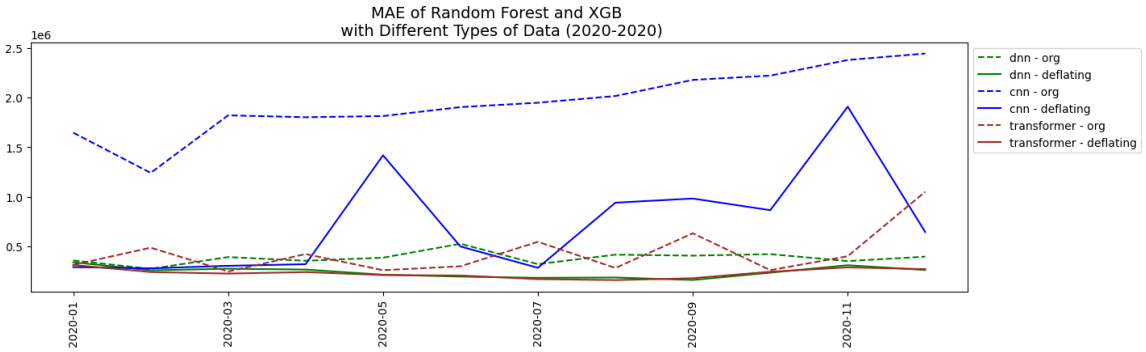
【圖 27、DNN、CNN、Transformer RMSE】



【圖 28、RandomForest、XGBoost MAE】

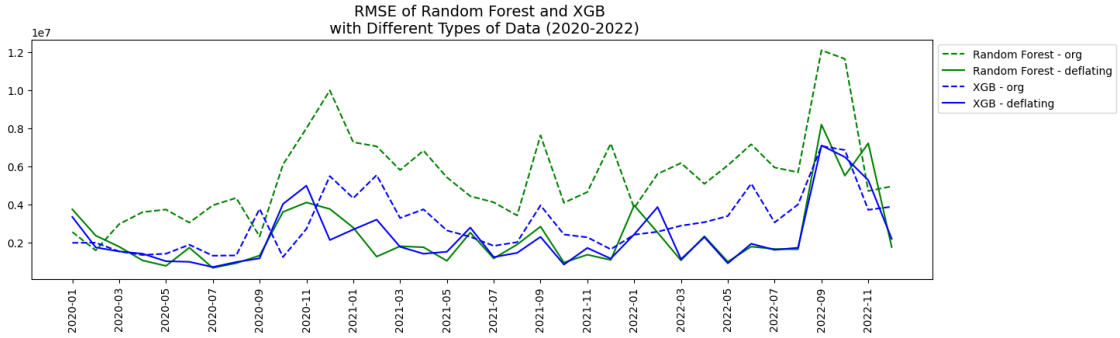


【圖 29、DNN、CNN、Transformer MAE】

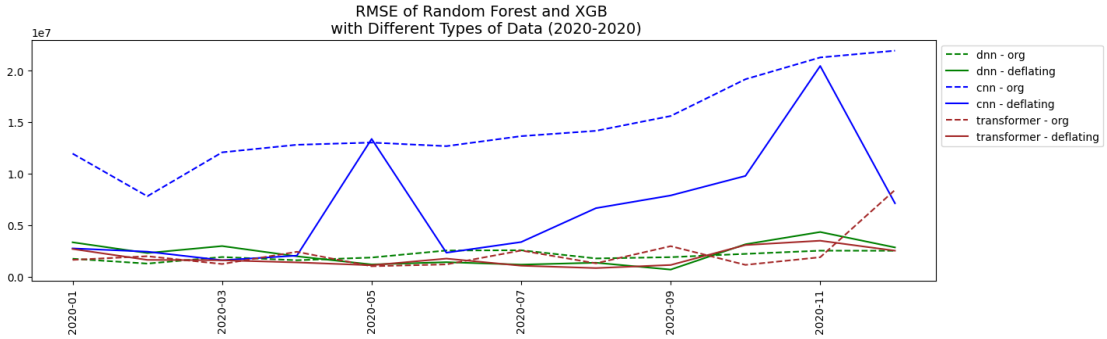


2. 相異:
    a. DNN、Transformer 準確度較高: 由【圖 30】~【圖 33】RMSE、MAE 圖片比較可知，無論資料是否經過平減，CNN、Transformer 表現大致比 RandomForest、XGBoost、CNN 還好。可能因為 DNN、Transformer 在進行數據預測的情境下較有優勢。

