泡沫浮選雜質預測與分析

一、背景與動機

1.1泡沫浮選作業簡介

泡沫浮選是一種廣泛使用的選礦技術，用於分離有價值和不需要的礦物顆粒，而不同的礦物具有不同的物理化學性質，如親水性或疏水性，因此在泡沫浮選廠中通過線上(on-line)監測或估算即時的礦物品質和回收率對於工程師來說非常的重要，且對於控制和優化有著密切的關聯性。而線上監測所需要的不僅僅有昂貴且複雜的機器，更需要耗費整段流程的時間來得到最終的結果，其成本非常之高，因此若能在較前面的步驟及能知道最終結果，必能減少非常可觀的成本。

1.2浮選作業流程概述

此專案所分析的為在鐵礦分析工廠中的浮選流程，其主要目的為分離二氧化矽顆粒以得到更高純度的鐵礦。而在此浮選廠中，鐵漿(400噸/小時)被送入三個平行運行的浮選槽，而此浮選槽可以將具有高濃度的二氧化矽泡沫從表面去除，並且將含有大量鐵的沉澱物流入下一個槽中重複相同的流程，由Fig1可以看到，浮選槽1、2、3為第一層的浮選槽，而其輸出會分別進入浮選槽4、5、6，最終再匯入到浮選槽7，要得到最後的二氧化矽比例需要考慮的因素眾多，且擁有不確定的數學關係，如線性或非線性的關係，因此需要藉由機器學習的方式來找出隱藏在這些步驟中的數學關係，以此來預測最終二氧化矽的雜質比例。

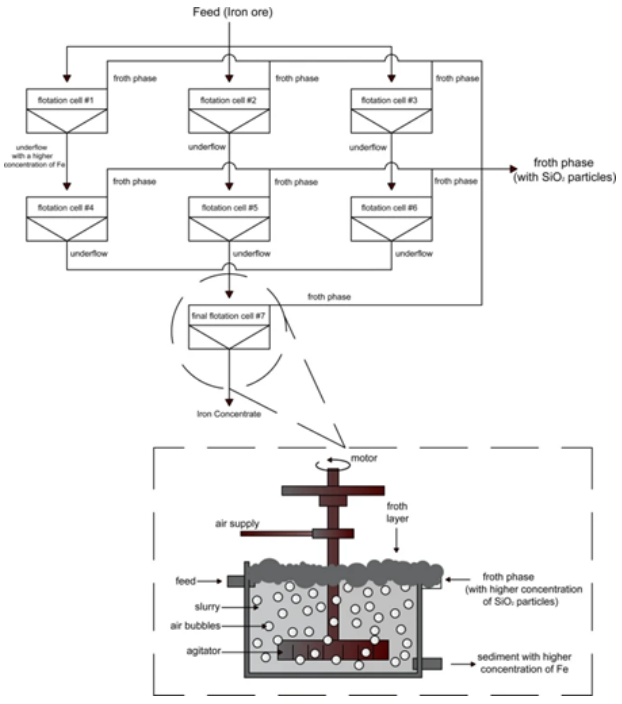


Fig1. 浮選廠運送流程圖from[1]

1.3問題定義

我們選用了在[1]中所提供的浮選廠數據做分析，我們的目標為用個歷史之環境與流程間的變量數據來預測二氧化矽雜質的含量(%Silica concentrate)，在得到含量後可以有相對應的方式調整參數以讓最終結果符合預期。

其預測的目的主要有4點:

1. 生產效率提升：透過浮選廠預測，可以更精確地估算出生產品質，並提前準備好相應的設備和人力，以便提升生產效率。

2. 成本控制:浮選廠預測可以幫助企業掌握市場變化，並提前準備好相應的生產計劃，以便控制成本。

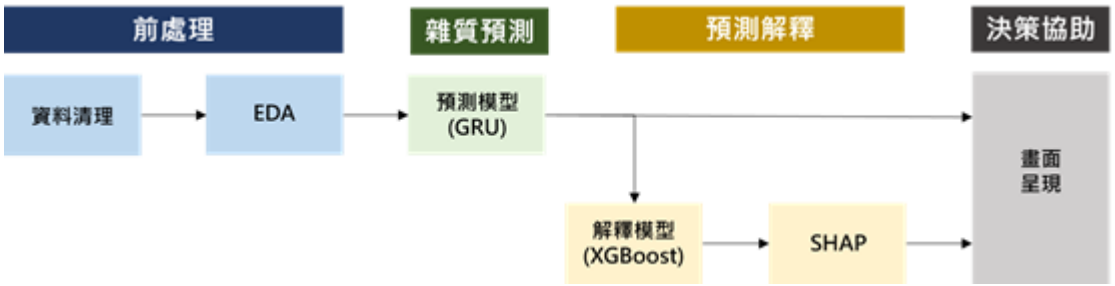
3. 品質控制:浮選廠預測可以幫助企業掌握產品品質，並提前準備好相應的生產計劃，以便控制產品品質。

