**G Research Crypto Forecasting Competition**

**Optimize winning model's performance in bear market**

葉力嘉、黃元裕、陳韋勳、胡祖望

台大財金碩二

r10723056@ntu.edu.tw

r10723057@ntu.edu.tw

r10723058@ntu.edu.tw

r10723059@ntu.edu.tw

**Introduction**

G Research是一家位在歐洲的量化金融研究公司，於2021年底時在Kaggle平台上舉辦了一場加密貨幣預測競賽[1]。 該比賽的參賽時間是從2021/11/2至2022/2/1，預測的區間是從2022/2/1至2022/5/4。然而，比特幣與其他加密貨幣在比賽結束後沒多久就連續發生了數次市場崩盤，加密貨幣的價格開始出現往下的趨勢，也就是進入了所謂的熊市。

於是我們好奇這些競賽中的得獎模型，是否在熊市的情況下表現也是很優良，所以我們決定去檢視以上想法並且嘗試去提出改善原有模型的方法。



(圖1)

**Method**

**-Task Definition**

我們從競賽中找到有公開的第三名模型[2]，並對此模型進行分析，並希望能找到此模型的漏洞，希望經過我們的改進在面臨2022/5/4後的幣災，能擁有比原模型更好的結果。

原資料有14檔加密貨幣的資料，但因為我們電腦的記憶體與時間的不足，所以我們只專注在其中兩檔最有名的加密貨幣( Bitcoin, ETH )進行分析。

**- Approach (features & 5 model definition)**

我們的方法大致可以分成兩種：1.加入新的feature：以捕捉到原模型未捕捉到的能預測幣災的因子 2. 加入新的model：原model雖然已經得獎，但在以往未發生過幣災的影響下，原模型可能無法應付，因此我們透過加入新的模型，加強對新資料的影響。

由於模型需要預測2種加密貨幣的結果，因此我們設計了6種類型的模型(由於其中兩種(NN, LSTM)表現不好，一種(Attention機制)時間不足沒有跑出分析結果，因此不放數據結果，但依然會在此作介紹)。

**1.原作者**：單一模型(Light GBM)，其一次學習與預測 2 種加密貨幣的結果。 這種模型設計上相對直觀，不須將加密貨幣分開來處理，因此它假設兩個加密貨幣之間有某種關係在。在此先比較它的優缺點，優點: 1.直觀，設計上簡單。2.模型訓練較快，預測也是。3.相較節省記憶體。4.若兩種貨幣間有相關性可學習到。缺點:忽略加密貨幣之間可能有不能共存的關係。

原model feature:

for lag = 60,300,900(單位:分鐘) id=0~13 i= start\_time~end\_time

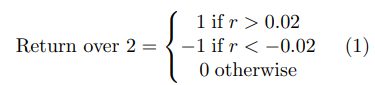
1. log\_close/mean\_lag[i] = log(close[i]/mean(close[i-lag:i]) 近期**平均**漲(跌)勢
2. log\_return\_lag[i] = log(close[i]/close[i-lag]) 近期的漲(跌)勢
3. mean\_close/mean\_lag[i] = mean( log\_close/mean\_lag[i]\_{id=0~13} )
4. mean\_log\_return\_lag[i] = mean( log\_return\_lag[i] \_{id=0~13} )
5. (3)-(1) 單論單一資產近期**平均**漲(跌)勢
6. (4)-(2) 單論單一資產近期的漲(跌)勢

**2.原作者＋新feature：**

新Feature介紹

for lag = 60,300,900 id=1,6

1. Return over 2 :



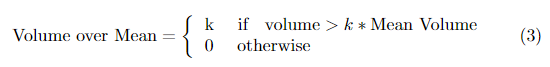
in last lag minutes

1. mean volume



mean is in last lag minutes

1. Volume over Mean:

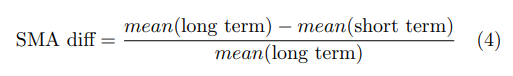


for k = 1,2,3,4

後面是股市技術分析常用並參考了一些論文選取的指標 [3] [4]：

[short\_term, long\_term] = [60,120], [120,180], [180,240]

1. SMA\_diff :解釋短長期K線差距，可解釋相較長期，短期是漲勢(SMA diff<0)或是跌勢(SMA diff>0)



1. bbands :解釋現在的價格與近k分鐘價格相比是percentile多少，代表現在相較近k分鐘是漲勢或是跌勢。



where upper is the former lag minutes (mean + 2 std)

,lower is the former lag minutes (mean - 2 std)

,and close is the close price of this minute

1. RSI 解釋市場熱度(value = 0~100, 0:近k分鐘都沒漲，近期大跌、100:近k分鐘都沒跌，近期大漲)



where A is the average of positive return in the former lag minutes,

B is the average of negative return in the former lag minutes.

