# 期末專題書面報告

## 摘要

價格最佳化有助於企業找到有效的定價平衡點，實現盈利目標，同時也滿足客戶需求。若單價過高，容易會失去客戶，若是單價過低，則會導致收入損失。因此，本文使用幾種迴歸模型，透過數據來訂定和預測最適當的產品價格。

## 1. 數據集描述

I. 資料來源：https://www.kaggle.com/datasets/suddharshan/retail-price-optimization/data

II. 資料名稱：retail price 零售價

III. 資料年分：2017-2018

IV. 資料內容：共 30 個變數，676 筆資料。產品單價、產品的相關資料、運費、競爭對手的相關資料等

V. 變數名稱：

| | |
|---|---|
| product_id | 產品名稱 |
| product_category_name | 產品類別 |
| month_year | 日期(以月為單位) |
| qty | 銷售數量 |
| unit_price | 物品單價 |
| total_price(qty * unit_price) | 總金額 |
| freight_price | 產品敘述的長度 |
| product_name_lenght | 產品的照片數量 |
| product_photos_qty | 產品的重量 |
| product_weight_g | 用戶對產品的評分 |
| customers | 此產品的需求量 |
| weekday | 工作日天數 |
| weekend | 假日天數 |
| holiday | 節日天數 |
| volume | 產品的大小 |
| comp_1、2、3 | 其他 3 個競爭對手對此產品的定價 |
| ps1、2、3 | 其他 3 個競爭對手的用戶對此產品的評分 |
| fp1、2、3 | 其他 3 個競爭對手的產品運費 |
| lag_price | 上個月的單價 |

## 2. 資料處理

I. 數據集內無缺失值。

II. 去除多餘變數：month_year、product_id、s 等等

III. 處理競爭對手資料。若只單獨看一個公司沒辦法瞭解市場的運作，因此將其他三個競爭對手的產品單價、運費、顧客評分取平均，做為新變數：
comp_mean：將其他三個競爭對手的產品定價取平均
ps_mean：將其他三個競爭對手的顧客對此產品的評分取平均
fp_mean：將其他三個競爭對手的運費取平均