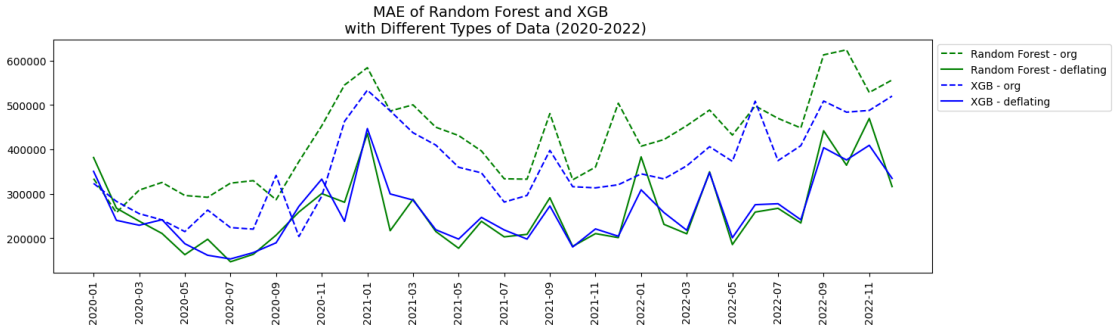
【圖 30、RandomForest、XGBoost RMSE】



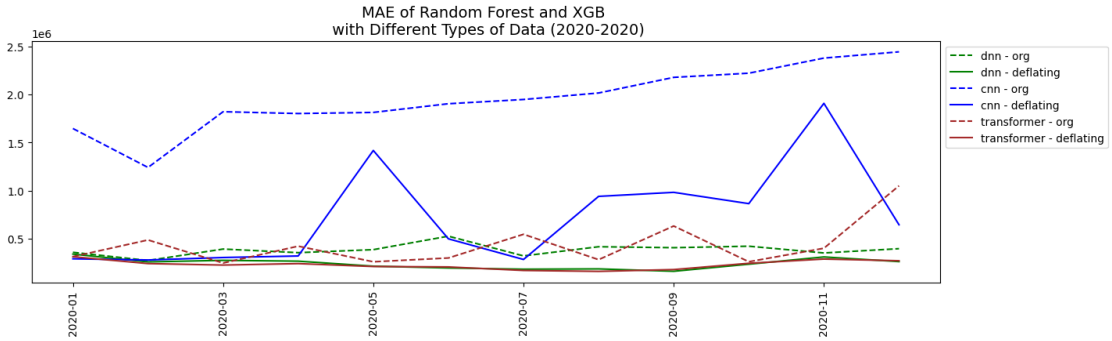
【圖 31、DNN、CNN、Transformer RMSE】



【圖 32、RandomForest、XGBoost MAE】



【圖 33、DNN、CNN、Transformer 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，說明資料集處理方式。

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。

3. 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。

4. 預測模型及衡量指標:

