Data Mining Term Project Report

Group 12

312551098 林佑庭, 0811295 黃俊翔, 109705001 陳以瑄, 109705003 吳振豪

1 計畫目標的問題 (Target problem)

1.1 目標問題描述（需包含資料輸入/處理過程/輸出）

本計畫的目標問題是預測股票實現波動率（Realized Volatility。實現波動度指的是某個股票在一段特定時間內實際價格變動的幅度，最常見的方法是計算收盤價格之間的日度變動的標準差，然後乘以相應的年化因子。

我們想預測實現波動度的原因在於，進入金融市場的目標無疑是追求盈利。想要賺錢就需在低點買進、高點賣出，也就是說我們會需要精準預測價格走勢，找到進出場的時機。但是，即便對股價有了一定的預測能力，卻並未衡量未來股價可能的變動程度 (即實現波動度)，在設計交易策略時，並不能同時考慮可能的風險，從而導致潛在的鉅額損失。舉個簡易的例子: 一模型預測某一資產會在 5 秒後漲到 100 元，這是對資產價格的點估計 (在價格的機率分布上，平均值是 100 元)，但價格機率分布的標準差 (實現波動度) 可能是 ±10 元，若交易者在 99 元進場，可能的收益範圍在-9 11 元，約莫是 ±10 頻交易是相對非常不穩定的收益。此外實現波動度也在投資組合管理及衍伸性金融商品定價中有非常重要的地位，如衍伸性金融商品的價格會直接受標的資產的實現波動度影響，針對不同的實現波動度也可組建不同的交易部位和標的，來賺取報酬或進行波動度套利。因此除了預測股價之外，精準預測實現波動度也是十分重要的任務。

目標問題的定義:

- 輸入: 給定歷史的最佳兩檔揭示檔與實際交易的價格與交易量資訊
- 處理過程: 先透過財金技術指標做特徵工程，提取股票數據的重要特徵。再利用 K-means clustering 將股票分為群組，並找出每個群組的共同特徵。然後使用 LGBM、FFNN 和 TabNet 建模，最後進行 ensemble 以提升整體預測效能。
- 輸出: 預測未來 10 分鐘內的實現波動度

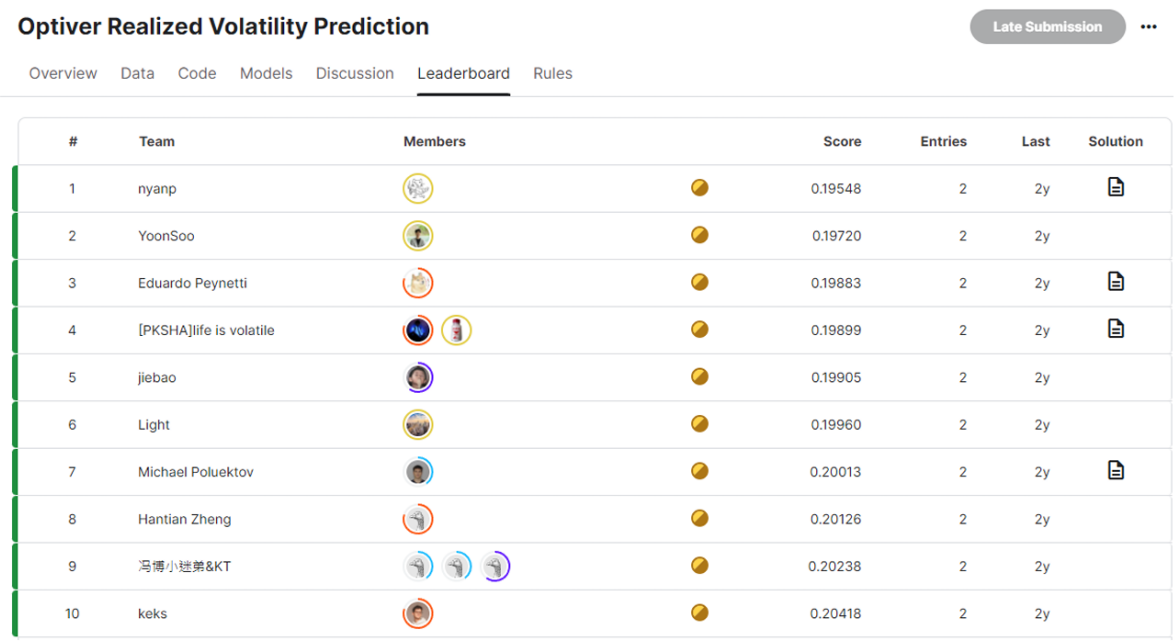
1.2 評估指標

由於本計畫的問題同時是 Optiver 在 Kaggle 上舉辦的一場競賽題目 (連結)，因此我們與該競賽採用相同的評估指標 RMSPE，計算公式如下:

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{y}_i}{y_i})^2}$$

1.3 原預估之模型效能與目標

我們在參考競賽的排行榜之後，決定將預期 RMSPE 設定在 0.2，而這個分數將會進到榜單的前十名。目前的排行榜如下圖所示:



#	Team	Members	Score	Entries	Last	Solution
1	nyanp		0.19548	2	2y	
2	YoonSoo		0.19720	2	2y	
3	Eduardo Peynetti		0.19883	2	2y	
4	[PKSHA]life is volatile		0.19899	2	2y	
5	jiebao		0.19905	2	2y	
6	Light		0.19960	2	2y	
7	Michael Poluektov		0.20013	2	2y	
8	Hantian Zheng		0.20126	2	2y	
9	冯博小迷弟&KT		0.20238	2	2y	
10	keks		0.20418	2	2y	

2 選用的資料集描述 (Descriptions of selected datasets)

2.1 目標問題描述（需包含資料輸入/處理過程/輸出）

我們採用的資料集來源是前面提及的 Kaggle 競賽 (連結)

2.2 資料集相關描述

此資料集共有三種檔案:train, book, trade，涵蓋 112 種股票的資訊，接下來將依序對這三種檔案做詳細的介紹。

2.2.1 Train

此檔案表示的是某股票在某時間段的實現波動度，即為本計畫預測的 ground truth。一共有 3 欄 * 428932 筆，且無缺失值。

資料欄位	型態	意義/用途
stock_id	整數	股票編號，共 112 種
time_id	整數	時間段編號，共 3830 種。非連續，但所有股票編號一致
target	小數	未來十分鐘內的實際波動度

2.2.2 Book

112 種股票有分開的 book 檔，此檔案表示該股票在某時間點的揭示訊息，即最佳兩檔的委買與委賣價量資訊。一共 11 欄 * 774671 2295344 筆 (每支股票筆數不同)，且無缺失值。

資料欄位	型態	意義/用途
stock_id	整數	股票編號
time_id	整數	時間段編號
seconds_in_bucket	整數	bucket 中位列第幾秒
bid_price1	小數	最佳的標準化買價
bid_price2	小數	次佳的標準化買價
ask_price1	小數	最佳的標準化賣價
ask_price2	小數	次佳的標準化賣價
bid_size1	小數	最佳的欲買量
bid_size2	小數	次佳的欲買量
ask_size1	小數	最佳的欲賣量
ask_size2	小數	次佳的欲賣量

2.2.3 Trade

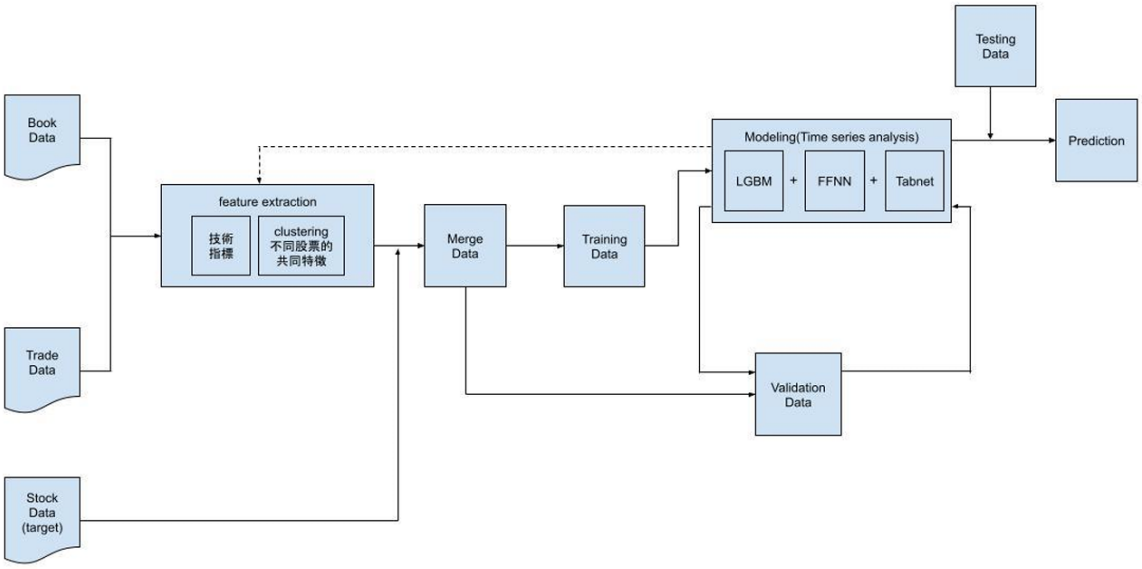
112 種股票有分開的 trade 檔，此檔案表示該股票在某秒的成交資訊。一共 6 欄 * 84294 1453087 筆 (每支股票筆數不同)，且無缺失值。