IV. 經處理後，剩餘 21 個變數

## 3. 探索式資料分析

I. Unit_price。如圖 1
集中在 20~150

II. Unit_price vs. lag_pric。如圖 2
有非常明顯的線性正相關，不過相關係數過大，後面做模型時需考慮是否加入。產品的單價不可能每個月做調整，所以大多產品的價格可能會維持和上個月一樣。

III. Unit_price vs. comp_mean。如圖 3
圖中可觀察到有一些正相關。

IV. Unit_price vs. freight_price。如圖 4

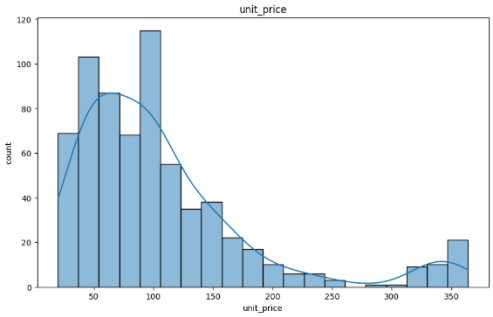



圖 1

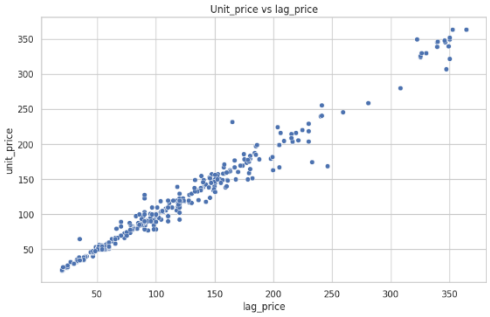



圖 2


圖 3


圖 4

I. Unit_price vs. year。如圖 5
圖中可觀察到年份和單價似乎無太大關係。

II. Unit_price vs. month。如圖 6
圖中可觀察到月份和單價似乎無太大關係。

III. Unit_price vs. product_category_name。如圖 7
圖中可觀察到產品類別和單價可能有些關係。

IV. Unit_price vs. product_score。如圖 8
圖中可觀察到顧客評分和單價可能有些關係。

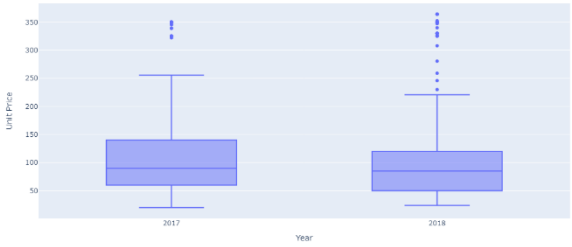



圖 5

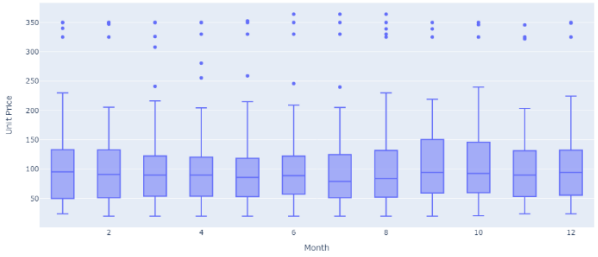



圖 6

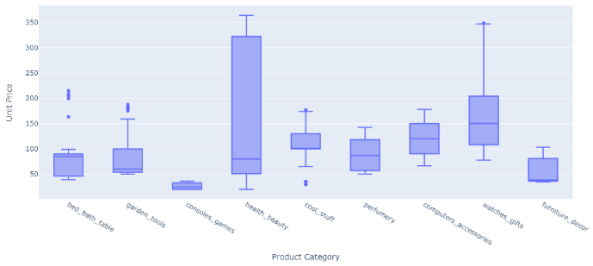



圖 7

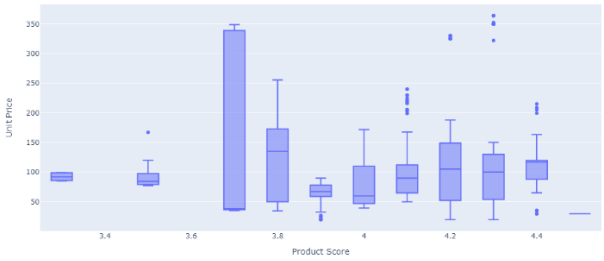



圖 8

I. 與其他對手單價比較。如圖 9
圖中可觀察到四家公司的單價有些微差異。

II. 與其他對手運費比較。如圖 10
圖中可觀察到四家公司的運費並沒有明顯差異。

III. 與其他對手評分比較。如圖 11
圖中可觀察到四家公司的顧客評分有明顯差異。

IV. 變數的熱力圖。如圖 12

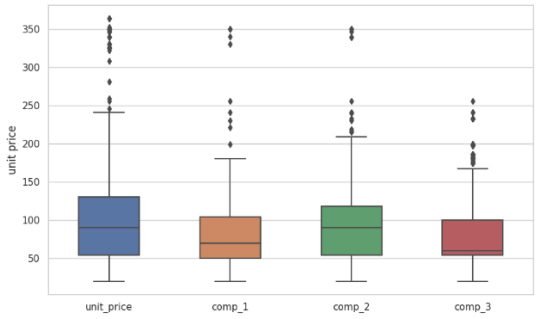



圖 9

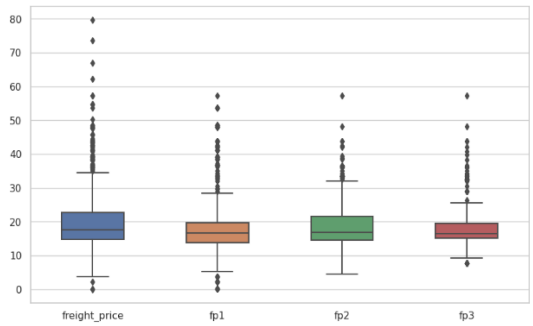



圖 10

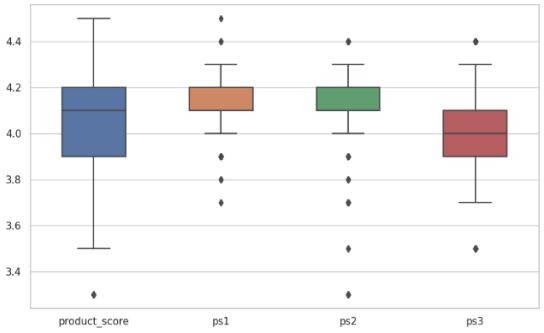



圖 11

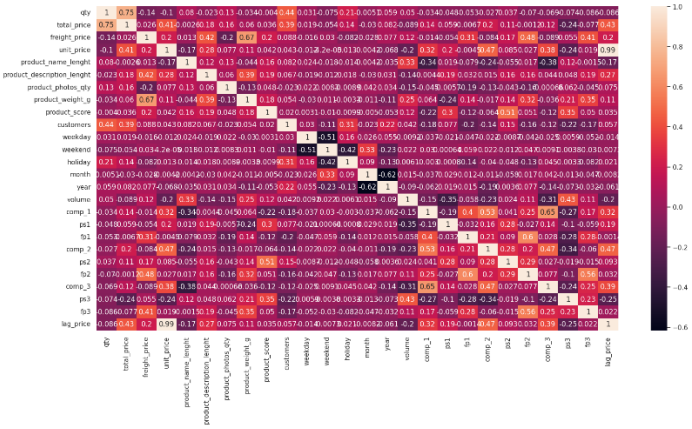
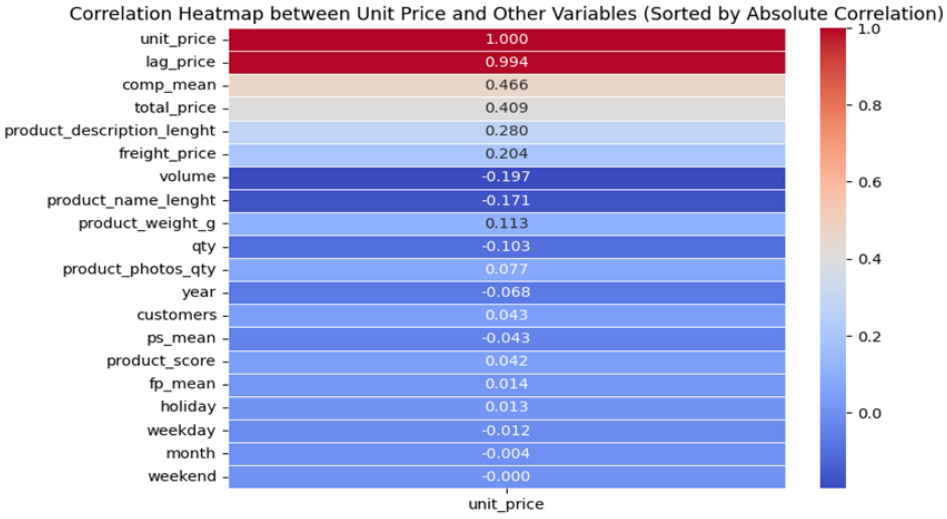


圖 12

## 4. 變數篩選

利用響應變數 unit_price 與其他變數的相關性來挑選變數，我們會分成兩種取法:第一種取法是有 lag price 變數的那這邊我們會取相關性大於 0.2 的變數包括 volume；第二種取法是沒有 lag price 變數那我們會取相關性大於 0.1 的變數。



## 5. 模型篩選

我們透過兩種不同的變數選取來分別帶入模型來看模型，那只比較同一種取法下不同模型的評分，這邊用的評分標準是用 Test MSE，而過程會利用交叉驗證去初步的篩選超參數。