使用 CNN 、 DNN 及 Transformer 模型，並以 RMSE、MAE 作為模型好壞的衡量指標。

5. 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。

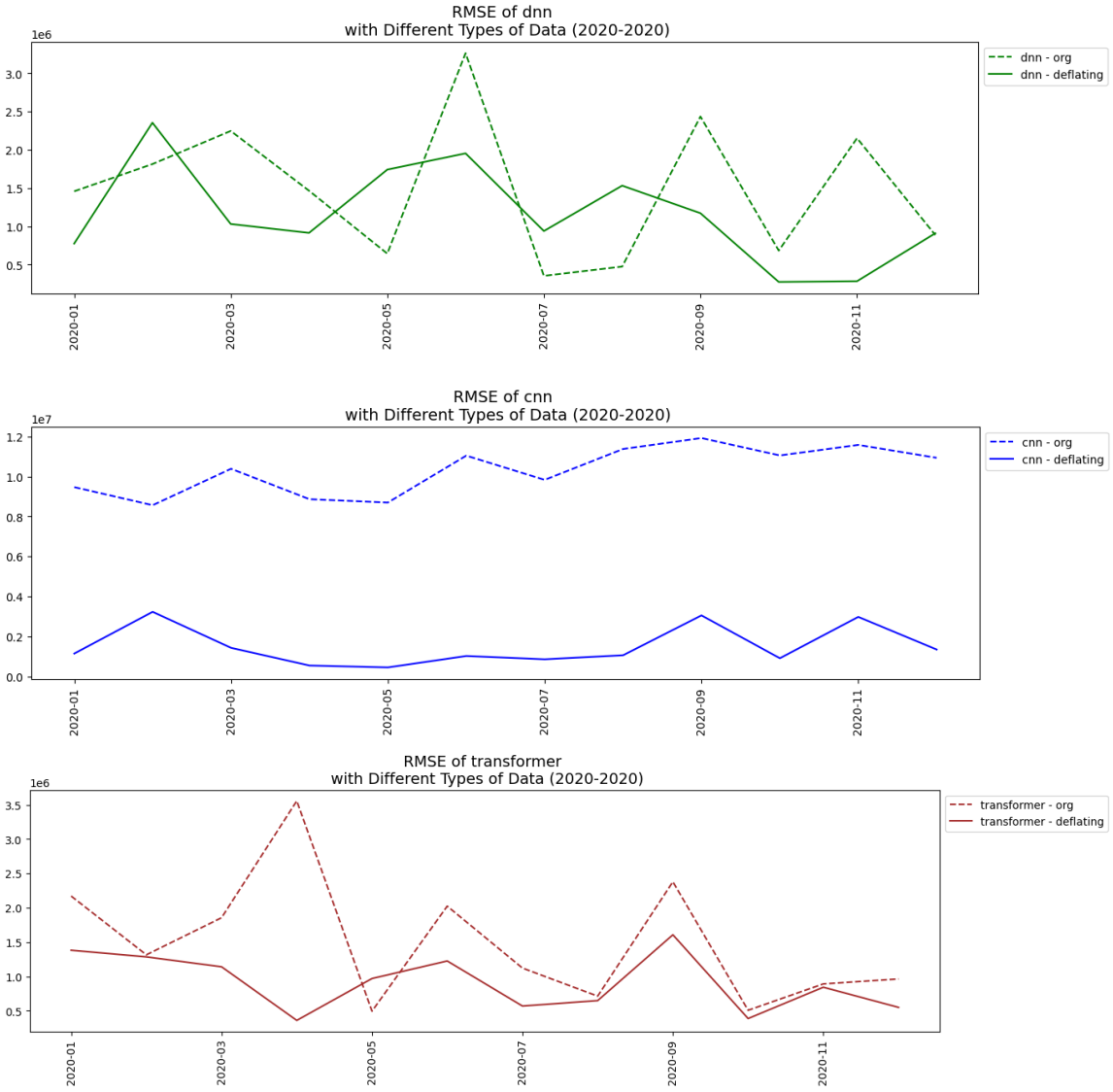
Q2: 分析 2020/1 至 2022/12 月營收金額的預測結果（須標明使用之模型和衡量指標

1.模型準確率預測結果:

