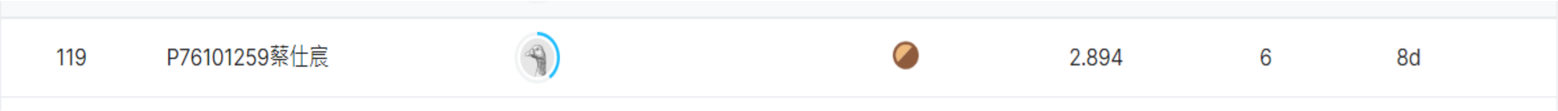
# DSAI Final Project Result Presentation

Member: 蔡仕宸-P76101259、鄭力維-NE6081080

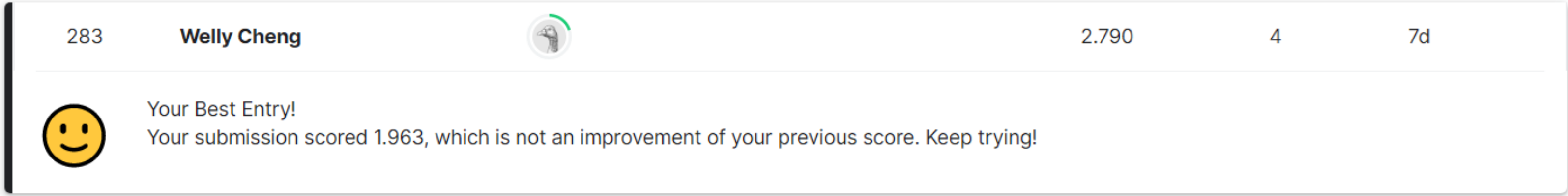
# Rank 2022/6/7

## Model without categorical feature



| | | | | |
|---|---|---|---|---|
| 119 | P76101259蔡仕宸 | 2.894 | 6 | 8d |

## Model without categorical feature



| | | | | |
|---|---|---|---|---|
| 283 | **Welly Cheng** | 2.790 | 4 | 7d |

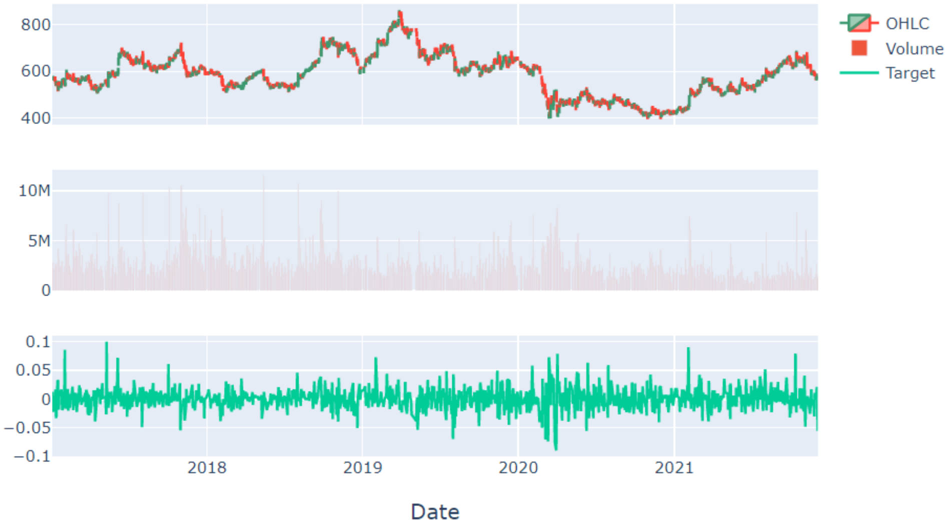
Your Best Entry!
Your submission scored 1.963, which is not an improvement of your previous score. Keep trying!