I. 取法一(有 lag price):

| | 變數名稱 |
|---|---|
| unit_price | 產品單價 |
| lag_price | 上個月單價 |
| comp_mean | 競爭對手的產品定價 |
| total_price | 總金額 |
| product_description_lenght | 描述產品的長度 |
| freight_price | 運費 |
| volume | 產品體積 |

檢查六個變數是否有共線性問題:

| | Feature | VIF |
|---|---|---|
| 0 | lag_price | 4.957114 |
| 1 | comp_mean | 4.464615 |
| 2 | total_price | 2.088379 |
| 3 | product_description_lenght | 3.123164 |
| 4 | freight_price | 5.602019 |
| 5 | volume | 2.156670 |

6 個模型的比較:

| model | MSE |
|---|---|
| linear regression | 78.99773486466738 |
| ridge regression | 78.99556743386326 |
| lasso regression | 78.9563646613325 |
| Decision Tree | 77.04369441236186 |
| Randomforest | 75.98905964434049 |
| XGBoost | 68.92049967658838 |

可以看出雖然 XGBoost 的模型 MSE 是最小的，但是其實每個模型之間差距都不是很大，這邊我們推測是放入 lag price 這個變數的關係，因為這個變數跟響應變數相關性太高，導致不管什麼模型預測都會很好。

II. 取法二(沒有 lag price):

| | 變數名稱 |
|---|---|
| unit_price | 產品單價 |
| product_name_lenght | 產品名稱長度 |
| comp_mean | 競爭對手的產品定價 |
| total_price | 總金額 |
| product_description_lenght | 描述產品的長度 |
| freight_price | 運費 |
| volume | 產品體積 |
| product_weight_g | 產品重量 |
| qty | 銷售數量 |

檢查八個變數是否有共線性問題:

| | Feature | VIF |
|---|---|---|
| 0 | comp_mean | 4.443830 |
| 1 | total_price | 4.815511 |
| 2 | product_description_lenght | 3.541274 |
| 3 | freight_price | 9.434446 |
| 4 | volume | 3.367886 |
| 5 | product_name_lenght | 12.641648 |
| 6 | product_weight_g | 3.536406 |
| 7 | qty | 5.031389 |