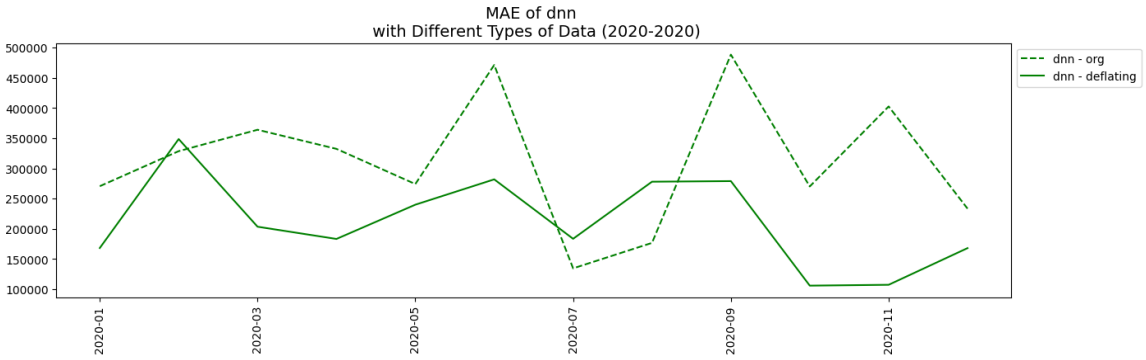
1.1 以 RMSE、MAE 作為衡量依據可得經過資料平減後模型表現較好:

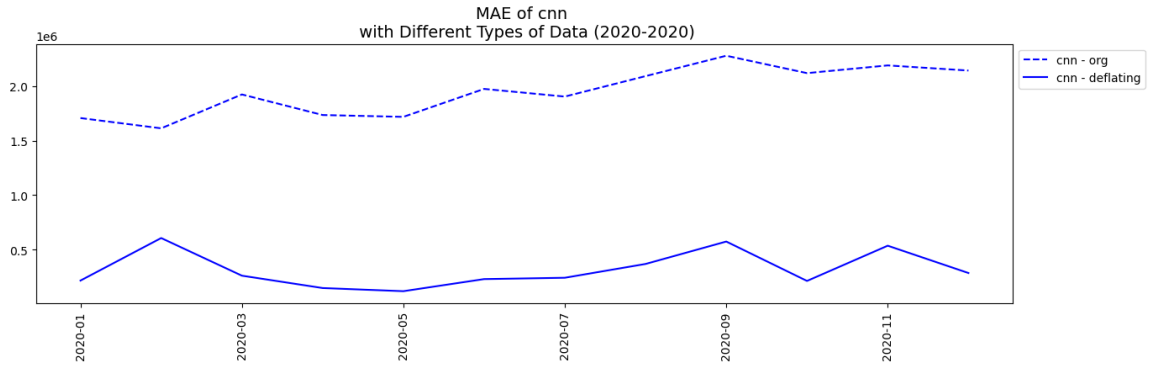
由【圖 34】、【圖 35】RMSE 及 MAE 圖形中可以發現資料經過平減後模型表現較佳。推測是因為平減過程可以去除數據當中的異常值，改善數據的質量，幫助模型更有效地進行學習和準確預測。

