期末報告 機器學習

Jane Street Market Prediction

組別：大地之蓋亞

系級：資訊管理系三A

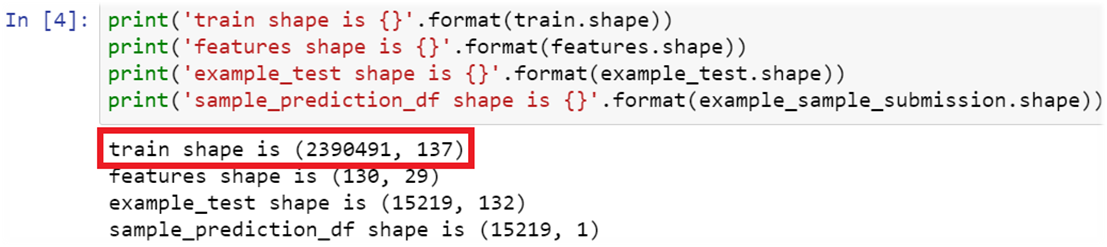
組長：B10756038施宗佑

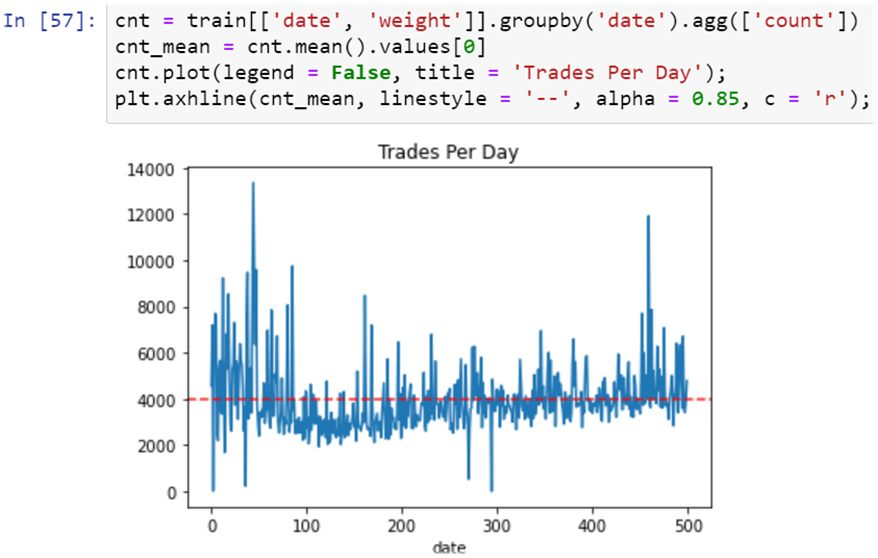
組員：B10756026林峻儀

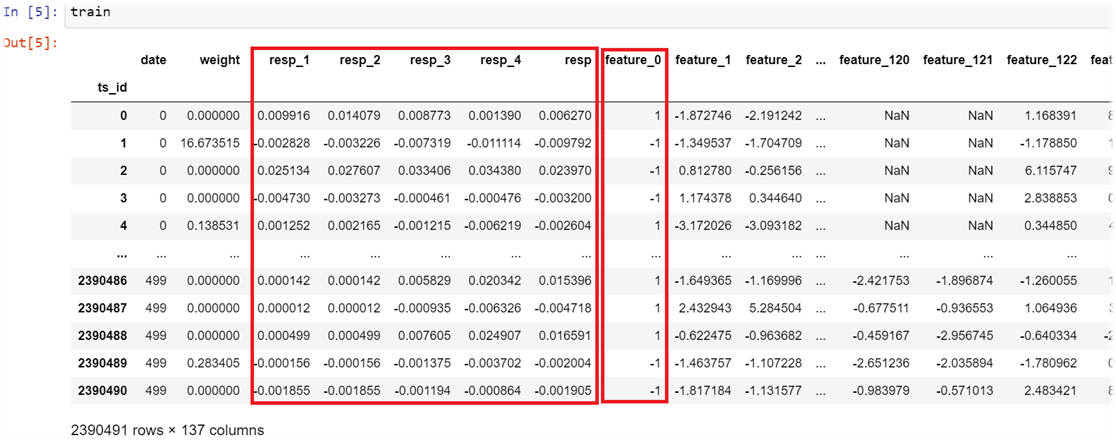
組員：B10756040郭家偉

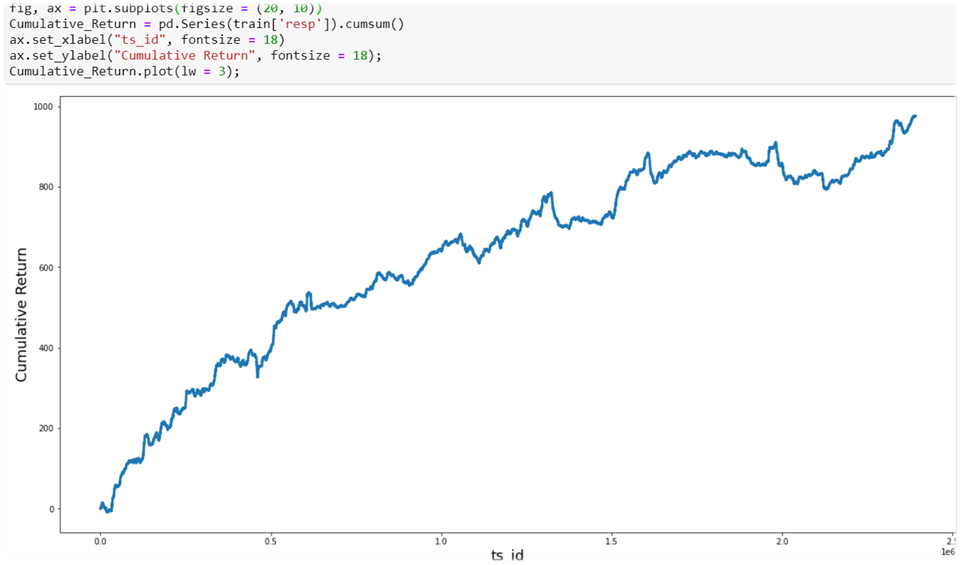
