團隊測驗報告

報名序號：112009

團隊名稱：慢靈魂

一、資料前處理 – 異常值修正

- 原始資料集內有2055年及2002年年分，與附近資料年份相距甚遠
- 經說明會表示這是誤植，我們團隊以上下資料推定正確年份應是2022年，先對資料集的年份做修正

2022/5/26	2B0	5	34	6272	1
2055/5/27	1B0	4	107	7254	1
2055/5/27	1C0	8	58	7531	1
2055/5/27	1C0	13	58	7371	1
2055/5/27	1C0	19	105	7224	1
2055/5/27	1.00E+00	5	102	6651	1
2055/5/27	1.00E+00	11	8	6171	1
2055/5/27	2C0	16	59	5752	1
2055/5/27	2D0	19	75_77_92_	5299	5
2022/5/28	1B0	4	57	7260	1
2022/5/28	1B0	9	72_73	7867	2
2022/5/28	1C0	1	46	7079	1
2022/5/28	1C0	9	38_79	7540	2

2022/5/13	2B0	8	97_115	5845	2
2022/5/13	2C0	5	18_89	5772	2
2002/5/14	1C0	1	5_96	6827	2
2002/5/14	1B0	14	18_20_37_	7512	4
2002/5/14	1.00E+00	4	36_62	6238	2
2022/5/15	1B0	4	109	7064	1
2022/5/15	1B0	11	23_39_70	7705	3
2022/5/15	1B0	12	36_107	7668	2

一、資料前處理 – 資料分割

- 以oven 2為例，將資料集分為 3 部分，訓練、驗證、測試分別為365、22、22天。

- 驗證集主要用於移動平均法和週期移動平均法，除了這兩種模型外，在其他模型中驗證集皆併入訓練集一起訓練。

二、演算法和模型介紹 – 移動平均法

- 首先使用驗證集決定最佳的窗口大小（window size）。

- 將測試集最後與最佳窗口大小相同的資料做平均，做為下一筆資料的預測值，以此類推預測後面幾天資料並與測試集做比較。

二、演算法和模型介紹 – 週期移動平均法

- 首先使用驗證集決定最佳的週期(period)和最佳窗口大小（window size）。

- 如右圖所示，接著將原本資料集整理為最佳週期(period)後，並設定最佳窗口大小（window size）將最後幾期資料的平均作為下一期資料預測值，視測試集大小看要預測幾期數值。

Date	Anomaly number	Prediction
12/27 12/28	N1	-
12/29 12/30	N2	-
12/30 12/31	N3	-
1/1 1/2	N4	P1=(N1+N2+N3)/3
1/3 1/4	N5	P2=(N2+N3+N4)/3
. . .	. . .	. . .
6/1 6/2	N79	P76=(N76+N77+N78)/3
6/3 6/4		P77=(N77+N78+N79)/3
6/5 6/6		P78=(N78+N79+P77)/3
. . .		. . .
6/29 6/30		P90 = (P87+P88+P89)/3

二、演算法和模型介紹 – 雙指數平滑法（Holt’ s model）

- Holt’s model 又稱雙指數平滑法，主要用於處理具有線性趨勢的時間序列數據，即數據在一個持續的方向上呈現增加或減少。

- 通常用於較短期的預測，並且假設數據的趨勢和基本線在未來一段時間內保持不變。

- 將訓練集和驗證集合併，一起投入 Holt’s model 訓練，並預測與測試集比較。

二、演算法和模型介紹 – 三指數平滑法（holt-winters）

- Holt-Winters 又稱三指數平滑法，模型適用於具有趨勢和季節性的數據，並且可以用於中長期的預測。

- 將訓練集和驗證集合併，一起投入 Holt’s model 訓練，並預測與測試集比較。

二、演算法和模型介紹 – ARIMA

- ARIMA 模型（Autoregressive Integrated Moving Average Model）是一種常用於時間序列預測和分析的統計模型。它結合了自回歸（AR）模型、差分（I）運算和移動平均（MA）模型，用於捕捉時間序列數據中的自相關、趨勢和隨機性成分。

- ARIMA模型適用於各種時間序列數據，包括具有趨勢、季節性和隨機性的數據。

- 將訓練集和驗證集合併，一起投入 Holt’s model 訓練，並預測與測試集比較。

二、演算法和模型介紹 – SARIMA

- SARIMA模型（Seasonal Autoregressive Integrated Moving Average Model）是ARIMA模型的擴展，專門用於處理具有季節性的時間序列數據。它結合了自回歸（AR）、差分（I）運算和移動平均（MA）模型，同時考慮了數據的季節性成分。