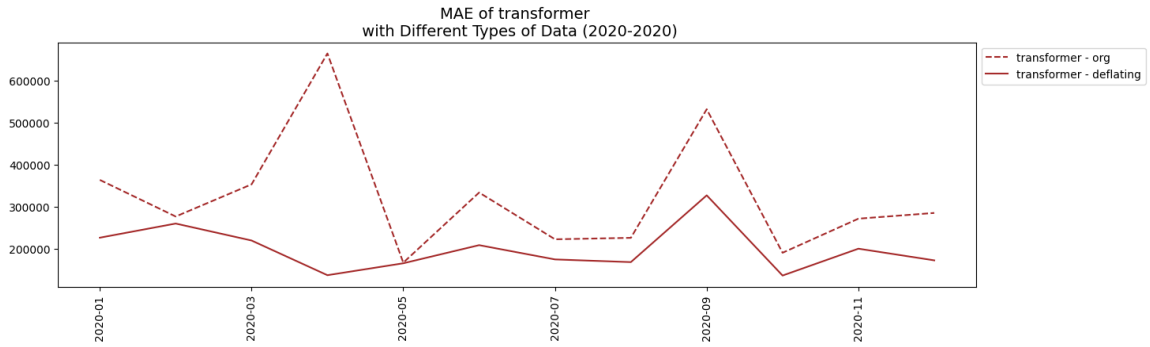
【圖 34、半導體產業 RMSE】



【圖 35、半導體產業 MAE】





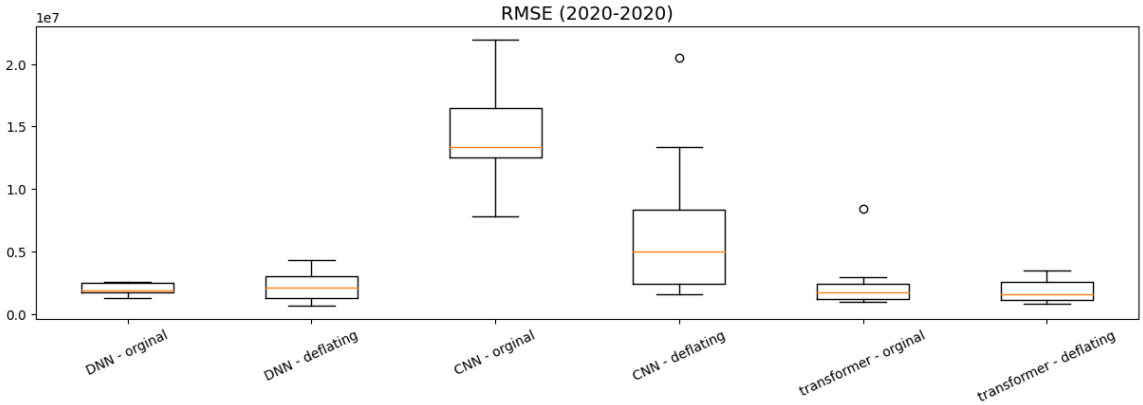


1.2 經資料平減的模型中，單一產業分析模型表現效能較好:

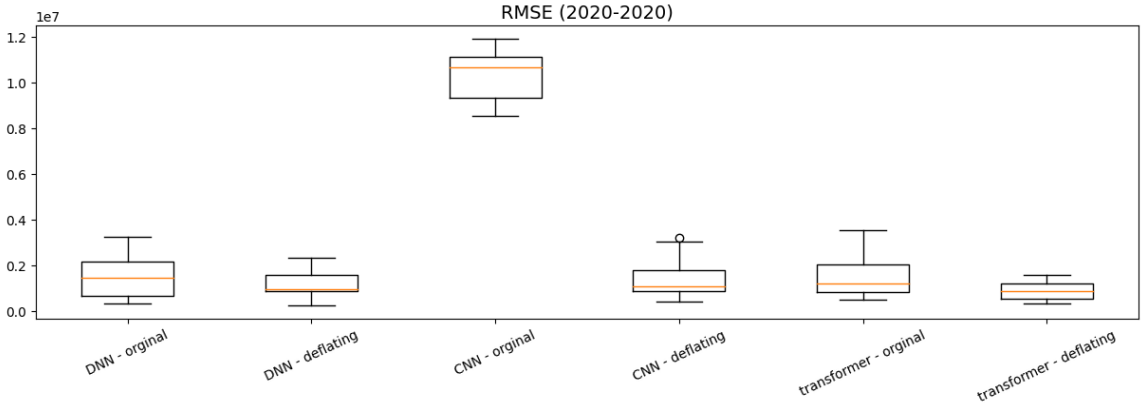
由【圖 36】~【圖 39】RMSE 及 MAE 圖形可看出，相比於全產業的營收預測，把產業限縮到半導體單一產業，平減模型的預測表現稍微較好，推測可能有以下幾個原因:

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。

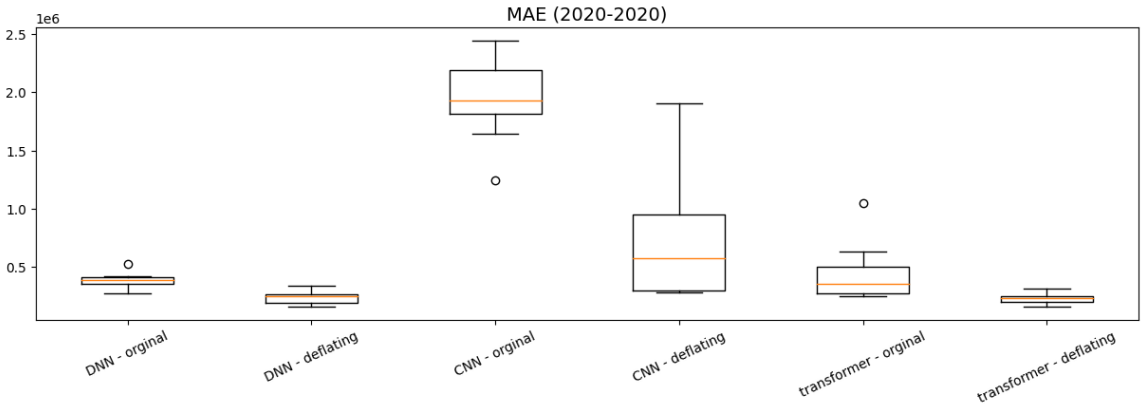
【圖 36、全產業各模型 RMSE】



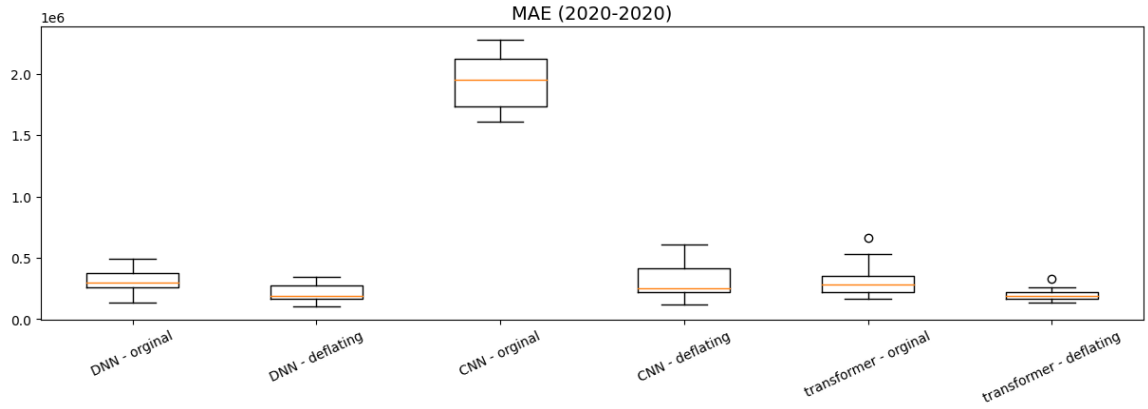
【圖 37、半導體產業各模型 RMSE】