4. 過程中調整:要是工程師透過我們所建立的模型提前知道了結果，那就可以提前對過程中的可控變量做調整，以此來將最後的品質調整為所期望的最終品質。

二、方法

2.1 整體流程架構

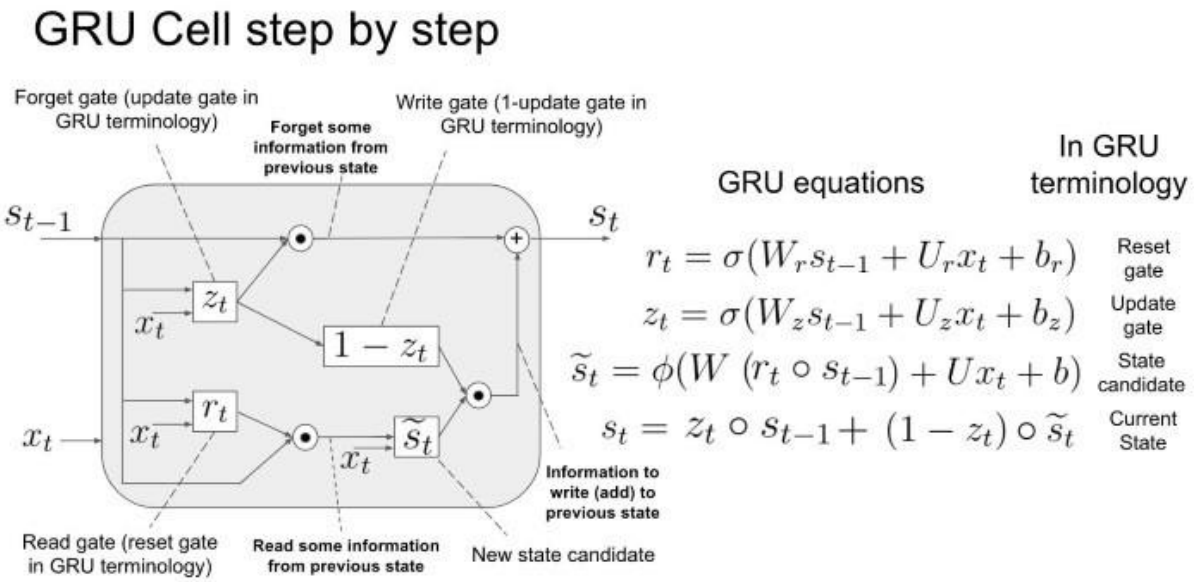
我們將處理的流程分為四個部分:分別為前處理、雜質預測、預測解釋、決策協助，如下圖。



- 前處理:首先會進行資料清洗，將資料的狀態清理成模型訓練前的狀況。之後則會進行EDA探索式的資料分析，讓我們了解各變數的關係與Y的分布等資訊。
- 雜質預測:為解決本專案欲解決的問題，我們使用考慮時間序列的深度學習模型GRU來進行下一小時二氧化矽雜質含量預測。
- 預測解釋:為提供工程人員預測值的解釋與提供補救措施的協助，我們將預測模型之結果值當作Y值，訓練一個較為簡單的模型做為解釋使用。並將該模型連接SHAP值，提供各樣本特徵貢獻值，協助工程人員調整各參數時的優先順序。
- 決策協助:結合預測值、解釋性模型訓練之重要變數趨勢、SHAP值，提供工程人員畫面做決策參考。

2.2 預測用模型介紹與合理性評估

有鑒於本次泡沫浮選資料集是帶有時間資訊的連續紀錄數據，我們認為使用考慮基於時間序列的模型能有效作為此次任務做預測。著名的時間序練模型有 RNN, LSTM, GRU。基於效能及效益，我們最後選用 GRU 作為此次深度模型的架構。



與 LSTM 同是基於 RNN 為變形，GRU 改良了 LSTM 中 output gate 和 forget gate 的設計。簡單來說，原本的 LSTM 模型需要基於此時此刻學習過往那些資料需要被紀錄及參考，而那些則可以被遺忘，GRU 將這兩個 gate 做合併，稱之為 update gate，他的概念即是只訓練原 LSTM 的 forget gate，ouput gate 則自動與 forget gate 的資訊互補。如此一來 GRU 就少了一個參數需要訓練，整體訓練效率也提升不少。

至於 GRU 模型的輸入資料，我們試了兩種

第一種選法是將 n-6 到 n-1 小時的所有資訊當作 feature，label 即是第 n 小時的二氧化矽的雜質比例，也就是我們期望末行最終能成功預測的目標。

第二種選法是將 n-6 到 n-1 小時的二氧化矽雜質比例當作 feature，label 即是第 n 小時的二氧化矽的雜質比例，我們想觀察前 6 小時的二氧化矽雜質有多大幅度的影響預測結果，且了解其他資訊的加入是否能真的增加模型預測的準確度。