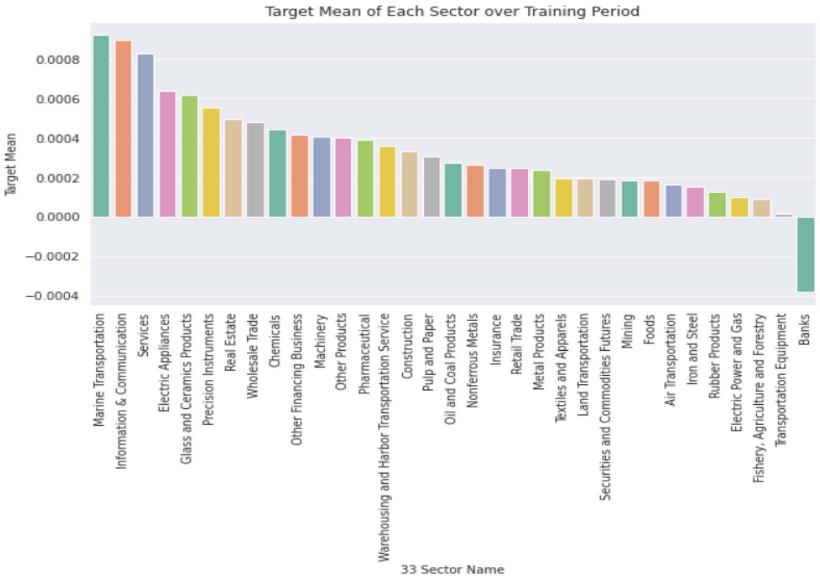
# 競賽目標

本次比賽由Japan Exchange Group, Inc. (JPX) 主辦，JPX是一家控股公司，經營著世界上最大的證券交易所之一、東京證券交易所 (TSE) 以及大阪交易所 (OSE) 和東京商品交易所 (TOCOM)。