報告日期：2021年1月11日

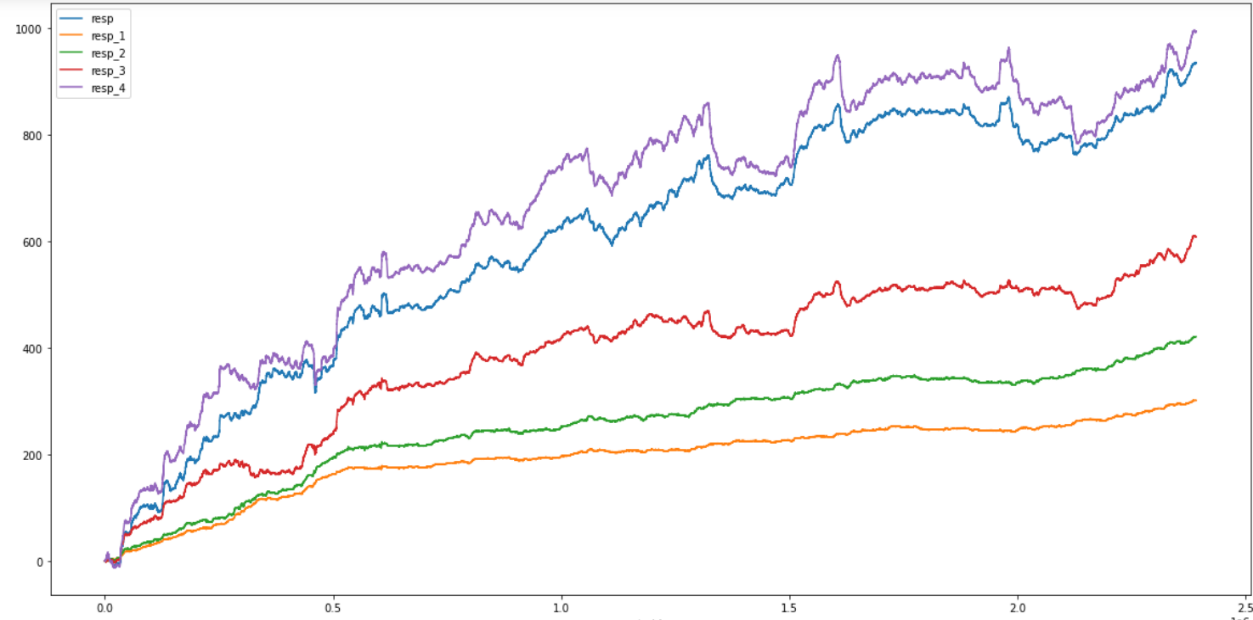
**1.資料分析**

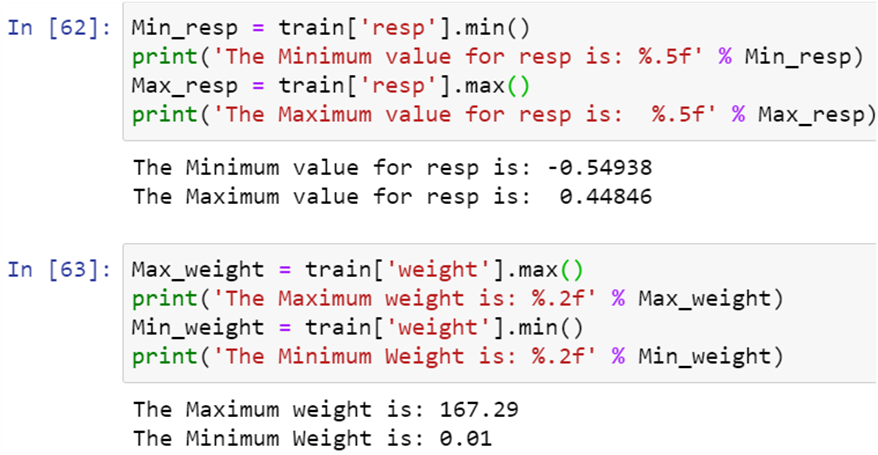
(1)觀察train.csv有將近5.8GB的資料量，且有破百萬筆的歷史交易資料。

(2)每日當天的平均交易量。

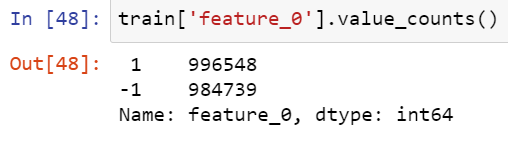
(3)資料中以resp~resp\_4與feature\_0當作我們觀測數據的指標。

