MACHINE LEARNING FINAL PROJECT

統計所 R26111086 楊文博
統計所 R26114042 何佩欣
統計所 R26111028 林虔毅

隊伍名稱：Vivian_hsu0619

指導老師：許志仲 老師

Institute of Stastic
National Cheng Kung University

1. INTRODUCTION

本次報告會著重在說明此次 Kaggle 預測股票的比賽過程中，我們這組參考別組想法後，在此競賽做了什麼事情，最後使用 Catboost 進行預測，以及未來可行方向。

2. DATA

本次報告的 data 來自 kaggle 公開資料集 https://www.kaggle.com/competitions/optiver-trading-at-the-close，目的是開發一個能夠使用股票的訂單簿和收盤拍賣數據預測數百支納斯達克上市股票的模型，有助於從拍賣和訂單簿中整合訊號，從而提高市場效率和可及性。資料集中包含 5237980 筆資料、17 個變項，其中”Target”這個變項將作為預測目標，代表股票 wap 的未來 60 秒變動，減去合成指數的未來 60 秒變動(以美元為單位)。

3. DATA INTERPOLATION

由於是時間序列資料，使用平均數或中位數對 NA 值插補不太合理，這邊會使用離此筆資料最近之上筆資料插補。而 ”far_price” 與 ”near_price” 之 NA 值已超過整體資料的 55%，會直接移除此兩種變數。

4. ADD VARIABLES

會使用扣掉目標變數之 16 個變數產生額外變數，例如 ”volume” 為買方和賣方的訂單量總和、“size_imbalance” 為買方和賣方訂單量的比例等等，另外也會對變數進行差分，總共有三種差分分別為：

1. Shift：此列的值將被往前平移 n 個時間步。
2. Ret：此列的值相對於前 n 個時間步的值的變化率。
3. Diff：此列的值的 n 階差分。

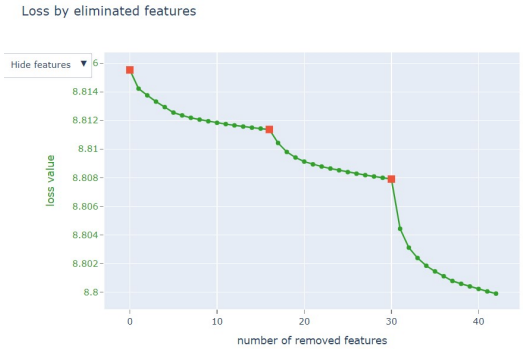
其中 n 會考慮放入[1,2,3,4,8,12]。接著再加入原有變數的統計特徵，包含平均值、標準差、偏度和峰度成為新變數，最後建立與時間相關變數包含：

1. 計算星期幾："date_id"除以 5 的餘數。
2. 計算秒數："seconds_in_bucket"除以 60 的餘數。
3. 計算分鐘數： "seconds_in_bucket"除以 60 的商。

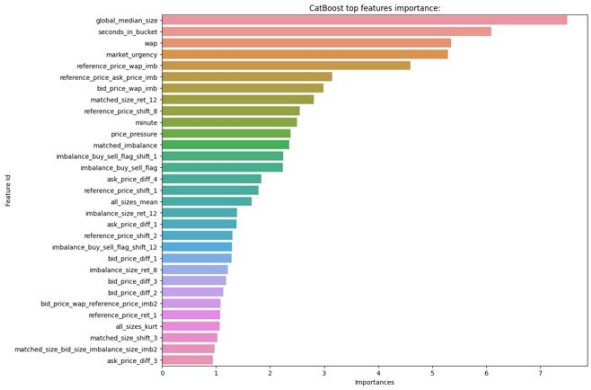
擴充後總共有 142 個特徵。

5. FEATURE SELECTION AND MODEL

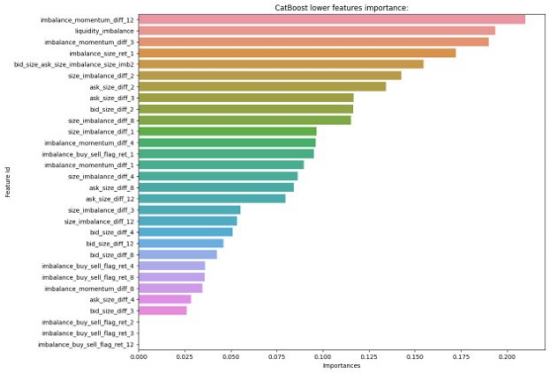
特徵選擇的算法是透過計算 SHAP 值，訓練過程中自動從資料裡篩選對模型預測有用的特徵，設定選擇特徵的步驟數為 3 步，每個步驟刪去 14 個特徵，最後只會剩下 142-14*3=100 個特徵當作 Catboost 的解釋變數，可以觀察 loss value 圖發現其的確是穩定往下。



畫出特徵重要性圖，分別為最重要及最不重要。



發現擴增 n=8,12 的變數會有較好的效果。



發現在 n=2,3 的 Diff 插分重要性程度明顯較差。
使用 catboost 模型

- 迭代次數=700
- 學習率=0.1
- 決策樹的最大深度=12
- 損失函數=RMSE

得到最終成績為 5.3641。

透過更改超參數

- 迭代次數=600
- 學習率=0.1
- 決策樹的最大深度=8
- 損失函數=MAE

得到最終成績為 5.3501，排名約 1900 左右。

6. FEATURE WORK

未來會考慮使用 k-fold cross-validation 使結果更加嚴謹，以及使用 XGboost 和 LGBM 模型並將結果 ensemble。

Github 連結：https://github.com/ywb0401/ml-final