2.3 解釋用模型介紹與合理性評估

首先介紹可解釋性人工智慧，簡稱XAI。主要目的是因為許多深度學習的模型都像黑盒子一樣，雖然可以有相當高的預測水平，所以預測上可以相當準確，但許多製造現場或是企業應用中會需要知道出現這個預測值的原因或是背後是哪些特徵的影響比較大，以協助進行流程

的控制與補救。所以可解釋性人工智慧，就是希望可以透過另一個較結構上簡單的模型或是其他統計方法去找到局部或是全部的特徵重要性。而我們在此處選擇使用Xgboost作為本次解釋用模型去擬合預測用模型的結果後，做整體特徵重要性的分析與後面單一樣本解釋之SHAP值背後的模型。

Xgboost為一boosting tree架構的模型，除了一般gradient boosting tree針對下一棵樹訓練是基於上一棵樹錯誤的修正特性外，它也會在每一棵樹進行訓練時隨機取不同的特徵而非全部特徵進行訓練。Xgboost作為regressior或是classifier在過去Kaggle的競賽中具有相當好的成效，應具有能力可以擬和預測用模型的結果。預測模型另外boosting tree的架構與特性，可以藉由計算每一個特徵作為節點的資訊增益去計算該特徵的重要性，也具有解釋性。

2.4 SHAP值介紹與合理性評估

除了用可解釋性模型得知平均而言各特徵與模型預測值之間的關係。我們也想知道模型如何預測樣本(local model-agnostic)。在本文，我們使用 SHAP來解釋單一樣本是如何被GRU預測。SHAP 是用來估計 Shapley values 的模型，而Shapley values 為賽局理論一種分配獎勵的方式，能透過每個參與者(變數)對結果(GRU預測值)的貢獻，來計算各參與者的獎勵。我們可以把每個特徵的 Shapley value 視為這個特徵對最終預測值的貢獻程度。透過 SHAP，現場的工作人員能理解複雜模型所捕捉到的資訊並將其利用再提高產能或輸出品質。

三、資料蒐集

3.1 原始資料集簡述

資料集取得於Kaggle上的Quality Prediction in a Mining Process的資料集[2]，裡面包含時間跨度2017/03/10到2017/09/09泡沫浮選廠中鐵礦浮選流程中輸入到輸出，多數為每20秒產生一筆相關資訊，共24欄737,453筆。

3.2 預處理流程

- 資料補值

  由於每筆資料為每20秒一筆，所以基本上每一個小時的資訊為180筆。而其中<日期時間>中未滿180筆，在此我們假設為該時間區段中前面的時間缺值，並利用backfill的方式，使用最近後面時間的資料補上缺失的時間區段。

- 資料時間標註

  資料時間欄位的資料同一小時的資料標註時間皆相同，並沒有切割確切時間。所以我們以原始資料的順序進行標註，增加時間欄位datetime，讓每筆資料有獨立的時間標註。

- 調整問題Y值

  原始資料集在Kaggle的敘述中表示Y值也就是二氧化矽雜質為每一小時由實驗室出來的數據，所以同一小時的資訊應相同。然而有部分時間雜質比例確實不是如此，為

20秒變動一次，我們懷疑為資料提供時有誤，所以我們將這些每小時內二氧化矽雜質不一致的資料進行每小時一次的平均，作為該小時代表之Y值。

- 以每小時為單位作為模型訓練樣本

  為提供浮選廠預測未來一小時的二氧化矽雜質比例， 我們將每20秒一筆的資訊取其第一筆資訊(即datetime分、秒部分為00:00者)當作代表該小時的資訊。選擇該取法的原因為我們希望能讓使用者在每小時最開始時拿到的資訊就與過去的歷史資訊即可以預測下一個小時的雜質。

3.3 乾淨資料集簡述

清理後的資料集，4097筆 25欄，欄位的資訊如下：

原始取樣類型	分類	數量	意義	欄位名
每1小時取樣	原料	2	原料輸入的比例	% Iron Feed % Silica Feed
	輸出	2	產品輸出成分比例(Y) → 實驗室取得	% Iron Concentrate % Silica Concentrate
	時間	2	資料紀錄之時間	date、datetime
每20秒取樣	環境參數	5	製造環境整體狀況變數	Starch Flow、Amina Flow、Ore Pulp pH…
	流程參數	14	各泡沫浮選機台狀況變數(共7台)	Flotation Column 01 Air Flow、Flotation Column 01 Level…