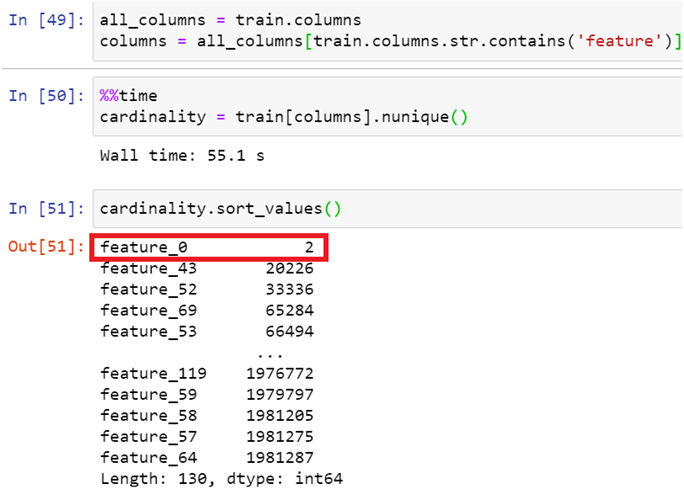
(4) train.csv中resp隨著時間的累積收益。

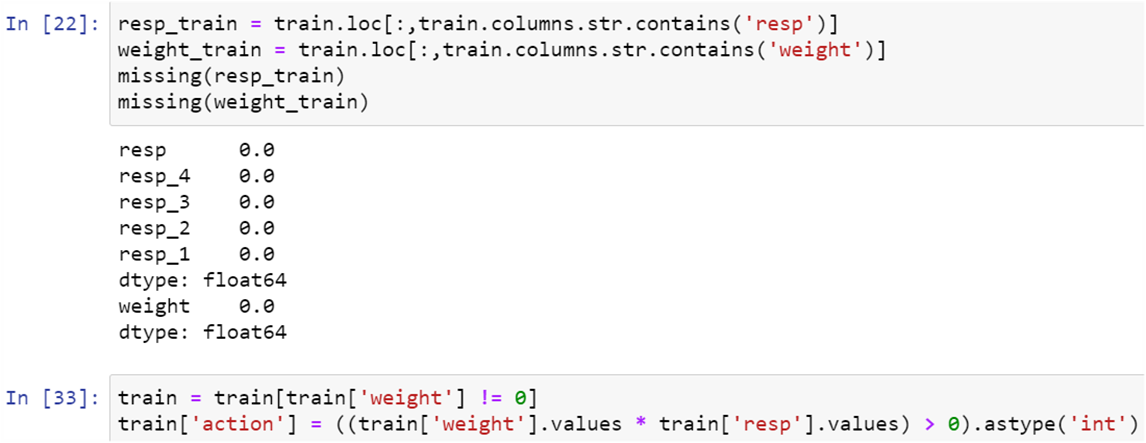
(5)其中藍色resp與紫色resp\_4的時間序列最接近。

(6)每筆交易都有關聯的權重(weight)和回報(resp)，代表每次交易的收益。

(7) train.csv 中的 feature 除了 feature\_0 明顯有分類外，其餘則沒有。





(8) resp和weight皆沒有缺失值，以此在train.csv最後一欄加入action column。

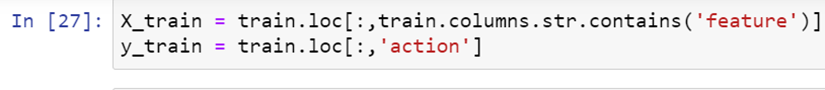
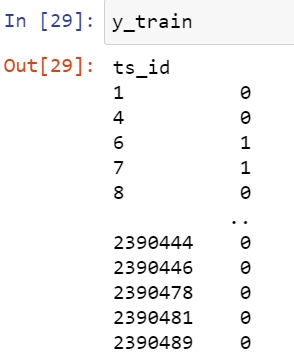
(9) feature\_0與其他特徵不同的點在於它是唯一沒有True的特徵。