比賽將涉及從符合預測條件的股票（約 2,000 隻股票）中建立投資組合。並對股票進行排名，最後選擇前 200 隻股票和後 200 隻股票的投資回報進行評估。

# Model - Feature





# Model - Feature

avg_price : (open+high+low+close)/4  => 當日開、收、高、低price平均

vol_amount = feature_avg_price*volume => 成交值

BOP: (open-close)/(high-low) => K棒實體棒的比例

wp : (open+high+low)/3 => 開、高、低price平均

TR : (high-low) => K棒距離

# Model - Feature

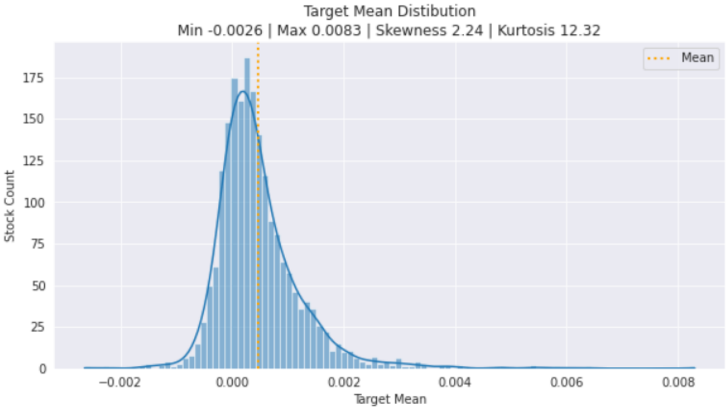
OC :  open*close => 開、收乘積

HL : high*low => 高、低乘積

logC: log(close+1) => 收取log

OHLCskew : skew(open,high,low,close) => 開、收、高、低price偏度

OHLCkur :kurtosis(open,high,low,close) => 開、收、高、低price峰度



# Model - Feature

Cpos : [ (close-low)/(high-low)]-0.5 => 收最低-0.5，收最高0.5

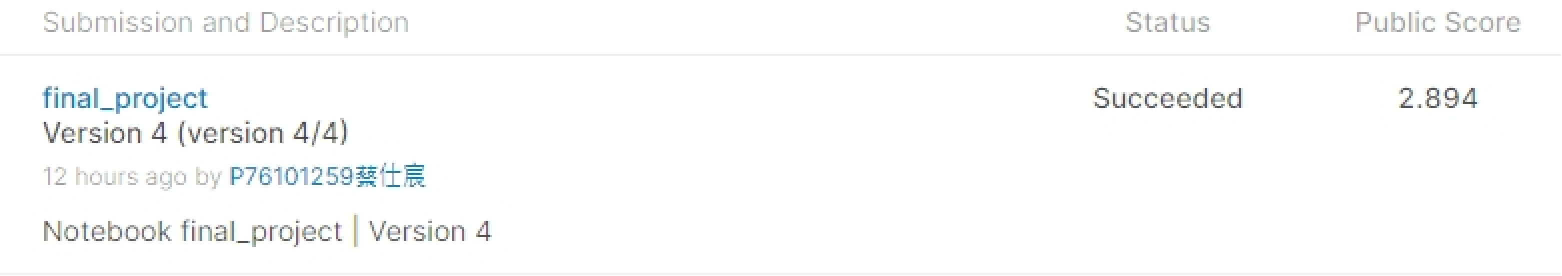
bsforce : feature_Cpos*volume => 上一項加入量

Opos: [(open-low)/(high-low)]-0.5 => 開最低-0.5，開最高0.5

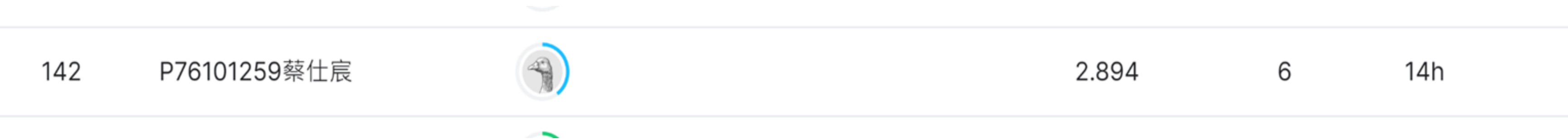
5_10_20_long = (close-feature_ro5>close-feature_ro10) & ( close-feature_ro10>close-feature_ro20) => 5日、週、月線多頭排列

5_10_20_short = (close-feature_ro5<close-feature_ro10) & ( close-feature_ro10<close-feature_ro20) => 5日、週、月線空頭排列

# Rank - Model1

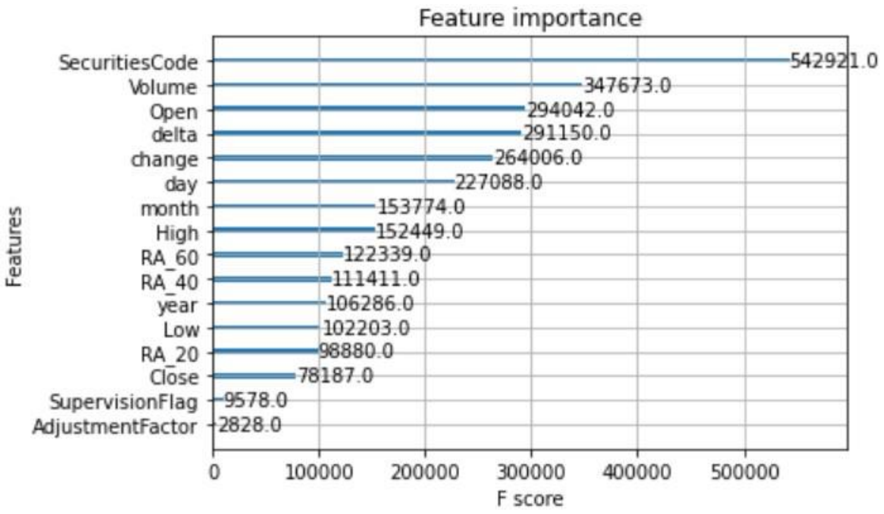


| Submission and Description | Status | Public Score |
|---|---|---|
| final_project<br>Version 4 (version 4/4)<br>12 hours ago by P76101259蔡仕宸<br>Notebook final_project \| Version 4 | Succeeded | 2.894 |