【圖 38、全產業各模型 MAE】



【圖 39、半導體產業各模型 MAE】



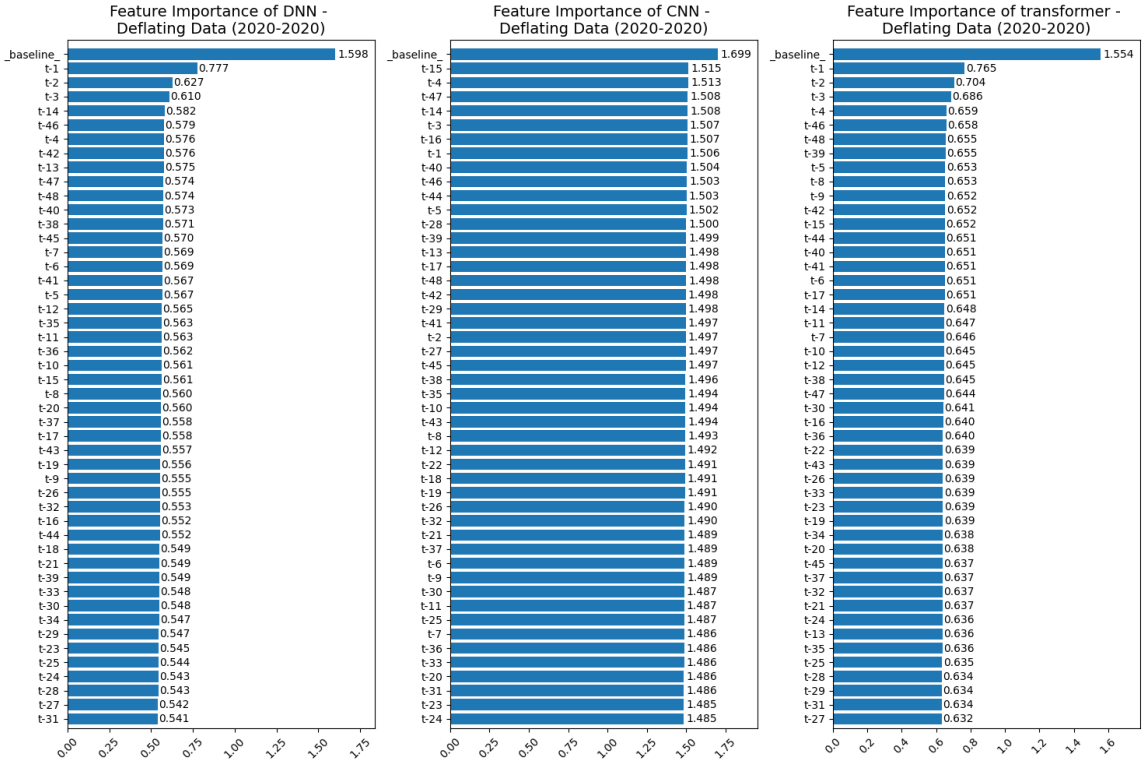
2.半導體產業最佳&最差模型特性及重要變數:

2.1 使用平減後資料進行預測，以及以 RMSE 作為衡量標準的最佳模型中， t-1、t-2、t-3、t-4 為重要變數:

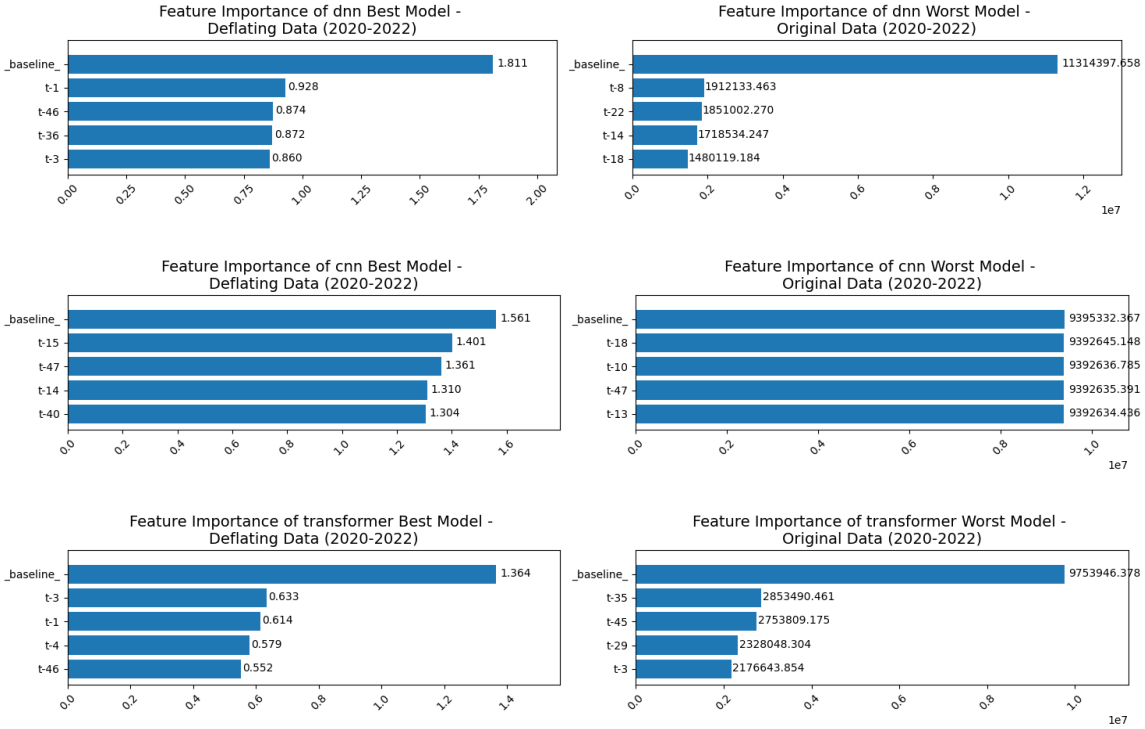
由【圖 40】、【圖 41】可以看出進行半導體產業營收預測時，重要變數包含 t-1、t-2、t-3、t-4 期資料，近期歷史資料重要性高，推測可能為以下原因:

- 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 40、使用平減後資料進行預測之各模型重要變數】



【圖 41、RMSE 標準下最佳&最差模型重要變數】



2.2 由【圖 42】、【圖 43】可知，使用 RMSE、MAE 作為衡量標準下，以 DNN、CNN、Transformer 進行半導體產業營收預測在 2020 年 4 月、6 月、9 月模型表現較差，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:

- 2020 年 4 月受全球疫情影響:
  - COVID 疫情大約在 2020 年 3 月開始在全球爆發，可能導致半導體業受到供應鏈中斷、生產停滯或需求下降等重大影響，而首當其衝應該是 4 月的營收波動，因此模型無法準確捕捉疫情對營收的影響。
  - 參考網址: https://zh.wikipedia.org/zh-tw/2019%E5%86%A0%E7%8B%80%E7%97%85%E6%AF%92%E7%97%85%E7%96%AB%E6%83%85%E6%99%82%E9%96%93%E8%BB%B8#4%E6%9C%88
- 2020 年 9 月 15 日美國華為禁令緩衝期的最後一天:
  - 美國進行的半導體制裁禁止向中國客戶提供尖端產品、設備或技術，提高了全球半導體產業的不確定性，因此模型較難準確預測半導體產業 9 月營收。
  - 參考網址: https://www.bbc.com/zhongwen/trad/business-54151243

【圖 42、RMSE 標準下 DNN、CNN、Transformer 最差模型】

modelName	dataType	scoreType	max_month	max_score
dnn	org	RMSE	2020-06	3264010.0

modelName	dataType	scoreType	max_month	max_score
cnn	org	RMSE	2020-09	11920471.0

modelName	dataType	scoreType	max_month	max_score
transformer	org	RMSE	2020-04	3555637.0

【圖 43、MAE 標準下 DNN、CNN、Transformer 最差模型】

modelName	dataType	scoreType	max_month	max_score
dnn	org	MAE	2020-09	488393.0

modelName	dataType	scoreType	max_month	max_score
cnn	org	MAE	2020-09	2276744.0

modelName	dataType	scoreType	max_month	max_score
transformer	org	MAE	2020-04	665506.0