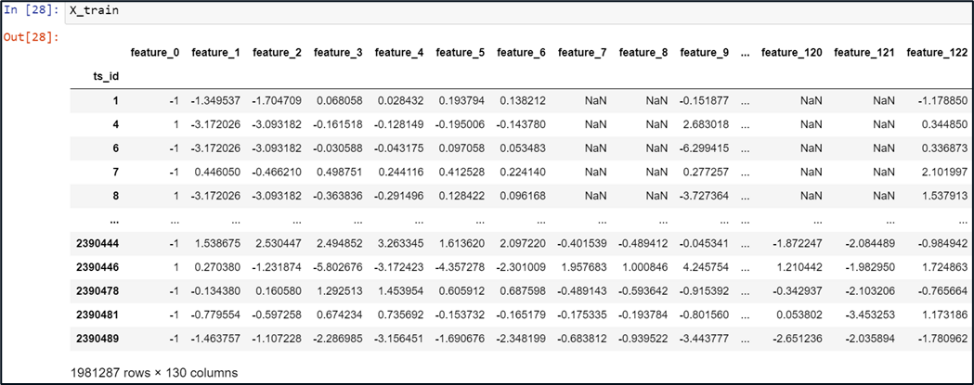


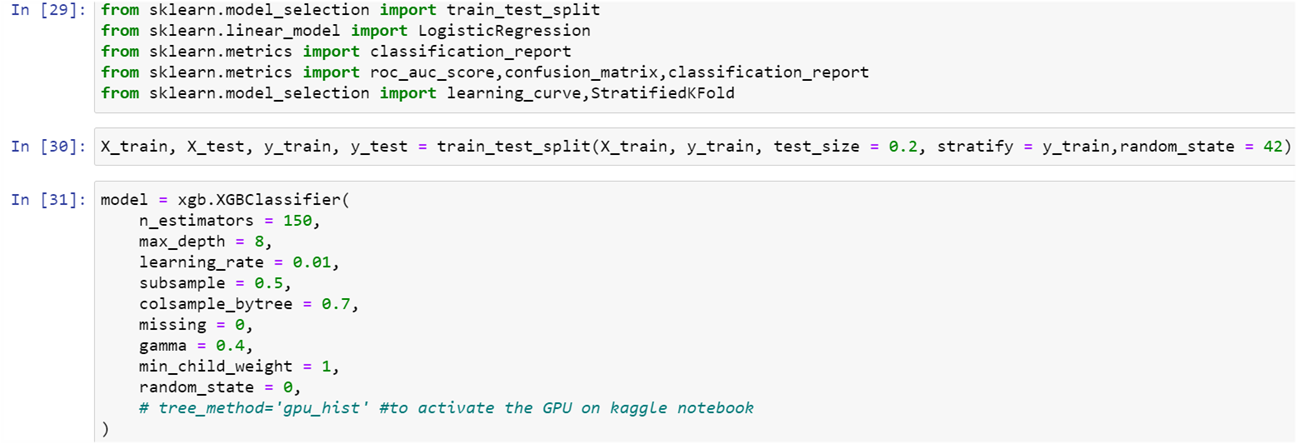
(10) features.csv中True為1，False為0，29個 Tag 關係著每筆 features。