資料欄位	型態	意義/用途
stock_id	整數	股票編號
time_id	整數	時間段編號
seconds_in_bucket	整數	bucket 中位列第幾秒
price	小數	一秒內交易的平均標準化價格
size	整數	該秒內總交易量
order_count	整數	該秒內成交訂單數量

3 針對問題設計的分析流程 (Analysis workflow)

3.1 流程圖

我們先對 book 跟 trade 的檔案進行兩階段的特徵工程，第一階段是計算財金技術指標，第二階段是利用 k-means clustering 將股票分群，並計算共同特徵。接著將資料整併與切分成 training dataset 與 validation dataest，進行三種模型 LGBM、FFNN、Tabnet 的訓練並用 ensemble 來優化表現。最後將訓練好的模型送進 Kaggle 競賽，對他們的 private testing dataset 做預測，並評估模型的表現。



3.2 選用之資料探勘方法以及選用原因

我們的特徵工程一共有兩個階段: 取得技術指標的特徵與找出類似股票的共同特徵。

3.2.1 財金技術指標

原始資料中的資訊非常有限，每個時間點僅提供了 11 個與價量相關的數據。然而，在實際的交易中，我們除了觀察原始價格波動，還會從這些數據中推導出多種技術指標，例如移動平均線等。因此我們參考了常見的金融技術指標，從僅有價格和成交量資訊的揭示檔與成交檔中提取了 226 種特徵。
特徵主要分為七類，考慮到本計畫涉及時間序列，每個技術指標都會有多種回溯窗口 (look-back window)，分別為前 100,200,300,40 0,500 筆資料。具體分類如下:

1. Weighted Average Price
   委託量加權買賣價是用來估計商品的真實價格，他的原始計算公式如下:

$$\mathrm{WAP} = \frac{\mathrm{BidPrice} * \mathrm{AskSize} + \mathrm{AskPriceBidSize}}{\mathrm{BidSize} + \mathrm{AskSize}}$$

   其中價量的資訊來源可以是最佳檔也可以是次佳檔，因此後續還延伸出比較最佳與次佳兩檔估值的 WAP Imbalance，作為觀察市場傾向的指標。

2. Price Spread
   價差是用來判斷市場流動性的指標，他的計算公式如下:

$$\mathrm{Spread} = \frac{\mathrm{AskPrice} - \mathrm{BidPrice}}{\mathrm{AskPrice} + \mathrm{BidPrice}}$$

   除了比較買賣方的價格之外，後續還延伸出比較兩檔賣價的 Ask spread、比較兩檔買價的 Bid spread，以及比較買價差與賣價差的 Bid-Ask spread。

3. Price to Moving Average
   價格相對於移動平均的差可以反映當前價格與趨勢的偏離程度，進而提供價格可能回撤或補漲的訊息。計算公式如下:

$$\text{Price to MA} = |\text{mid} - \text{mid\_MA}|$$

$$\text{mid} = \frac{\text{AskPrice} + \text{BidPrice}}{2}$$

4. Log Return
   這也是一個與價格相關的技術指標，他呈現出價格相較前一刻的變化，計算公式如下:

$$\text{log return} = log(\frac{price_t}{price_{t-1}})$$

5. Volume Imbalance
   除了價格可以延伸出技術指標，委託口數的資訊也很重要，VI 的功用就是可以呈現買賣量差，來衡量買賣力道的強弱。計算公式如下:

$$\text{VI} = |\text{AskSize} - \text{BidSize}|$$

6. RSI
   相對強弱指標是評估股票買賣熱度時一個常用的指標，主要用於判斷市場是處於超買還是超賣條件。他是通過比較最近一段時間內的平均收盤漲幅和跌幅的大小來計算，公式如下:

$$\text{RSI} = \frac{\text{N 天平均漲幅}}{\text{N 天平均漲幅} - \text{N 天平均跌幅}}$$

7. Size Tau
   除了價量資訊，時間資訊對交易而言也是相當重要,size tau 的主要功能是衡量交易頻率，公式如下:

$$\text{Size tau} = \sqrt{\frac{1}{\text{count of unique seconds in each time bucket}}}$$