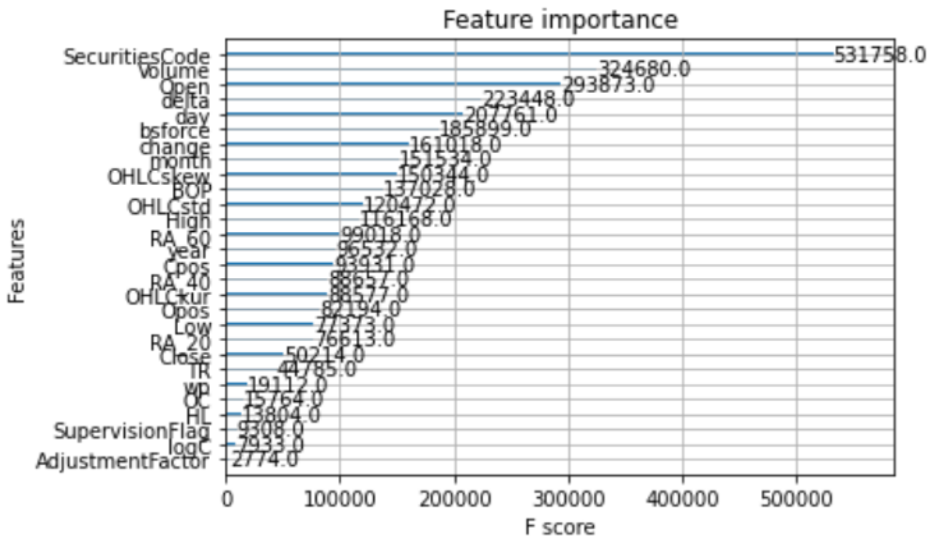


| | | | | |
|---|---|---|---|---|
| 142 | P76101259蔡仕宸 | 2.894 | 6 | 14h |

# Result Observation

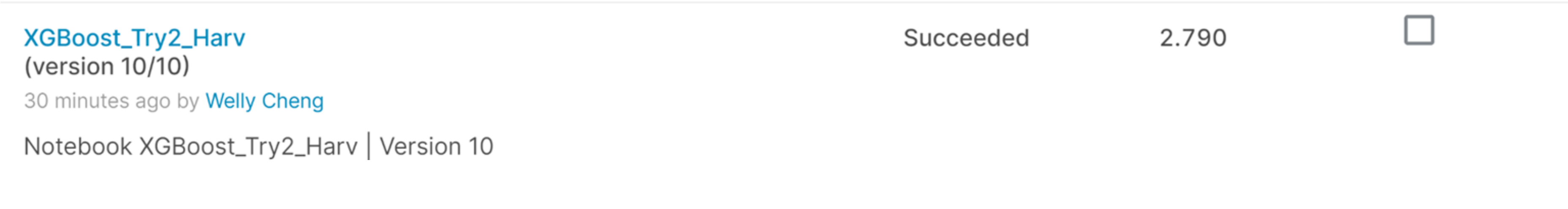


Benchmark



Our model1

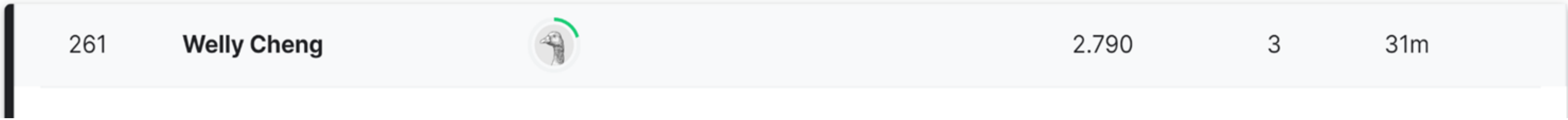
# Rank - Model2



XGBoost_Try2_Harv (version 10/10) Succeeded 2.790 □

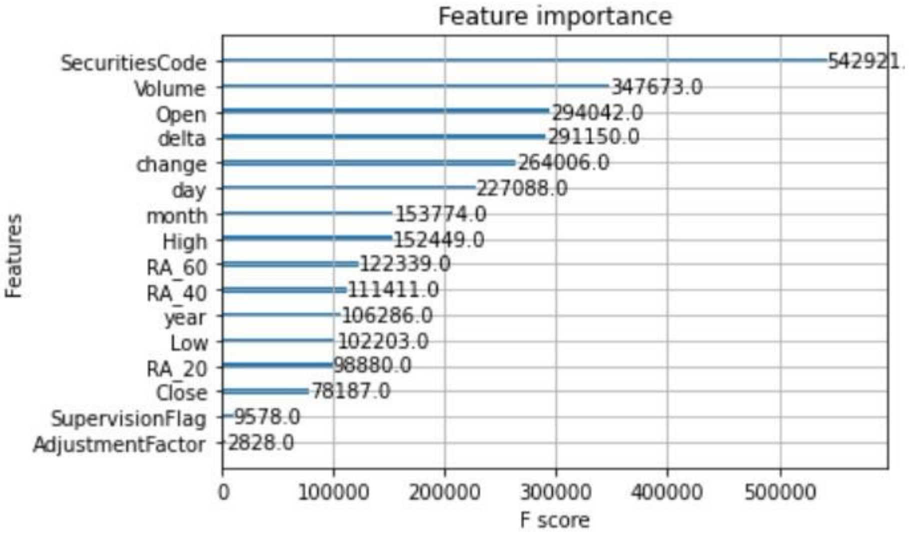