- 將訓練集和驗證集合併，一起投入 Holt’s model 訓練，並預測與測試集比較。

二、監督式學習

- 使用變數(以oven1為例)

特徵名稱	說明
number_sum	前27天異常燈管總數
number_max	前27天異常燈管最大值
number_min	前27天異常燈管最小值
number_mode	前27天異常燈管眾數
days	前27天異常燈管出現天數
oven_encoder	前27天的平均損壞數量
cooler_max	前27天水冷板溫度最大值
avg_accumulation	前27天累積時數平均
max_accumulation	前27天累積時數最大值
min_accumulation	前27天累積時數最小值
avg_power_setup	前27天對應累積時數平均的功率輸出設定

二、監督式學習

- 以oven1為例，在新增未來27天的新特徵後進行標準化
- 由於模型計算上可能會進行特徵間的互相比較，因此先對所有特徵進行標準化，如此才能正確比較各特徵的影響力
- 接著投入Lasso regression、XGBoost、LightGBM中預測未來27天可能的anomaly_total_number並加總

二、監督式學習 – Lasso regression

- Lasso regression（L1 正則化線性回歸）是一種線性回歸模型，它在擬合數據時除了最小化成本函數外，還對模型的權重進行正則化。

- Lasso 正則化的主要效果是使得模型的某些權重變為零，進行特徵選擇，從而簡化模型並避免過度擬合。因此，它通常用於高維數據集或具有多個特徵的數據集，以選擇最重要的特徵。

二、監督式學習 – XGBoost

- XGBoost 是一種集成學習算法，通過組合多個弱學習器（通常是決策樹）來建立一個強大的模型。

- 另外， XGBoost可通過梯度提升來逐步改善模型的預測能力，它以分類錯誤（分類任務）或預測誤差（回歸任務）為損失函數，不斷優化模型，使其在每一輪迭代中都更接近實際值。

二、監督式學習 – LightGBM

- LightGBM（Light Gradient Boosting Machine）是一個高效的梯度提升樹（Gradient Boosting Tree）框架，專為處理大規模數據集和高維特徵而設計。

- 另外，LightGBM使用基於直方圖的算法來加快訓練過程，並且能夠充分利用多核 CPU 資源，是一種高效的機器學習算法，具有出色的訓練和預測速度。

二、監督式學習 – CatBoost

- CatBoost 的名稱中的 “Cat” 指的是分類特徵的處理，而 “Boost” 是梯度提升決策樹算法的一部分。CatBoost 已經在許多數據科学競賽和實際應用中取得了顯著的成功，特別是在需要處理大量分類特徵的問題上。

- CatBoost 可以自動選擇合適的超參數，並在大規模數據集上訓練快速而高效的模型。

三、預測結果 – OVEN1

- 以資料集的後27天作為測試集進行預測

	Actual number	Holts	Holt-Winters	MA	Period MA	ARIMA	SARIMA
1B0	58	49	46	75	42	43	39
1C0	38	49	48	76	81	34	45
1D0	42	55	54	56	68	49	67
1E0	26	48	47	49	56	42	62
1G0	0	2	2	0	0	2	3
RMSE		13.1072	12.9074	22.1721	27.1330	10.4881	21.6333

三、預測結果 – OVEN1

- 以資料集的後27天作為測試集進行預測

	Actual number	Lasso	LGBM	XGB	CatBoost
1B0	58	49	41	35	54
1C0	38	43	34	29	36
1D0	42	48	36	33	39
1E0	26	38	30	21	29
1G0	0	0	0	0	0
RMSE		7.5631	8.4499	11.9666	2.7568

三、預測結果 – OVEN2

- 以資料集的後22天作為測試集進行預測

	Actual number	Holts	Holt-Winters	MA	Period MA	ARIMA	SARIMA
2B0	53	50	52	65	53	52	44
2C0	51	44	45	53	60	45	36
2D0	50	45	45	47	23	40	41
2E0	39	41	42	49	74	52	56
2G0	0	1	1	0	0	1	1
RMSE		4.1952	3.7947	7.1694	20.1742	7.8358	11.6362