3.2.2 股票分群

當我們真正進行交易時，除了會參考目標股票的技術指標外，也很常參考同類型股票的資訊，例如投資人們常將股票區分為金融股、科技股、航運股等，因為我們相信同產業的股票有同樣的市場趨勢，所以波動度應該會有類似的變化。
由於我們只知道股票編號，並不曉得實際對應到的股票為何，因此我們決定利用分群技術，區分不同類型的股票並取得他們的共同特徵。我們採用的算法是 K-means clustering，主要原因是我們的股票數量很多，而 K-means 是個簡單且計算效率高的分

群算法。透過 K means 我們將股票分成七群，並分別計算各組在每個 time_id 下的平均數據作為共同特徵，與前一階段的結果整併後，最終有 280 個特徵。

3.2.3 特徵迴歸

透過資料的前處理，我們可以得到一個維度為 (428932, 280) 的資料，而我們所要做的就是利用每項資料的 280 個 feature 來做迴歸分析得出 10 分鐘內的波動度。
由於 competition 的要求，我們必須使用 kaggle notebook 完整的進行 compile，並且要在規定的記憶體與時間 (9 小時) 內完成，因此在模型選擇上我們必須捨去部分模型，例如 LSTM。也因此，我們在模型選擇上大多選用輕量化和訓練速度較快的模型，利用 k-fold cros s validation 的方式來增加 robustness 和準確度。以下是我們所使用的三種模型:

1. Tabnet
   TabNet 是一種由 Google Cloud AI 研究團隊開發的深度學習模型，專門用於處理表格數據。這種模型的設計是為了結合深度神經網絡的能力和決策樹模型的直觀解釋性。TabNet 的一些關鍵特點: 自適應特徵選擇、解釋性、深度學習的優勢、效率等優勢，讓 tabnet 可以透過特徵選擇來讓使用者可以去對訓練資料的特徵有更近一步的認知，並且進一步提升效率，讓我們能夠在訓練過程更有餘裕去做多餘的操作。

2. lgbm
   Gradient Boosting Decision Tree 是一種基於決策樹的學習算法，利用梯度提升（gradient boosting）的方法進行模型訓練。這種方法在處理各種結構化數據時非常有效。結合多個弱分類器去得到更好的分類效果，並且支援大量的 hyperpa rameter 調整來適應多種資料可能。

3. ffnn
   即為全連階層的神經網路，但對於部分資料，如 stock id，我們會特別將其撈出轉換成 embedding 後併回原始資料再餵入神經網路，最後再利用 linear activation 轉換為結果得到波動度。

3.3 模型評估方法（如：K-Fold cross validation）

我們使用 k-fold, k=5 來拆分 training data 和 validation data，由於 t esting data 只能透過 submit competition 來得到結果，因此沒有 da ta leakage 的問題。其中，針對 Tabnet 採用了 5 個不同的 seed，從中挑選一個平均 validation 結果最好的模型作為最終模型。最後針對我們使用的三個模型賦予不同的權重，透過 ensemble 的方式得到一個最終的答案。