30 minutes ago by Welly Cheng

Notebook XGBoost_Try2_Harv | Version 10

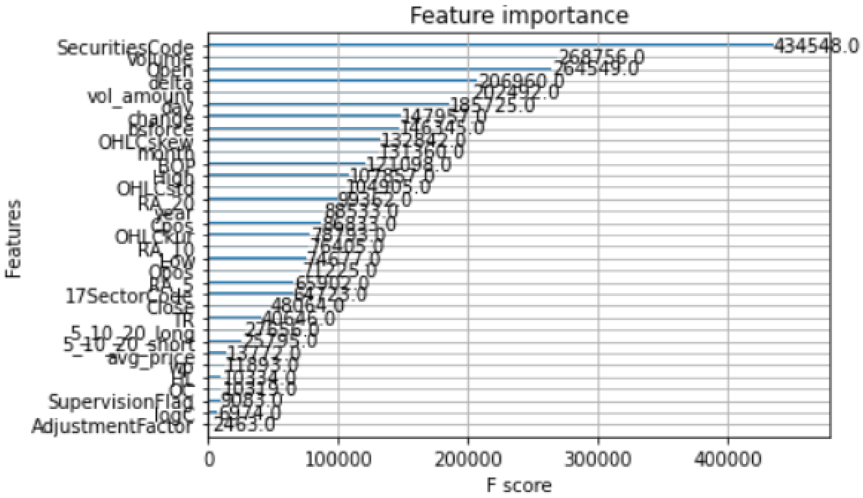


261 Welly Cheng 2.790 3 31m

# Result Observation

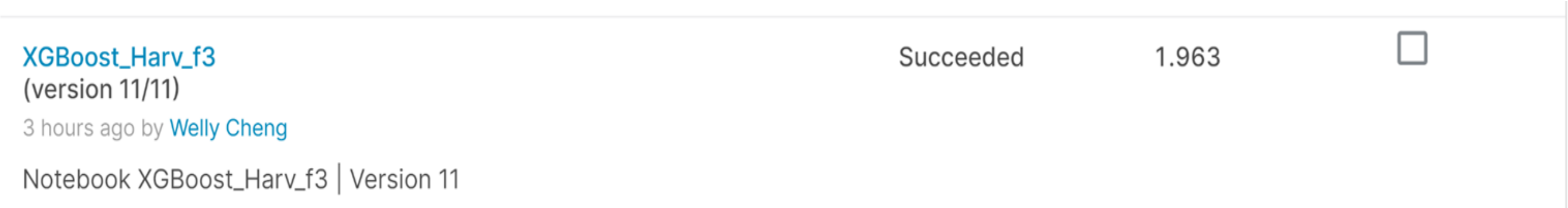


Benchmark



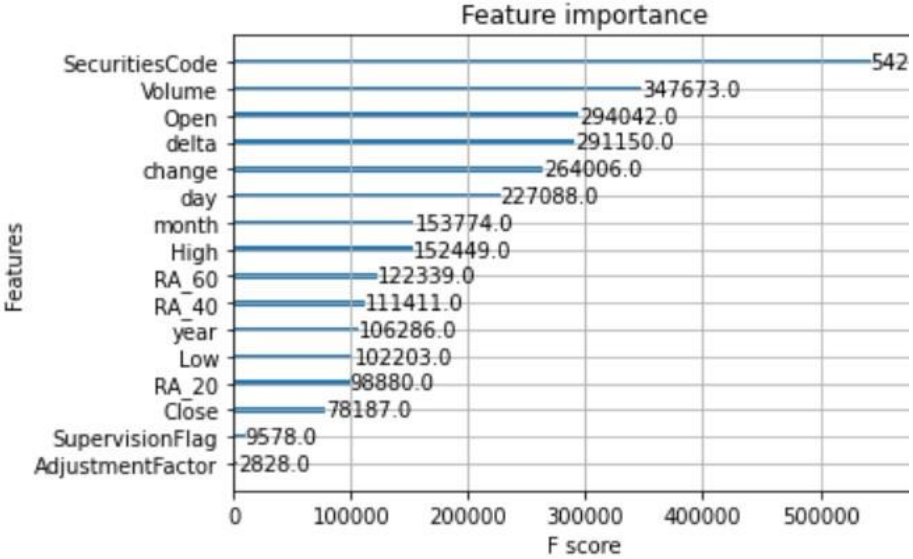
Our model2

# Rank - Model3

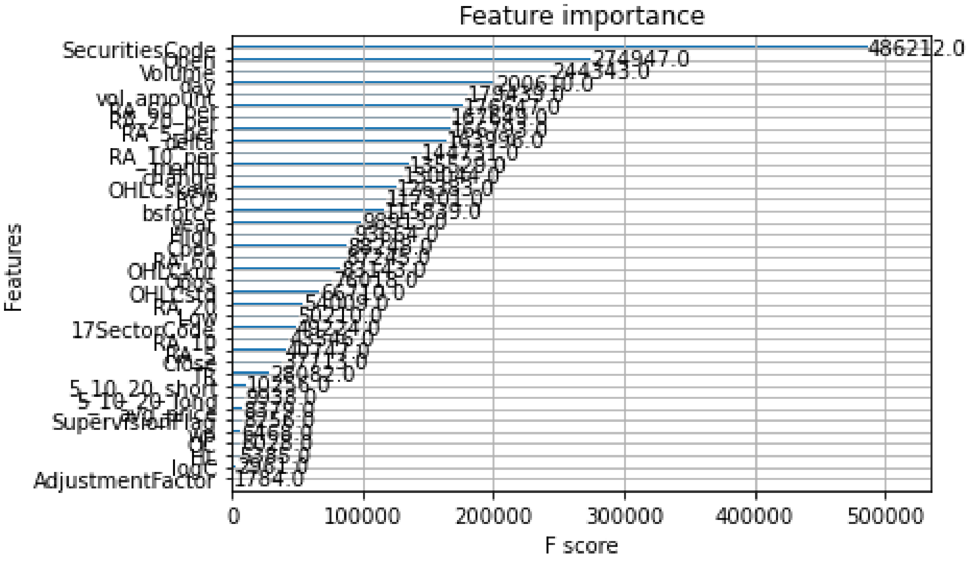


|  |  |  |  |
|---|---|---|---|