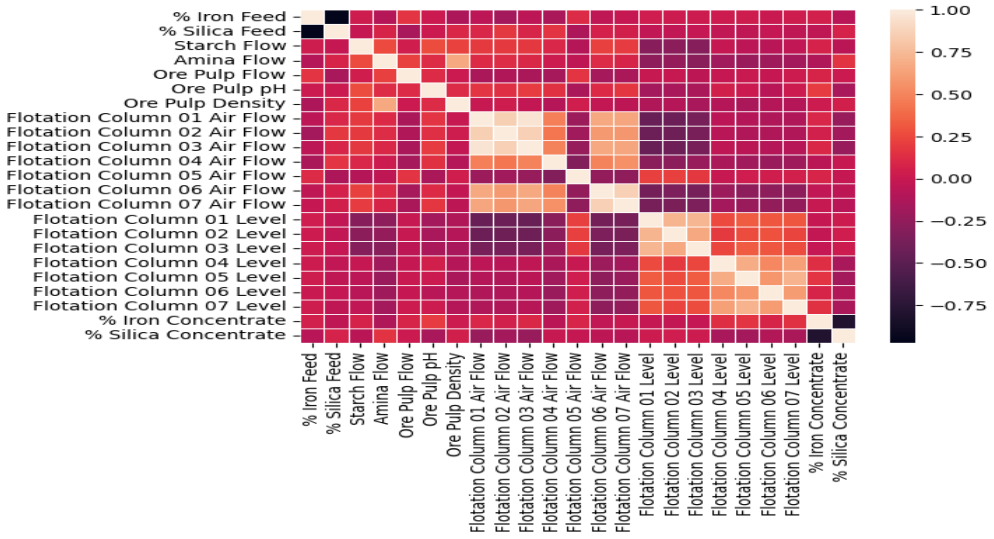
3.3 EDA

3.3.1資料間的共線性問題

i. 變數相關性視覺化:

由皮爾森相關係數初步觀察與判定不同變數之間的相關性是如何，如下相關性熱

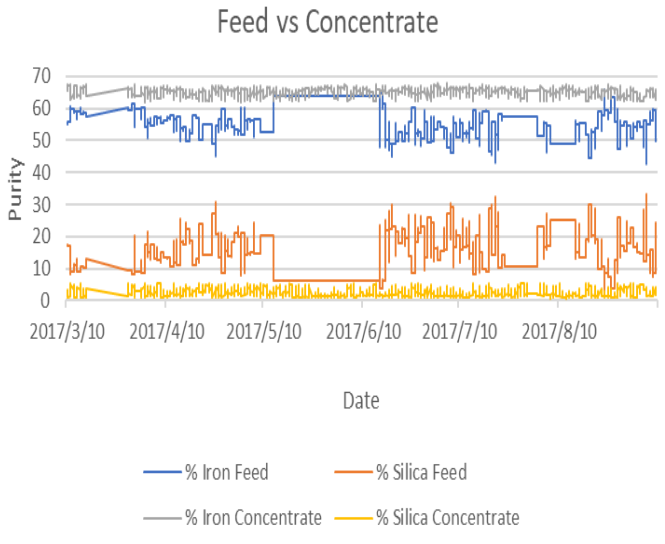
力圖可以發現，Iron Feed&Silica Feed、Air Flow (1,2,3),(6,7) 、Level(1,2,3),(4,5,6,7)和Iron Concentrate&Silica Concentrate之間有高度的線性關係，如Air Flow (1,2,3)為原物料分配到平行的三個第一層浮選槽，他們有高度的相關性可以讓我們知道這個數據集的合理程度，因為他們皆未受到加工，因此他們在剛進入浮選槽有相同的特性是正確的。



ii.由變異數膨脹因子(Variance Inflation Factor, VIF)診斷共線性問題:

變異數膨脹因子為判別此變數是否有共線性問題的數值，通常超過10即判別此變數和其他變數擁有高度的線性關係，可能是其他變數的倍數或是可以以線性組合的方式用其他變數代表此VIF很高之變數，由下表可以看到IronFeed&Silicon Feed和Air Flow1&Air Flow3擁有很高的VIF數值，有上面的相關係數熱力圖可以得知他們VIF數值高的問題，應該就是浮選槽第一層的影響還有輸入原物料雜質程度(鐵含量和二氧化矽含量成高度反比)間的高相關性所導致的，但因為我們的數據數量相較於特徵來說是足夠的，我們認為這些輕微共線性的狀況是可以接受的，因此我們最後沒有做特徵選擇一方面也讓我們可以有更完善的特徵解釋性。(見下左)

variable	vif
% Iron Feed	19.382871
% Silica Feed	19.190670
Starch Flow	1.264975
Amina Flow	2.250004
Ore Pulp Flow	1.257789
Ore Pulp pH	1.230154
Ore Pulp Density	2.040420
Flotation Column 01 Air Flow	12.387707
Flotation Column 02 Air Flow	4.355539
Flotation Column 03 Air Flow	13.831862
Flotation Column 04 Air Flow	1.680906
Flotation Column 05 Air Flow	1.228301
Flotation Column 06 Air Flow	4.371520
Flotation Column 07 Air Flow	4.281479
Flotation Column 01 Level	2.918093
Flotation Column 02 Level	2.362083
Flotation Column 03 Level	2.436730
Flotation Column 04 Level	2.070157
Flotation Column 05 Level	2.733284
Flotation Column 06 Level	1.816860
Flotation Column 07 Level	2.457509
% Iron Concentrate	3.081682
% Silica Concentrate	3.129639