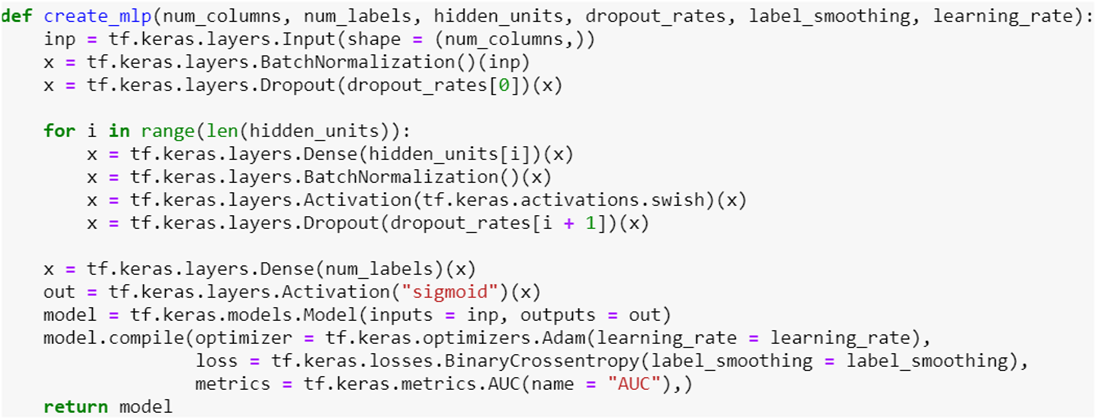


**2.建立訓練模型**

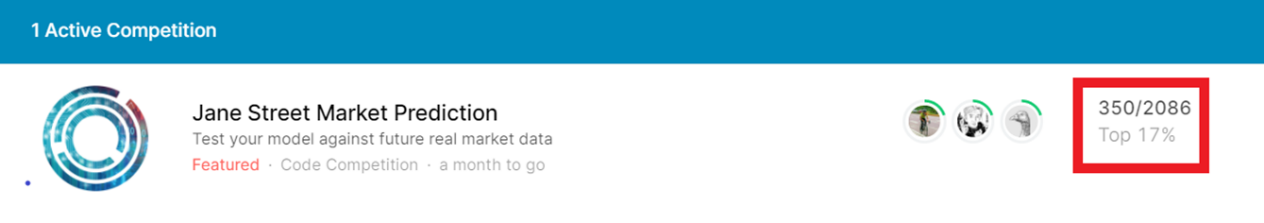
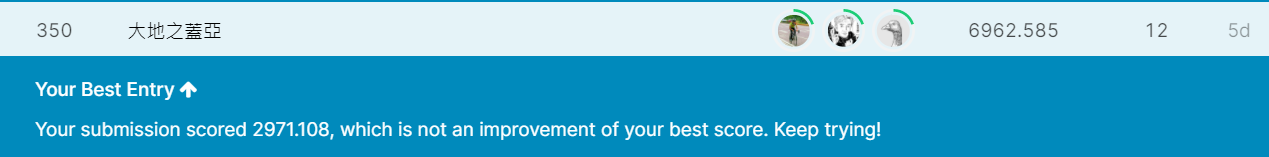
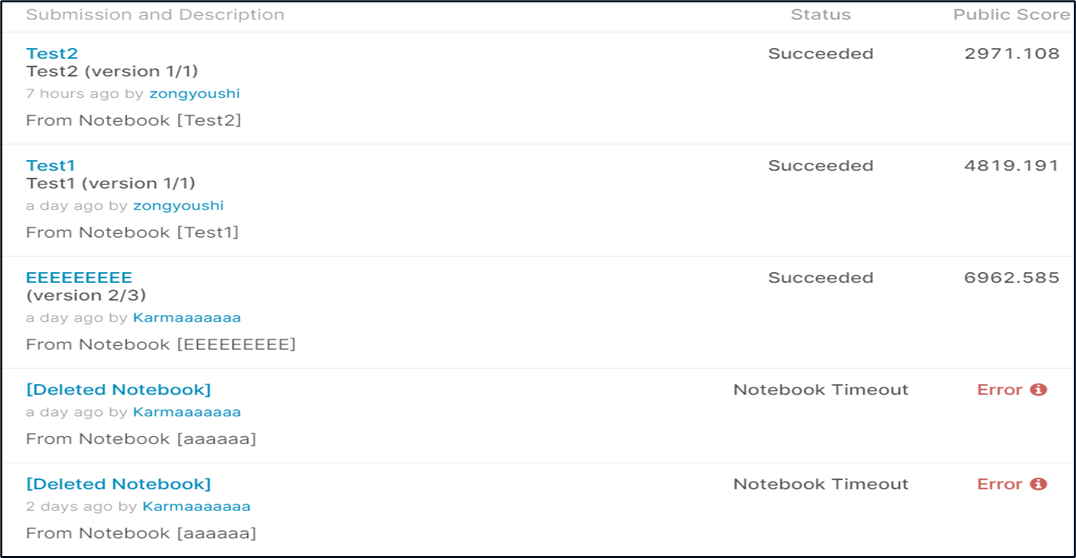
(1)將train.csv中的feature和action分別切割成 X\_train與y\_train。



(2)使用XGBClassifier模型對X\_train和y\_train進行訓練。

(3)使用Neural Network模型預測。

**3.最佳分數**

總共上傳12次，成功4次，失敗8次，由於每次上傳到Kaggle都需要2-3個小時，還不一定能成功，所以每次都要耗費時間做模型調整，光是上傳就耗費了多天。

**結論**

這次的期末專案，一開始我們是專注在預測模型的抉擇，比賽中公開的Notebook給了我們很多選擇，最後選取了XGBoost和Keras，主要是課程中有學到，也比較好上手，兩種方法都有各自的差異，最明顯的就是速度，也許是我們的參數設定各有不同，直接影響最後的預測跑分，而一開始我們的方向就是把分數衝高，最後才做資料的特徵分析，最終分數也讓我們覺得不錯，有前20%的名次，過程中看到各國的高手做出的預測模型，不同的思維有著多種的寫法，值得我們學習。

**組員分工**

|  |  |  |  |
| --- | --- | --- | --- |
|  | B10756026林峻儀 | B10756038施宗佑 | B1075640郭家偉 |
| Coding | ✓ | ✓ | ✓ |
| 資料蒐集 | ✓ | ✓ | ✓ |
| Word |  | ✓ |  |
| PPT |  | ✓ |  |

**Github**

<https://github.com/zongyoushi/NPUST_ML_Final_Project_Jane_Street_Market_Prediction_20210111.git>

**Kaggle**

<https://www.kaggle.com/c/jane-street-market-prediction>