| XGBoost_Harv_f3<br>(version 11/11)<br>3 hours ago by Welly Cheng<br>Notebook XGBoost_Harv_f3 \| Version 11 | Succeeded | 1.963 | □ |

# Result Observation



Benchmark



Our model3

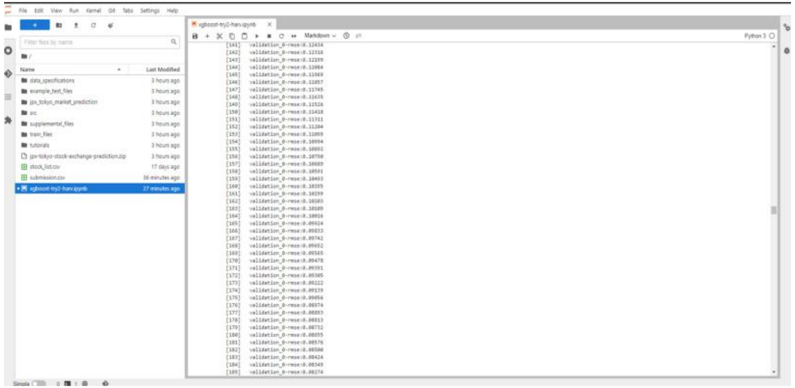
# Try and Error - Feature

How we deal with categories:

- Drop Categorical Feature
- Ordinal Encoding
- One Hot Encoding: Too many columns=> Can’t run in Kaggle