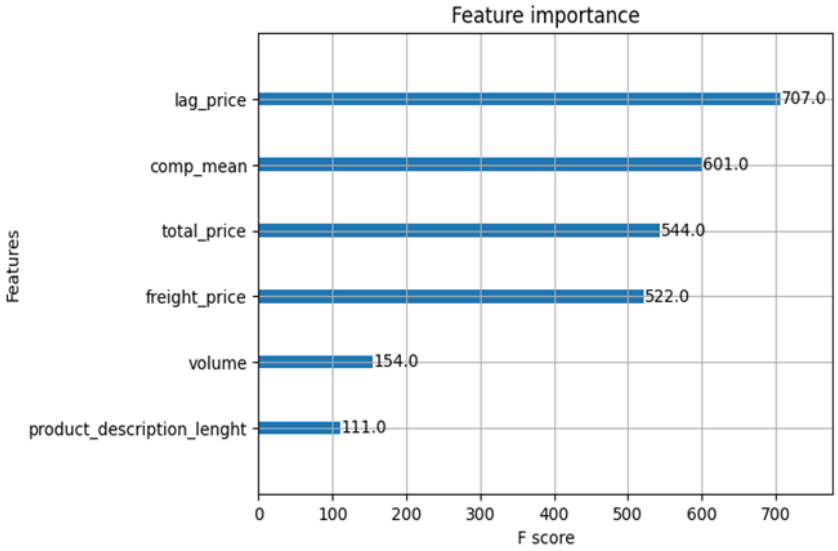
6 個模型的比較:

| model | MSE |
|---|---|
| linear regression | 1945.5833366877976 |
| ridge regression | 1945.4686283447945 |
| lasso regression | 1945.3254837138495 |
| Decision Tree | 1182.5794313585207 |
| Randomforest | 298.5825480408068 |
| XGBoost | 160.82715994659713 |

可以發現 lag price 拿掉之後模型之間就有明顯的差異了，尤其是兩個集成學習的模型的表現相對其他模型好很多，這是比較合理的，但模型最好的還是 XGboost。

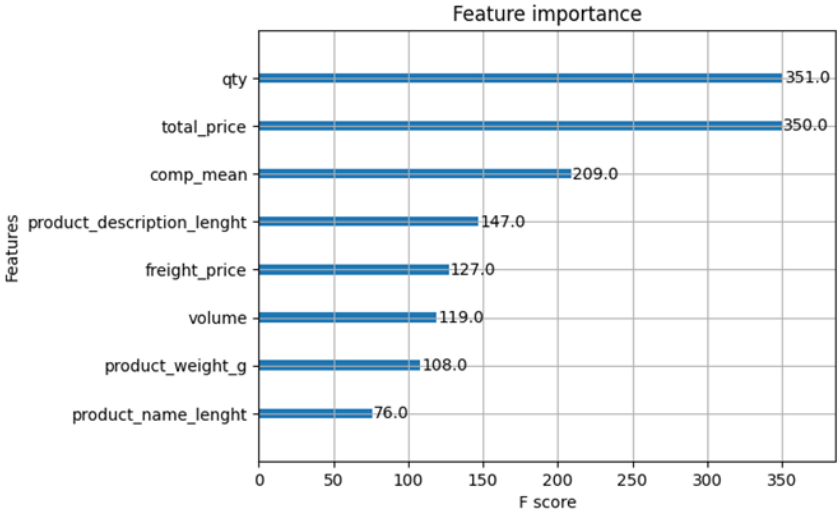
III. 比較兩種變數取法 XGboost 的 feature importance:

A. 有 lag price:




跟我們想的一樣，lag_price 的相關性太高導致模型會根據這個變數會被模型判斷成最重要的變數，但以實際上來看用 lag price 來預測不太合理

B. 沒有 lag price:




可以看到產品數量、總金額跟競爭對手的平均定價對於模型來說是最重要的，如果未來想要制定方案可以去利用這幾個變數去做決策。

## 6. 結論

1. 以預測單價而言，無論是有 lag price 或沒有 lag price 的模型，XGboost 的模型 MSE 皆是最小的，也就是最好的。

2. 比較兩種取法(有無 lag price)，雖然有 lag price 的每個模型預測出來結果都不差，但是因為 lag price 這個變數只是將月份的資料作平移而已，所以用這個變數來預測會很不合理。

3. 如果未來要繼續分析這個模型我們會使用沒有 lag price 的取法，可以透過更多更進階的模型來讓模型預測更準確，也可以去研究對模型來說那些重要變數，可以對這個領域會有更好的理解。

4. 模型覺得最重要的變數為 qty、total_price，但這兩個變數大家都知道其重要性，所以在做價格預測時，這類「直接構成價格的變數」可能會掩蓋其他潛在的影響因子。未來，或許可以控制這些變數，以探討其他因素對價格的額外貢獻。

Team Contributions:
吳俞憲:負責整體流程、書面報告後半部分
張家輔: 提供意見增加細節、做簡報上台報告、書面報告前半部分