1. ATR:代表真實波動區間，與bbands的概念很像，但多考慮了價格並非連續變動，有跳空現象發生，因此做出調整



where HMC = (this minute high price) - (last minute close price)

CML = (last minute close price) - (this minute low price)

HML = (this minute high price) - (this minute low price)

Three of them are three columns within the former lag minutes (size=lag)

1. NATR: Normalize the ATR with close price



接下來我們會再加上新的model，以下介紹New model的設計方式:將原作者＋新feature的 model predict出來的 y hat當作一種feature，因為predict出來的 y hat 已經捕捉到了一些趨勢，再將他與其他我們新設計的crach-relevant features結合，訓練一個新model(NN，LSTM，LGB)

**3.原作者＋新feature + New model(LGB):**結合原有模型的優勢，再疊上一個新的模型來使結果更好。再次使用lgb模型訓練速度快、高效率、低記憶體使用率及更好的準確度等等優勢。但lgb的缺點就是過擬和，可能會訓練到一些雜訊資料等等。

**4.原作者＋新feature + New model(NN):** 我們想說做NN model 當作 baseline與其他model做比較但做出來後發現預測力極差，可能原因是因為模型過於簡單，難以預測如此複雜的金融市場。

**5.原作者＋新feature + New model(LSTM):**我們想說做LSTM model 當作第二個比較結果的model，會選LSTM是因為他蠻適合做timeseries的工作因此選他。但做出來後發現預測力也很不好，可能原因是因為1.模型layer太少無法有效的捕捉到金融市場時間序列的特爭，2.由於資料的跳動幅度很大且毫無規律，在輸出結果時發現model為了降低loss會訓練出將預測結果走在平均值的附近，也就是training時 model 的 predict y 會在true y 的平均值上游走，偶爾跟者上下起伏，我也不太確定造成此結果的原因，我猜是有可能是optimizer的問題，我選的是Adam，未來可能要在特別去挑optimizer。

**6.原作者＋新feature + New model(Attention):** 原本預計會嘗試另一種跟Attention機制有關的模型，但因為時間關係，沒辦法取得資料並訓練得到結果。Attention機制新模型的架構已設計好且程式碼已完成。

設計直覺：原始訓練資料中有包含18年的熊市，但我們觀察到原作者模型在熊市時的表現不佳，因此希望可以設計出一個判斷牛熊市的因子。所以我們使用Attention機制取得訓練資料集的加權Target，搭配原作者訓練好的模型所產生的y predict，將這兩個features當作新的因子，放入一個NN model中去預測Target，希望能透過這種Attention機制讓模型多考慮牛熊市狀況。

**Data collection and validation :**

在競賽中，除了競賽期間給的training data：2018-01-01~2021-03-12及supplemental data：2021-03-12~2022-01-24

我們將兩者重疊的日期先處理掉，再合併兩個檔案為一個檔案，時間為：2018-01-01~2022-01-24

還另外去Binance(加密貨幣交易所)爬蟲抓了從2022-01-24~2022-12-5的data準備當作testing data

Data given:

Asset ID: 0~13 e.g. Bitcoin, ETH, Binance coin

timestamp: 時間(年月日)-> 整數 (每1分鐘一筆數據)

Open: 一分鐘的開盤價

Close: 一分鐘的收盤價

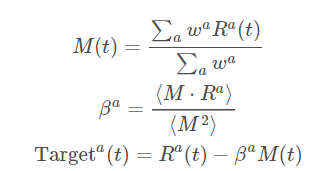
High: 一分鐘內的最高交易價格

Low: 一分鐘內的最低交易價格

Volume: 一分鐘內的交易量

Target : **預測目標y** (依照CAPM公式，這代表了單一資產未來15分鐘的獨有風險溢酬 )

(1)

(2)