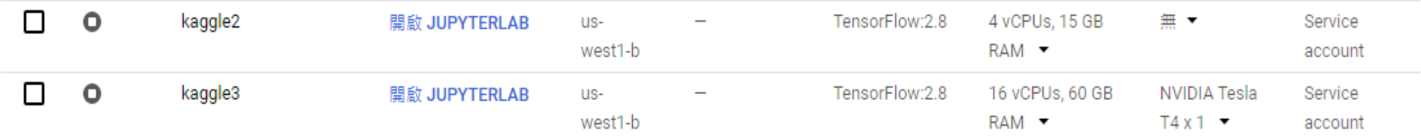
# Try and Error - Kaggle Environment

- Scoring - File submission/Code notebook
- Hardware limit
- Upgrade to Google Cloud AI Notebook



| Name | Last Modified |
|---|---|
| data_specifications | 3 hours ago |
| example_test_files | 3 hours ago |
| jpx_tokyo_market_prediction | 3 hours ago |
| src | 3 hours ago |
| supplemental_files | 3 hours ago |
| train_files | 3 hours ago |
| tutorials | 3 hours ago |
| jpx-tokyo-stock-exchange-prediction.zip | 3 hours ago |
| stock_list.csv | 17 days ago |
| submission.csv | 36 minutes ago |
| xgboost-try2-harv.ipynb | 27 minutes ago |

```
[141]	validation_0-rmse:0.12434
[142]	validation_0-rmse:0.12318
[143]	validation_0-rmse:0.12199
[144]	validation_0-rmse:0.12084
[145]	validation_0-rmse:0.11969
[146]	validation_0-rmse:0.11857
[147]	validation_0-rmse:0.11745
[148]	validation_0-rmse:0.11635
[149]	validation_0-rmse:0.11526
[150]	validation_0-rmse:0.11418
[151]	validation_0-rmse:0.11311
[152]	validation_0-rmse:0.11204
[153]	validation_0-rmse:0.11099
[154]	validation_0-rmse:0.10994
[155]	validation_0-rmse:0.10892
[156]	validation_0-rmse:0.10790
[157]	validation_0-rmse:0.10689
[158]	validation_0-rmse:0.10591
[159]	validation_0-rmse:0.10493
[160]	validation_0-rmse:0.10395
[161]	validation_0-rmse:0.10299
[162]	validation_0-rmse:0.10203
[163]	validation_0-rmse:0.10109
[164]	validation_0-rmse:0.10016
[165]	validation_0-rmse:0.09924
[166]	validation_0-rmse:0.09833
[167]	validation_0-rmse:0.09742
[168]	validation_0-rmse:0.09652
[169]	validation_0-rmse:0.09565
[170]	validation_0-rmse:0.09478
[171]	validation_0-rmse:0.09391
[172]	validation_0-rmse:0.09305
[173]	validation_0-rmse:0.09222
[174]	validation_0-rmse:0.09139
[175]	validation_0-rmse:0.09056
[176]	validation_0-rmse:0.08974
[177]	validation_0-rmse:0.08893
[178]	validation_0-rmse:0.08813
[179]	validation_0-rmse:0.08732
[180]	validation_0-rmse:0.08655
[181]	validation_0-rmse:0.08576
[182]	validation_0-rmse:0.08500
[183]	validation_0-rmse:0.08424
[184]	validation_0-rmse:0.08349
[185]	validation_0-rmse:0.08274
```



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| □ | ● | kaggle2 | 開啟 JUPYTERLAB | us-west1-b | – | TensorFlow:2.8 | 4 vCPUs, 15 GB RAM ▾ | 無 ▾ | Service account |
| □ | ● | kaggle3 | 開啟 JUPYTERLAB | us-west1-b | – | TensorFlow:2.8 | 16 vCPUs, 60 GB RAM ▾ | NVIDIA Tesla T4 x 1 ▾ | Service account |

# Conclusion

- 與量有關的feature為顯著特徵
- 乖離率特徵顯著
- Categorical feature並沒有想像中的影響顯著
- 訓練模型時有許多memory操作上的眉角
- Data preprocessing上有分general與個股的feature
- cuDF做EDA非常快