2025/ 11/ 12、11/ 19

20:30 - 21:30

ML Group 4th Meeting Minutes

Topic	checkpoint discussion
Place	Google meet
bullet points	分享各自模型做了哪些優化，優化後的表現和可能的原因，列舉如下： 1. KNN：加入特徵工程跟美股後，accuracy 才約在 50%上下。推測可能原因：KNN 的核心假設是「特徵相似 → 行為相似」，但股價具備高度噪音和受市場情緒影響，使得相似特徵無法保證相似走勢。特徵維度太高也使 KNN 的距離度量在高維空間中失去辨識力。 2. LightGBM： 經過大量特徵工程 ex：加入美股(NASKAK 等)、交叉信號、Lag 特徵等，以及由 SHAP 可視化進行特徵篩選，經 OPTUNA 調參後，預測一天後的股價漲跌 accuracy 約可達 67%。但預測五天後因市場噪聲太強、可預測性下降，約只能達 50%左右的準確率。 3. LSTM： 除了使用台積電本身的股票資料以外，還加入了台股指數、台幣兌美元匯率、美股等一系列相關的特徵，在使用前一天美股的資料加入 input 來進行預測後，能夠達到 70%以上的準確率。 4. XGBoost： 由預測股票價格改成預測股票漲跌，加入 NASDAK 指數作為特徵(其他測試過的美股相關指數皆降低模型的準確率)，使用前 N 天的資料預測，經測試 N=50 的時候可以達到最高的準確率: 預測一天後的準確率約 58%，預測五天後的準確率則有 60%。 5. RF：加入特徵希望增加準確率，除了加入美股當天的相關數據之外同時也將 MA5、MA10 這類帶有絕對股價的特徵，改為 close 是否大於 MA5 這種相對的二元特徵，加入後預測隔天的準確率可以到 71%，預測後天的如果沒有調參數的話有 61%，調參數後可以到 65%左右，比起原本的 51%好很多。 6. Logistic Regression/ DNN:透過前 30 天的股價資料預設 1 天後(D1)或 5 天後(D5)的股價，使用 logistic regression 時 train 跟 valid 準確度都只有 50%左右，因此換成 DNN，提升 unit 數跟 layer 數後，train 的準確度可以達 90%以上，而 valid 的準確度卻依然只有 50%左右，即使使用 dropout 和 L2 等 regularization 方法，D1 的 valid 準確度依舊只有 52%左右，但是 D5 的準確度能達到 69%左右。

	7. Naive Bayes: 在進行大量特徵工程與加入更多美股資料後，模型在 train 與 valid 的準確度仍只有 55% 上下。可能原因為 Naive Bayes 假設特徵間都是獨立的，然而在真實的股票市場上特徵間都是高度相關的，甚至部分數據是由其他特徵得來（如部分指數、趨勢線）。另外因為此模型本身對噪聲敏感，不排除在數據的高噪音對模型有產生顯著的影響。然而，我發現開盤價與前一天漲幅有顯著關係，之後可能會朝者此方向做研究。 討論發現的問題： 1. 資料洩漏：假設有台股、美股 11/20 開盤資料，拿來預測 11/21 資料，因美股 11/20 資料在台灣時間 11/21 凌晨才開盤，模型可以由此作弊。因此，應去美股 11/19＋台股 11/20 開盤資料預測 11/21 漲跌可能。 2. 由於台積電的股價整體趨勢是在不斷上漲，因此對於模型來說，驗證集或測試集中往往會出現模型在訓練階段 沒有見過的新高股價數值，這可能會讓模型無法達到很好的表現。對此需要加原本一些價格的特徵轉變為漲跌的百分比。
In attendance	郭宣汝(Present), 高予謙(Present), 謝尚錡(Present), 曾煜展(Present), 郭子維(Present), 盛樺(Present)
Task Assigned	1. 各自模型持續優化。可做特徵工程、調參等嘗試。試著理解為何模型能力會提升或下降。 2. 解決資料洩漏問題。 3. 交易模擬：通過模型輸出的漲跌機率和確信程度，進而決定當天交易決策，通過模擬交易，得到每天的交易結果、賺／賠和帳戶餘額，展示 2025/01/01 至 2025/11/20。 4. 視覺化呈現：將交易模擬和模型的輸出作為輸入，視覺化交易決策決定原因（如模型預測漲且確信程度 80%，當天則可能大量買入）、賺／賠金額、帳戶餘額變動等，有助於 final presentation 的呈現。 5. Ensemble：將各模型的輸出整合，試著通過類似 stacking 的手法，讓預測漲 or 跌的準確值再提升。同樣需完成交易模擬和視覺化呈現。

Next meeting	Date: 2025/11/26 Time: 20:30 Objective:討論各自模型進展、final presentation 形式。 Location: Google Meet