將數據以時間做排序並以圖形化顯示(見上右)，可以發現有以下兩個性質:

i.輸出純度不單單只受輸入純度影響

由5月至6月的數據可以發現雖然輸入原物料的純度沒有波動，但輸出的純度卻一直在改變，有此特性可以發現，輸出的鐵礦不單單只受到原物料的影響，也有其他變數會影響最終的輸出純度如環境變數等等。

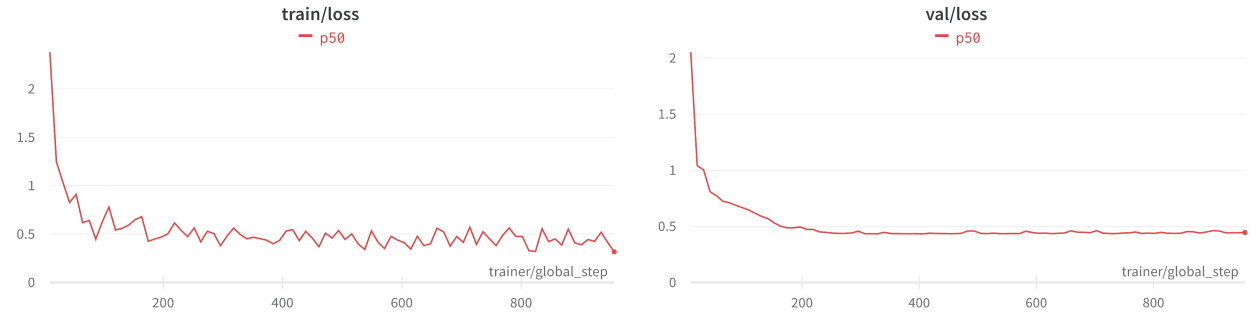
ii.輸出純度的暫存性

由7月至8月的數據可以發現當輸入原物料穩定後，但輸出並不會跟著穩定，而是會過一陣子才跟著穩定，這代表輸出的鐵礦純度有一定的暫存性，並不會因應變數的變化而有立即的波動。

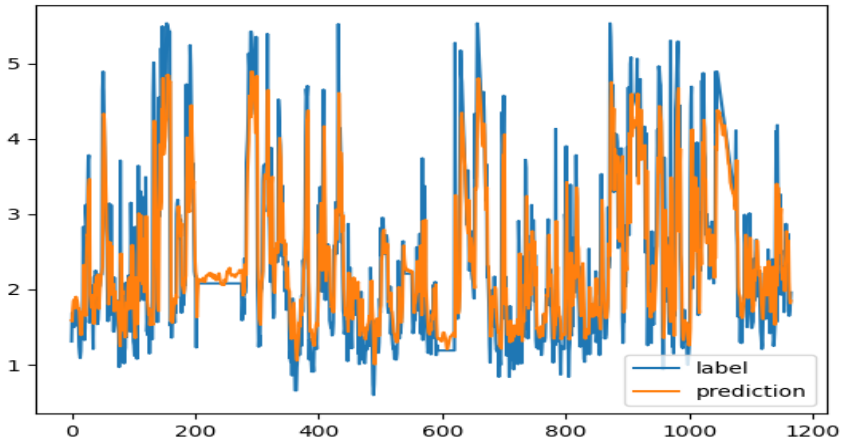
四、結果

4.1 預測模型訓練與結果

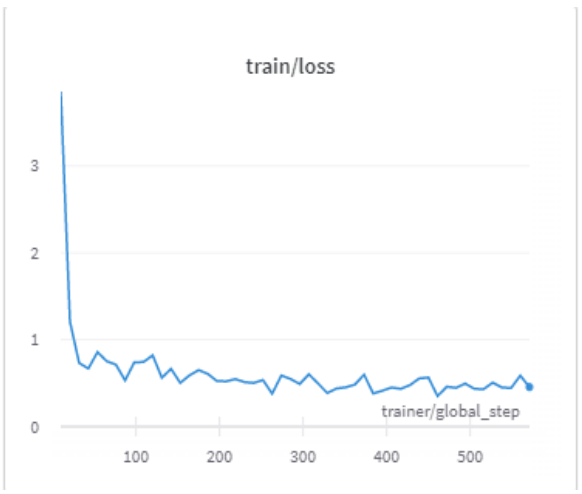
GRU 基於第一種資料即作為訓練集的結果:

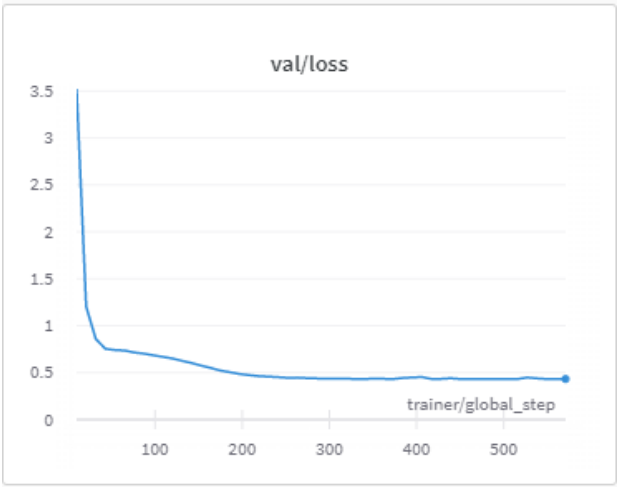


比對結果:　validation mAE: 0.442

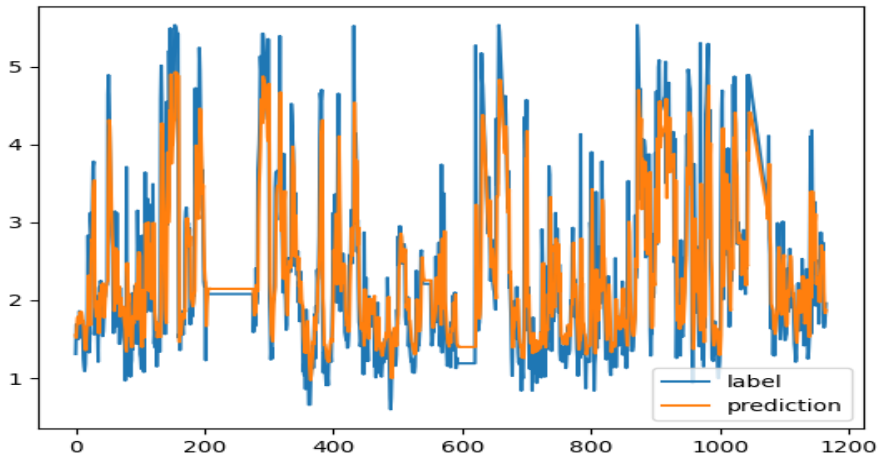


GRU 基於第二種資料即作為訓練集的結果:





比對結果:　mAE: 0.432



可見其實前 6 小時的二氧化矽雜質比例dominate整個深度學習的訓練過程，在 validation loss的數值上也可以發現存二氧化矽雜質比例數值相去不多，收斂的也越快。我們分析很有可能在這麼模型中其他 feature 能提供的幫助有限，模型還是大部分都依據過往的二氧化矽雜質比例來做預測。

4.2 解釋用模型訓練與結果

儘管二氧化矽雜質比例主要影響Y值的預測，但因解釋用模型需要協助補救性調整，所以我們仍使用其他的環境參數等進行預測，才可以提供工程師針對較重要的環境與流程參數做調整。另外，因預測用模型GRU取前六小時作為X值預測下一小時的Y，所以解釋用模型的部分在訓練前做了特徵工程，將各變數取前2~6小時進行rolling window的平均，並命名為X變數名_n(n為window size)。之後同預測用模型取前面7成作為訓練資料剩下3成為測試資料。結果如圖一，可以發現儘管成效佳，但重要變數皆為歷史的二氧化矽雜質(% Silica Concentrate_n)，若太依賴歷史的二氧化矽雜質做訓練則無法有效提供補救性措施調整參數的依據， 所以之後則嘗試僅留下前兩小時的二氧化矽雜質作為代表訓練，結果如以下:

i. 訓練與測試結果

	全部特徵 **(Train/test)**	僅留下前兩小時二氧化矽雜質 **(Train/test)**
RMSE	0.006 / 0.09	0.10 / 0.27
MAPE	0.002 / 0.03	0.02 / 0.06
R Square	0.99 / 0.98	0.98 / 0.90
重要變數	% Silica Concentrate % Silica Concentrate_2 % Silica Concentrate_3 Flotation Column 03 Air Flow_4	% Silica Concentrate_2 Flotation Column 06 Level_6 Flotation Column 03 Air Flow_4 Flotation Column 02 Air Flow_6

ii. 訓練時各Fold 驗證

	全部特徵**(RMSE)**	僅留下前兩小時二氧化矽雜質 **(RMSE)**
Fold 1 validation	0.205	0.618
Fold 2 validation	0.069	0.235
Fold 3 validation	0.044	0.170
Fold 4 validation	0.043	0.213
Fold 5 validation	0.051	0.176