三、預測結果 – OVEN2

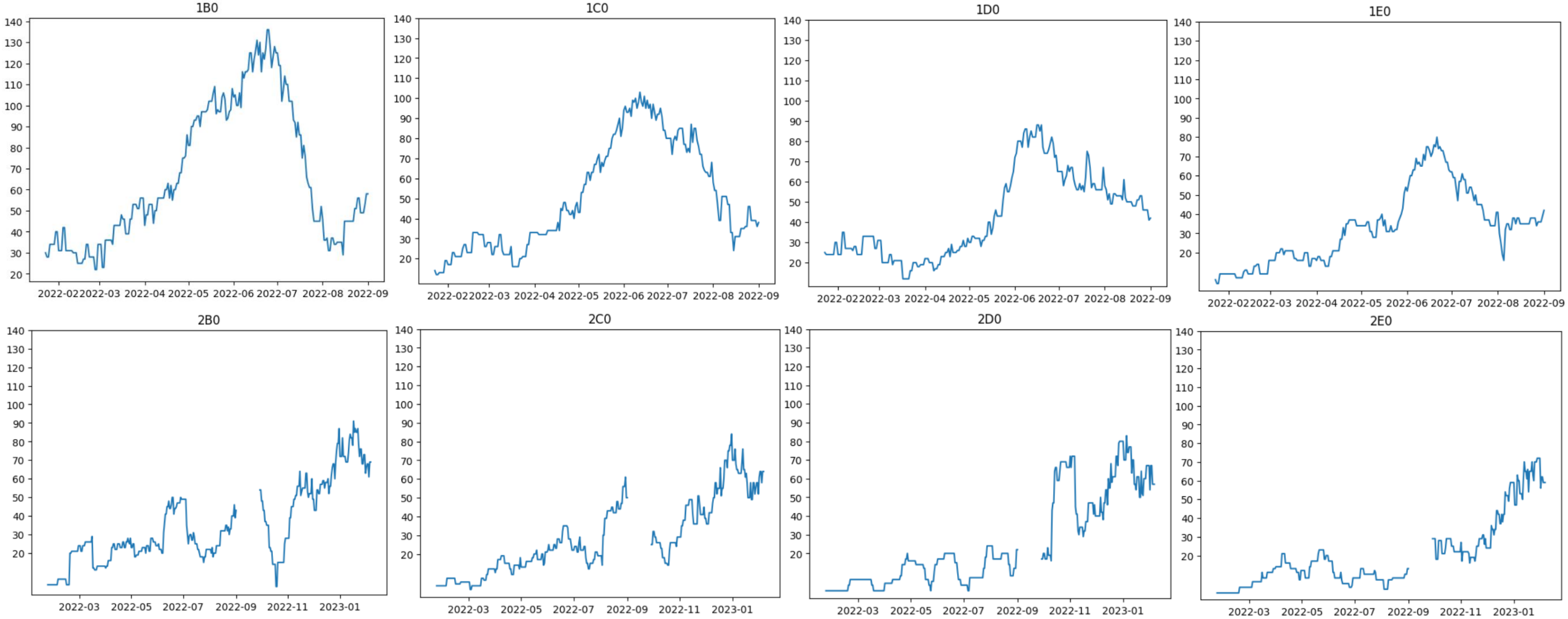
- 以資料集的後22天作為測試集進行預測

	Actual number	Lasso	LGBM	XGB	CatBoost
2B0	53	45	33	22	23
2C0	51	39	33	29	34
2D0	50	36	34	37	60
2E0	39	37	32	26	51
2G0	0	0	0	0	0
RMSE		9.0333	13.3457	18.8839	12.9293

四、資料分析與決策 – 分析一

- 此次資料有其特殊性，第一是爐層若累積使用時數達10000小時即翻新，而剛翻新的爐層幾乎不會出現異常
- 第二，每筆資料都與先前有關，不會短時間內起起伏伏
- 因此我們在選擇預測值時，主要以折線圖的趨勢以及未來爐層的狀況做評估為主，模型測試集RMSE為輔

四、資料分析與決策 – 異常曲線圖



各爐每一週期累積異常燈管數 (以預測天數為一週期)