R^a為Target，P^a(t)為時間t的收盤價，M(t)為市場風險溢酬，beta為市場與單一資場的相關係數

**Evaluation**

**-Metrics**

我們的評估方式是將我們predict 出來的 y跟 true y 去做 RMSE跟correlation。由於我們預測的數據是跳動幅度很大的並非線性，因此我覺得不太適合用correlation當作判斷標準，所以最後選擇RMSE，RMSE越小代表我們預測出來的結果跟真實資料越像。

### **-Time setting**

使用 2018-01-01 到 2022-12-05 之間所有Bitcoin跟 Ethereum的數據。我們將時間切成三種，來看各種模型在不同的時間下的預測能力。

1.預測幣災前

train & validation: 2018\_01\_01 ~ 2022-01-24

test: 2022-01-24 ~ 2022-05-24

2.預測幣災

train & validation: 2018\_01\_01 ~ 22022-01-24

test: 2022-05-24 ~ 2022-12-04

3. 加新資料預測幣災

train& validation: 2018\_01\_01~ 22022-05-24

test: 2022-05-24 ~ 2022-12-04

# 

(圖2)

**-Baselines**

**1. 原作者(lgb) 在 三個不同時間下的預測**

1.預測幣災前bit RMSE = 0.001403743

eth RMSE = 0.001525924

2.預測幣災 bit RMSE = 0.00148702

eth RMSE = 0.001799466

3. 加新資料預測幣災 bit RMSE = 0.00139661 eth RMSE = 0.001745011

這邊可以看到 model 1 在預測幣災時的能力變差，RMSE皆變大，但在新增了額外的data後，預測能力有變好。至於貨幣方面的預測是：比特幣表現比以太幣好

**Results**

**2. 原作者(lgb)＋新feature 在三個不同時間下的預測**

1.預測幣災前bit RMSE = 0.001402193 eth RMSE = 0.001525162

2.預測幣災 bit RMSE = 0.00146021 eth RMSE = 0.00176843

3. 加新資料預測幣災bit RMSE = 0.00139561 eth RMSE = 0.001743911

這邊可以看到 model 2 在預測幣災時的能力變差，RMSE皆變大，但在新增了額外的data後，預測能力有變好。至於貨幣方面的預測是：比特幣表現比以太幣好，結果與 model 1 相似。

**3. 原作者(lgb)＋新feature＋多model(lgb)**

1.預測幣災前 bit RMSE = 0.001401184 eth RMSE = 0.001521367

2.預測幣災 bit RMSE = 0.00145767 eth RMSE = 0.001766942

3. 加新資料預測幣災 bit RMSE = 0.00139176 eth RMSE = 0.001741172

一樣的可以看到 model 3 在預測幣災時的能力變差，RMSE皆變大，但在新增了額外的data後，能力有變好。至於貨幣方面的預測是：比特幣表現比以太幣好，結果與 model 1 和 model 2 相似

**4. 預測幣災前 比較3種 model**

1. model 1: 原作者(lgb)：bit RMSE = 0.001403743 eth RMSE = 0.001525924
2. model 2: 原作者(lgb)＋新feature： bit RMSE = 0.001402193 eth RMSE = 0.001525162
3. model 3: 原作者(lgb)＋新feature＋多model(lgb)： bit RMSE = 0.001401184 eth RMSE = 0.001521367

這邊可以看到3個model 在預測幣災的能力差不多，而在RMSE上model 3 < model 2 < model 1

。然後貨幣的預測是比特幣表現比以太幣好

**5. 預測幣災 比較3種 model**

1. model 1: 原作者(lgb)： bit RMSE = 0.00148702 eth RMSE = 0.001799466
2. model 2: 原作者(lgb)＋新feature：bit RMSE = 0.00146021 eth RMSE = 0.00176843
3. model 3: 原作者(lgb)＋新feature＋多model(lgb)： bit RMSE = 0.00145767 eth RMSE = 0.001766942

這邊可以看到3個model 在預測幣災時的能力差不多，而在 RMSE上model 3 < model 2 < model 1。至於貨幣的預測是：比特幣表現比以太幣好

**6. 加新資料預測幣災 比較3種 model**

1. model 1: 原作者(lgb)： bit RMSE = 0.00139661 eth RMSE = 0.001745011
2. model 2: 原作者(lgb)＋新feature：bit RMSE = 0.00139561 eth RMSE = 0.001743911
3. model 3: 原作者(lgb)＋新feature＋多model(lgb)：bit RMSE = 0.00139176 eth RMSE = 0.001741172