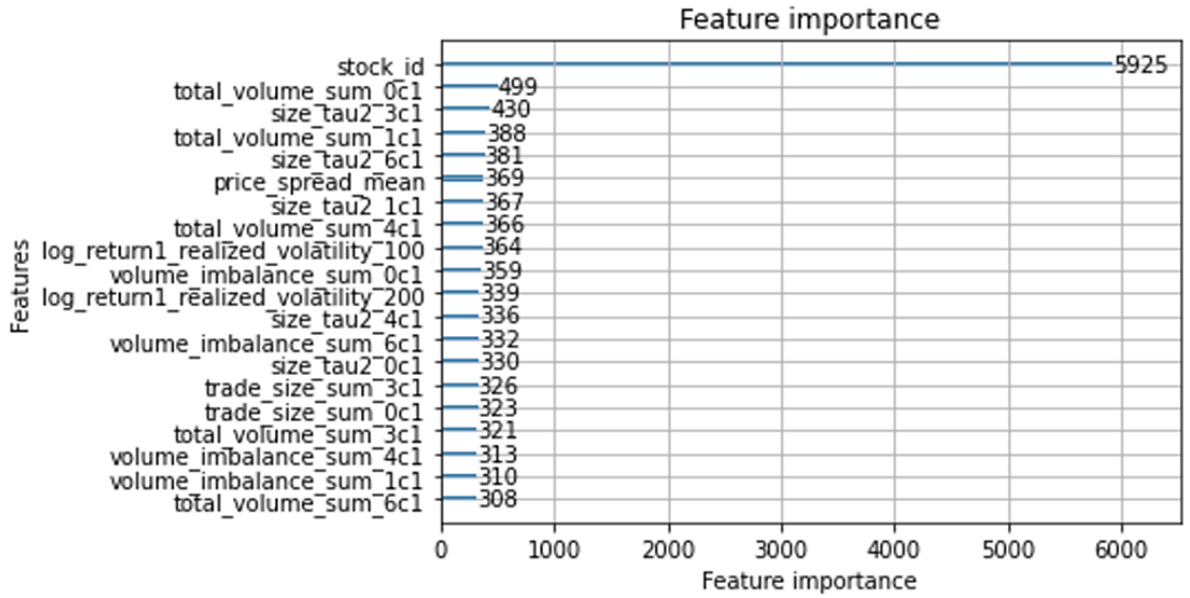
3.4 所採用之平台/工具

我們使用了兩個硬體平台進行開發：1. 2080 ti (jupyter notebook) 2. GPU T4 x2 (kaggle notebook); 軟體部分則是使用 python 以及相關套件：Python, keras, tensorflow 等。

4 分析結果 (Analysis results)

4.1 實驗結果（需有圖表，以及針對圖表的闡釋）

就 LGBM 所得出的 feature importance 我們可以發現，決定波動度最大的一個條件便是股票本身。其次則對 volume 與 tau size 有比較大的敏感度。由委託量及交易頻率可能可以推斷當前市場的熱度，在較熱門的市場時，由於流動性更高，價格變動會相對較穩定，導致可能較低的實現波動度；此外可發現較重要的特徵有多數為 K-means 結合相近股票的平均數據，或許也可以依此結論推斷，有相近市場熱度的股票，其實現波動度可能也比較相似。



在我們的測試中，利用了不同的權重參數來調整三個模型的比例。結果如下：

TabNet	LGBM	FFNN	Results
1	0	0	0.25473
0	1	0	0.22274
0	0	1	0.24043
0.3	0.4	0.3	0.22593
0.3	0.3	0.4	0.22421
0.25	0.5	0.25	**0.21994**

在最終的結果，我們達到了 leaderborad 前 4 用多個模型來對單一回歸任務 ensemble 可以達到更精確的效果。

5 過程中遭遇的挑戰以及總結 (Discussion and Conclusion)

5.1 挑戰一、些許誤差對排名的影響

在實作的觀察中，我們發現預測目標的數量級往往落在 1e-3 上下，這代表當我們使用 RMSPE 作為模型指標時，少許的誤差都會使得分數急遽的變化。也因此我們認為使用多個模型加權可以減少單一模型 out lier 的出現，來讓整體的結果更加穩定。

5.2 挑戰二、缺乏逆向工程技術

在我們搜尋 Leaderboard 上其他的可行辦法時，發現部分頂尖 kaggle r 會利用逆向工程得到真實股票的資訊，透過在 local 預先對相似資料進行大量訓練，得到精煉過後的 feature pattern 後再加入與 test ing data 一同訓練，來讓 feature engineering 中 clustering 的結果可以更加準確。惟討論區並沒有公開他們是如何對這些額外的 data 做處理，我們也沒有太多的時間去研究如何去做逆向工程，所以只能用主辦方提供的 public dataset 來做 data mining。

5.3 總結

即便沒有透過額外的資料獲取更加精確的資料，我們仍能夠透過正確的選擇模型、財經背景的特徵分析、以及適當的參數調整來得到 l eaderboard 前 4% 的成績。

6 參考文獻 (Reference)

1. Optiver Realized Volatility Prediction
2. TabNet: Attentive Interpretable Tabular Learning
3. LGBM document
4. Technical indicators for trading stocks]

7 組員分工與各自執行細項 (Work distribution chart)

1. Feature extraction Data Analysis: 陳以瑄、吳振豪
2. Model selection Training: 林佑庭、黃俊翔