各爐各層最後一日的累積時數

剛翻新爐層

即將翻新爐層

2022/9/1	1B0	1	9367
2022/9/1	1B0	2	8140
2022/9/1	1B0	3	9430
2022/9/1	1B0	4	8425
2022/9/1	1B0	5	9027
2022/9/1	1B0	6	9389
2022/9/1	1B0	7	9061
2022/9/1	1B0	8	9166
2022/9/1	1B0	9	9052
2022/9/1	1B0	10	9274
2022/9/1	1B0	11	9130
2022/9/1	1B0	12	9120
2022/9/1	1B0	13	9234
2022/9/1	1B0	14	8993
2022/9/1	1B0	15	8996
2022/9/1	1B0	16	9050
2022/9/1	1B0	17	9061
2022/9/1	1B0	18	8907
2022/9/1	1B0	19	4435

2022/9/1	1C0	1	8185
2022/9/1	1C0	2	556
2022/9/1	1C0	3	347
2022/9/1	1C0	4	8849
2022/9/1	1C0	5	539
2022/9/1	1C0	6	8698
2022/9/1	1C0	7	8883
2022/9/1	1C0	8	283
2022/9/1	1C0	9	8602
2022/9/1	1C0	10	8495
2022/9/1	1C0	11	8346
2022/9/1	1C0	12	8644
2022/9/1	1C0	13	8495
2022/9/1	1C0	14	8511
2022/9/1	1C0	15	150
2022/9/1	1C0	16	8419
2022/9/1	1C0	17	8370
2022/9/1	1C0	18	8312
2022/9/1	1C0	19	8322

2022/9/1	1D0	1	8003
2022/9/1	1D0	2	8391
2022/9/1	1D0	3	8507
2022/9/1	1D0	4	8324
2022/9/1	1D0	5	8329
2022/9/1	1D0	6	8210
2022/9/1	1D0	7	8325
2022/9/1	1D0	8	7614
2022/9/1	1D0	9	8147
2022/9/1	1D0	10	8048
2022/9/1	1D0	11	8012
2022/9/1	1D0	12	7829
2022/9/1	1D0	13	7810
2022/9/1	1D0	14	7760
2022/9/1	1D0	15	7711
2022/9/1	1D0	16	7608
2022/9/1	1D0	17	7346
2022/9/1	1D0	18	7598
2022/9/1	1D0	19	7394

2022/9/1	1.00E+00	1	7717
2022/9/1	1.00E+00	2	7844
2022/9/1	1.00E+00	3	7743
2022/9/1	1.00E+00	4	7580
2022/9/1	1.00E+00	5	7689
2022/9/1	1.00E+00	6	7532
2022/9/1	1.00E+00	7	7470
2022/9/1	1.00E+00	8	7393
2022/9/1	1.00E+00	9	7404
2022/9/1	1.00E+00	10	7376
2022/9/1	1.00E+00	11	7163
2022/9/1	1.00E+00	12	7052
2022/9/1	1.00E+00	13	6943
2022/9/1	1.00E+00	14	6926
2022/9/1	1.00E+00	15	6406
2022/9/1	1.00E+00	16	6543
2022/9/1	1.00E+00	17	6926
2022/9/1	1.00E+00	18	6671
2022/9/1	1.00E+00	19	6799

2023/2/6	2B0	1	1018
2023/2/6	2B0	2	994
2023/2/6	2B0	3	1011
2023/2/6	2B0	4	533
2023/2/6	2B0	5	565
2023/2/6	2B0	6	9591
2023/2/6	2B0	7	21
2023/2/6	2B0	8	0
2023/2/6	2B0	9	8265
2023/2/6	2B0	10	0
2023/2/6	2B0	11	9384
2023/2/6	2B0	12	9837
2023/2/6	2B0	13	9443
2023/2/6	2B0	14	9805
2023/2/6	2B0	15	8337
2023/2/6	2B0	16	9568
2023/2/6	2B0	17	9846
2023/2/6	2B0	18	0
2023/2/6	2B0	19	9791

2023/2/6	2C0	1	515
2023/2/6	2C0	2	9351
2023/2/6	2C0	3	512
2023/2/6	2C0	4	8960
2023/2/6	2C0	5	9162
2023/2/6	2C0	6	22
2023/2/6	2C0	7	10206
2023/2/6	2C0	8	9927
2023/2/6	2C0	9	9930
2023/2/6	2C0	10	9729
2023/2/6	2C0	11	9798
2023/2/6	2C0	12	9824
2023/2/6	2C0	13	0
2023/2/6	2C0	14	9746
2023/2/6	2C0	15	7371
2023/2/6	2C0	16	7569
2023/2/6	2C0	17	9706
2023/2/6	2C0	18	9704
2023/2/6	2C0	19	9514

