

# 資料分析與學習基石

## Store Sales - Time Series Forecasting

Use machine learning to predict grocery sales

第 11 組

何子安 E44065020 | 蔡東霖 F74071166 | 林千祺 N46084036 | 吳定洋 F74076213

# Outline

- 資料前處理
- 模型介紹
- 改進方法
- 結果與討論
- 結論
- 未來工作
- 參考資料

# 資料前處理

- Train.csv (資料筆數 3000888 筆，沒有缺失值)

| Column      | Range/Unique  | Description    |
|-------------|---|----------------|
| id          | 3000888   | 索引值            |
| date        | 2013.01.01~2017.08.15   | 日期             |
| store_nbr   | 54 間  | 商店編號           |
| family      | 33 種產品  | 產品類別           |
| sales       | 124717  | 該產品於該商店當日的總銷售額 |
| onpromotion | 741   | 該產品於該商店當日促銷的數量 |
| 備註          | 發薪日為每月15日及月底； 2016.04.16 厄瓜多大地震，地震發生後數週，人們齊心協力捐贈水和其他急需物資，極大地影響了超市的銷售。 |                |

# 資料前處理-資料挑選

- 挑選 [2017/4/30 - 2017/8/15] 之間的數值
- 平均油價、油價 lag
- Holiday
- School session
- Blending

# 模型介紹 – Base Case

'SCHOOL AND OFFICE SUPPLIES'

r1 = ExtraTreesRegressor

r2 = RandomForestRegressor

b1 = BaggingRegressor(base\_estimator=r1, n\_estimators=10)

b2 = BaggingRegressor(base\_estimator=r2, n\_estimators=10)

model = VotingRegressor([('et', b1), ('rf', b2)])

else:

Ridge

SVR

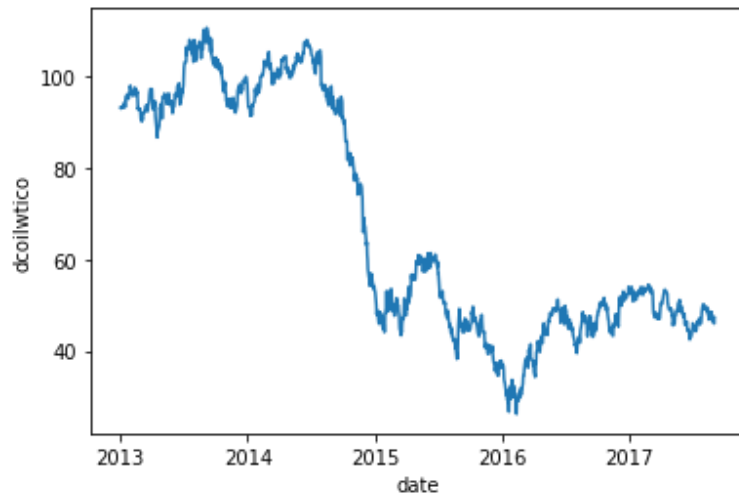
VotingRegressor([('ridge', ridge), ('svr', svr)])

# 模型介紹 – Case 1

Base Case 使用了 train.csv 中 [2017/4/30 - 2017/8/15] 的資料，

我們嘗試將時間區段拉長，並且加入 seasonal 的週期作為 feature，

但預測結果並沒有提升。



# 模型介紹 – Case 2

Base Case 的 holiday\_event.csv ,

只有使用 National 的 holiday 加入 calendar 做為 feature ,

我們嘗試將 regional 和 local 也加入 calendar 做為訓練模型的 feature ,


預測結果也**沒有提升**。


# 模型介紹 – Case 3

Base Case 中的 Model 有部分使用了 RandomForestRegressor，  
我們經過多次嘗試不同的組合後發現使用 XGBRegressor 時，  
預測結果有所提升。



# 結果與討論

|    |          |   |         |   |    |
|----|----------|---|---------|---|----|
| 21 | HoeZiOnn |  | 0.40334 | 6 | 2m |
|----|----------|---|---------|---|----|

 Your Best Entry!  
Your most recent submission scored 0.40334, which is the same as your previous score. Keep trying!

- 在最終的 Final Case 中得出了
  - 分數為 0.40334 的最佳結果
  - 位於 Public Leaderboard 上第 21 名
- 期間最差成績
  - 分數為 2.03583
- 並針對此結果的改進進行分析與檢討

# 結論

# 未來工作

利用 GridSearch 找出最佳的  
模型組合

篩選出更具影響力的特徵項

嘗試不同 Blending 的作法

# 參考資料

- [Store Sales TS Forecasting - A Comprehensive Guide | Kaggle](#)
- [Store Sales - Time Series Forecasting - Hyperparameters | Kaggle](#)
- <https://wizardforcel.gitbooks.io/ntu-hsuantienlin-ml/content/25.html>

Thank you for listening

Q&A