註：切fold採取TimeSeriesSplit的方式，做法是每fold的training data永遠早於validation data，第一fold的training data較少，最後一fold的最多，所以validation的RMSE也逐漸降低。

iii. 重要變數分析

下表為最後解釋性模型的各變數平均增益，大到小排序：

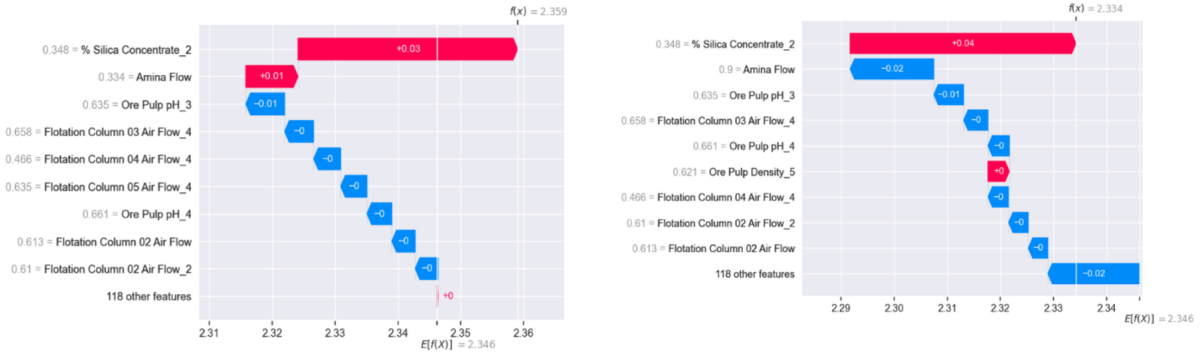
Feature	Importance
% Silica Concentrate_2	0.496738
Flotation Column 06 Level_6	0.010676
Flotation Column 03 Air Flow_4	0.008356
Flotation Column 02 Level_6	0.007359
Flotation Column 01 Air Flow_5	0.007244
Flotation Column 01 Air Flow_2	0.007211
Flotation Column 06 Air Flow_6	0.006742
Flotation Column 06 Air Flow_3	0.006641
Flotation Column 03 Air Flow_5	0.006405
Flotation Column 03 Air Flow_6	0.006167

基於上表的結果，有以下幾點觀察：

- 過去% Silica Concentrate的值仍然為重要預測特徵
- 各浮選槽中6號特別關鍵(氣流與水位皆為重要特徵)
- 流程變數重要程度大於環境變數
- 關於rolling window，時間跨度大的變數重要程度排序較跨度小前

4.3 SHAP結果與補救性調整案例

下圖，為將單一樣本輸入SHAP並進行補救性調整的案例。



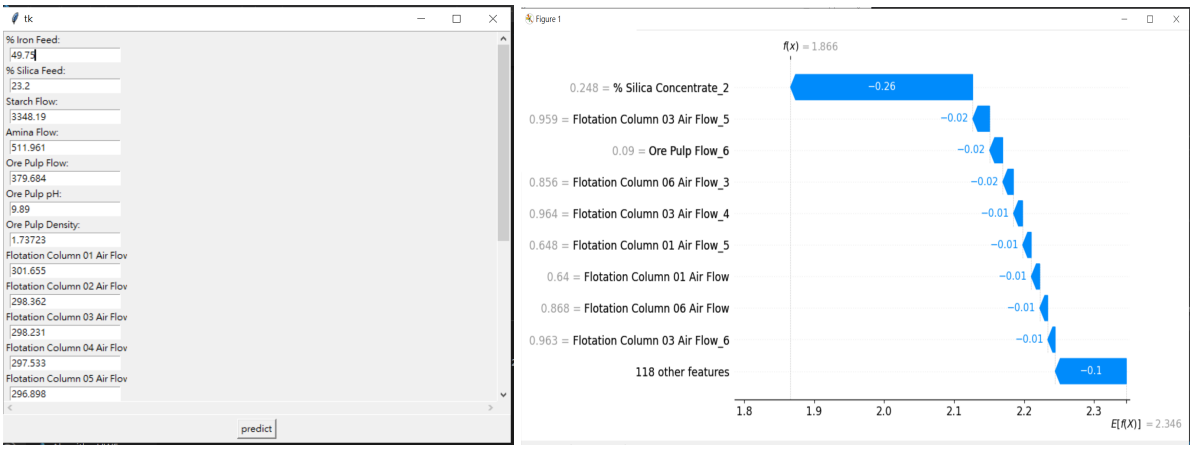
從圖，我們可以看到單一樣本是因為什麼特徵的緣故，而與樣本預測平均有所差異。製造現場人員能依照各變數對預測值的正負貢獻來決定調整現場的製造參數。舉例來說，我們可以看到當 Anima Flow 從 0.334 (左圖) 增為 0.9 (右圖)，預測值將降低至 2.334；如此，人員便能用自身的知識判斷是否能透過此資訊有效的將低雜質濃度。