可以看到 三個model 在預測幣災的能力就有比較明顯一點的差別，RMSE上model 3 < model 2 < model 1。貨幣的預測是比特幣表現比以太幣好

以下是上述結果的整理表格

**Bitcoin**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 幣災前 | 幣災 | 加新資料預測幣災 |
| model 1 RMSE | 0.001403743 | 0.00148702 | 0.00139661 |
| model 2 RMSE | 0.001402193 | 0.00146021 | 0.00139561 |
| model 3 RMSE | 0.001401184 | 0.00145767 | 0.00139176 |

(表1)

**Ethereum**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 幣災前 | 幣災 | 加新資料預測幣災 |
| model 1 RMSE | 0.001525924 | 0.001799466 | 0.001745011 |
| model 2 RMSE | 0.001525162 | 0.00176843 | 0.001743911 |
| model 3 RMSE | 0.001521367 | 0.001766942 | 0.001741172 |

(表2)

1. 可以看到不管是什麼model跟時間點下，比特幣的預測表現都比以太幣好。
2. 我們也可以看到不管事什麼model在預測幣災前都差不多，但在幣災時表現都會變差一點，在加新資料預測幣災時表現都會變好。
3. 我們可以看到作者的model在三個時間下的表現比較差，加上feature後的model 2表現有明顯進步（RMSE變小），再加上新的model後的model 3表現也有進步（RMSE變小，但變小幅度比加feature時來得小）。

model 3中的額外加model除了上面顯示的LGB外我們其實還有做NN跟LSTM來當我們的疊加model，但由於NN跟LSTM的表現不理想因此並沒有秀出數據結果，但程式碼依然會附上。

**Conclusion:**

經過這次專案後，我們知道產生一個好的模型並不是一件容易的事情，即使我們絞盡腦汁做了很多個模型，也做了很多個feature，最後的成果也只有一些模型(model2, model 3)比原模型的結果好上一些，讓我們知道深度學習的困難之處。

另外我們也想到幾個方法，也許能夠提升預測能力，也許未來能夠嘗試：

**1.用全部14種幣來訓練模型**

由於資料量太過龐大，記憶體無法處理

我們只選出了其中兩種幣：比特幣 以及 以太幣，來train model，

但其實有可能其他貨幣也會影響比特幣和以太幣的走勢，但我們就沒有考慮到。

故之後的改進可能會想加入其他具代表性的Model進來train。

**2. 模型的改善**

聽完上次演講者所說再經過討論後，我們這邊想到了一些模型的改進方法

像是用CNN or RNN，

CNN用來做pattern recognition；

RNN可以加入time series, Autoregression跟LSTM，

多一些判斷趨勢的部分，來讓模型能夠學習得更完善。

**分工**

陳韋勳: 規劃專案執行架構(都在嘴組員)、爬取新資料並整理成競賽格式、協助處理胡祖望的程式碼問題(只負責看，胡祖望沒有感受到任何幫助)、2nd model + attention機制去提升原始模型解釋力(等報告結束了才做)

葉力嘉: Designing new feature、書面報告統整

胡祖望: 跑所有模型程式, 設計新疊加的2nd model(LSTM), 設計新疊加的2nd model(LGB)

黃元裕: 資料前處理, 設計新疊加的2nd model(NN)，結果數據分析

**貢獻度**

陳韋勳（貢獻5%）

葉力嘉、胡祖望、黃元裕（貢獻95%）

Reference:

[1] Kaggle 競賽網址： https://www.kaggle.com/competitions/g-research-crypto-forecasting/overview

[2] 第三名得獎模型參考：https://www.kaggle.com/competitions/g-research-crypto-forecasting/discussion/323703

[3] Manish Agrawal, Asif Ullah Khan, Piyush Kumar Shukla “Stock Price Prediction using Technical Indicators: A Predictive Model using Optimal Deep Learning” (IJRTE)(2019)

[4] Aryendra Singh, Priyanshi Gupta, Narina Thakur “An Empirical Research and Comprehensive Analysis of Stock Market Prediction using Machine Learning and Deep Learning techniques” (2021 IOP Conf. Ser.)