雖然 SHAP 能直觀的告訴工程師現在哪個變數能夠最有效降低預測值，但其缺點為需要多次調整樣本和重複模型預測，相當耗時。因此尚須現場人員已領域背景知識挑出具代表性的樣本來跑 SHAP才能發揮其效用 (代表性樣本也可透過分析樣本分布獲取，如 mmd-critic) 。

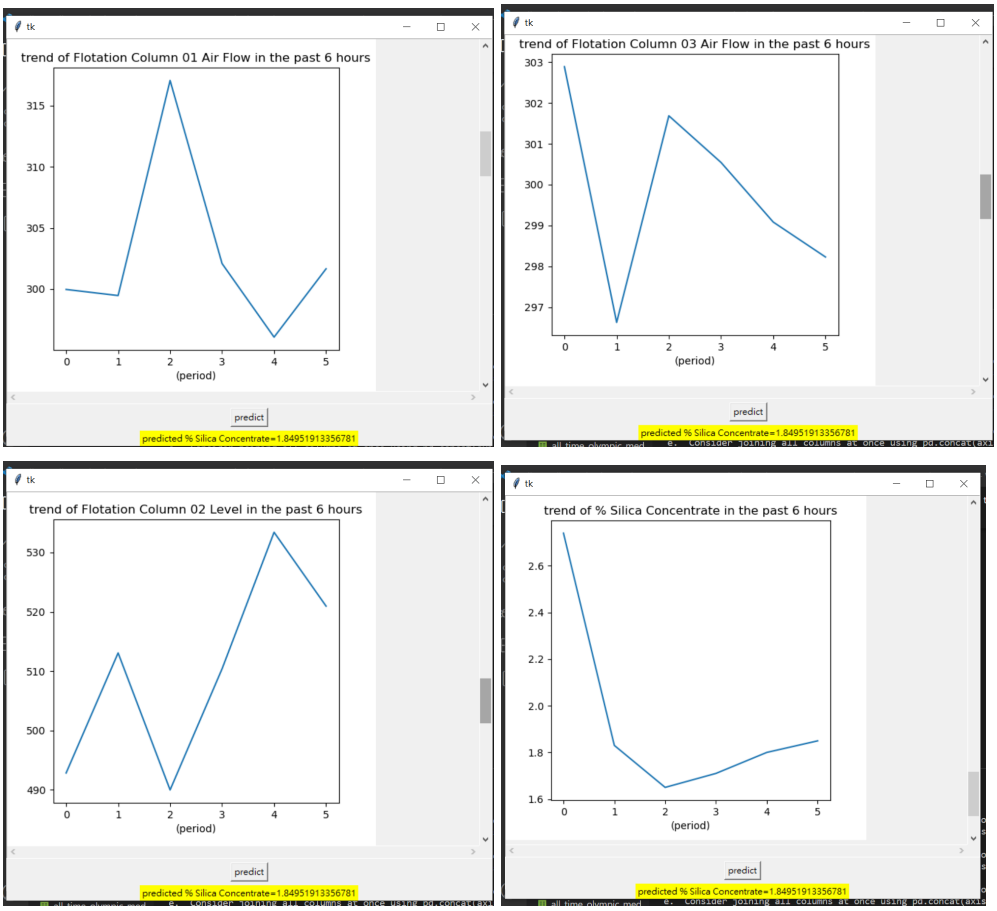
4.4 預測介面功能與成果展示

i. 輸入資料

ii.按下predict鍵後會呈現輸入資料的shap值，讓工程師了解每個特徵的貢獻程度以作為後續調整的參考



iii. 黃色框框內為下一小時% Silica Concentrate的預測值，並且畫出前四個重要變數在過去六小時的趨勢圖以供參考。



2.設計說明

i. 使用套件：此次呈現主要使用的模組為tkinter和matplotlib，並且使用xgboost、shap等模組作背後的計算

ii. 取得輸入：使用者輸入資料並且按下predict鍵後，程式會抓取個特徵的數值

iii. 預測值：將使用者輸入的資料載入GRU模型，並預測下一小時的% Silica Concentrate值(由於系統相容性問題無法再入訓練好的GRU模型，因此默認的預測值為範例資料的預測值)

iv. shap瀑布圖：以rescale後的使用者輸入資料和訓練好的XGboost解釋性模型為依據，繪製含有各特徵shap值的瀑布圖

v. 趨勢圖：將使用者輸入資料整合歷史資料庫，並且針對重要變數繪製過去六小時的趨勢圖

Reference

[1]Pu, Yuanyuan, Alicja Szmigiel, and Derek B. Apel. "Purities prediction in a manufacturing froth flotation plant: the deep learning techniques." *Neural Computing and Applications* 32.17 (2020): 13639-13649.

[2] EDUARDOMAGALHÃESOLIVEIRA. “Quality Prediction in a Mining Process” Kaggle 2017