2023/2/6	2D0	1	9989
2023/2/6	2D0	2	9861
2023/2/6	2D0	3	9825
2023/2/6	2D0	4	9658
2023/2/6	2D0	5	9677
2023/2/6	2D0	6	9801
2023/2/6	2D0	7	9472
2023/2/6	2D0	8	9432
2023/2/6	2D0	9	9500
2023/2/6	2D0	10	9109
2023/2/6	2D0	11	8451
2023/2/6	2D0	12	9257
2023/2/6	2D0	13	9127
2023/2/6	2D0	14	8000
2023/2/6	2D0	15	8827
2023/2/6	2D0	16	8956
2023/2/6	2D0	17	8769
2023/2/6	2D0	18	8118
2023/2/6	2D0	19	8830

2023/2/6	2.00E+00	1	8713
2023/2/6	2.00E+00	2	8498
2023/2/6	2.00E+00	3	1304
2023/2/6	2.00E+00	4	8299
2023/2/6	2.00E+00	5	8153
2023/2/6	2.00E+00	6	8391
2023/2/6	2.00E+00	7	7989
2023/2/6	2.00E+00	8	7384
2023/2/6	2.00E+00	9	7544
2023/2/6	2.00E+00	10	8000
2023/2/6	2.00E+00	11	7240
2023/2/6	2.00E+00	12	6777
2023/2/6	2.00E+00	13	7196
2023/2/6	2.00E+00	14	7185
2023/2/6	2.00E+00	15	7090
2023/2/6	2.00E+00	16	5712
2023/2/6	2.00E+00	17	6409
2023/2/6	2.00E+00	18	6172
2023/2/6	2.00E+00	19	6231

四、資料分析與決策 – 說明二

- Last欄位代表最後一週期異常燈管總數
- 第一爐在未來27天應不會翻新，且從趨勢判斷，累積時數的增加並不會顯著增加異常數量，因此都猜測未來應是持平或略增略減
- 唯1E0已經持平一個月了並有向上趨勢，我們認為可能會上升得多一些
- 將有可能的選擇以綠色螢光筆標出後，最終決定以紅框框起，第一爐都選擇其中測試集RMSE較低的Holt-Winters法

四、資料分析與決策 – 預測結果和選擇 (產線一)

	Last	Lasso	LGBM	XGBoost	CatBoost
1B0	58	40	85	106	72
1C0	38	29	34	45	31
1D0	42	26	24	39	25
1E0	26	24	22	16	22
1G0	0	3	1	1	1

1B0	59	59	49	75	60	53
1C0	40	37	26	40	30	33
1D0	46	45	26	28	40	61
1E0	36	36	37	32	32	37
1G0	2	2	0	0	2	1

四、資料分析與決策 – 說明三

- 第二爐的異常資料受累積時數影響較大，也常有爐層翻新問題，是故分開討論
- 2B0有8個爐層剛翻新、4個爐層快翻新； 2C0有4個爐層剛翻新、5個爐層快翻新，我們認為未來異常數會有比較明顯的下降
- 2D0有4個爐層快翻新，但考量大部分爐層的累積時數皆處於異常燈管可能出現的高峰，猜測會持平或略高
- 最後2E0只有1個爐剛翻新，未來應會有一段上升，而2G0資料太少，從模型表現看來LGBM符合這個趨勢

四、資料分析與決策 – 預測結果和選擇 (產線二)

	Last	Lasso	LGBM	XGBoost	CatBoost
2B0	53	25	39	33	44
2C0	51	28	29	40	38
2D0	50	22	56	70	68
2E0	39	16	14	16	16
1G0	0	0	0	1	1

2B0	51	50	54	60	50	44
2C0	47	46	56	59	47	52
2D0	46	46	45	51	42	52
2E0	38	38	46	21	27	42
2G0	1	1	0	0	1	1

五、分析與預測流程

1. 讀取初賽檔案

1. 將資料依據日期, 爐, 層的順序排列

1. 空值與異常值修正

1. 將資料分成訓練、驗證資料集，訓練模型時使用訓練資料集

1. 訓練模型並持續調整參數

1. 為每一爐選擇最適模型(根據爐層新舊以及趨勢判斷)
   - 1B0, 1C0, 1D0, 1E0, 1G0以Holt Winter模型預測
   - 2B0, 2D0, 2G0以LGBM模型預測
   - 2C0以CatBoost模型預測
   - 2E0以MA模型預測

1. 預測10個爐的結果

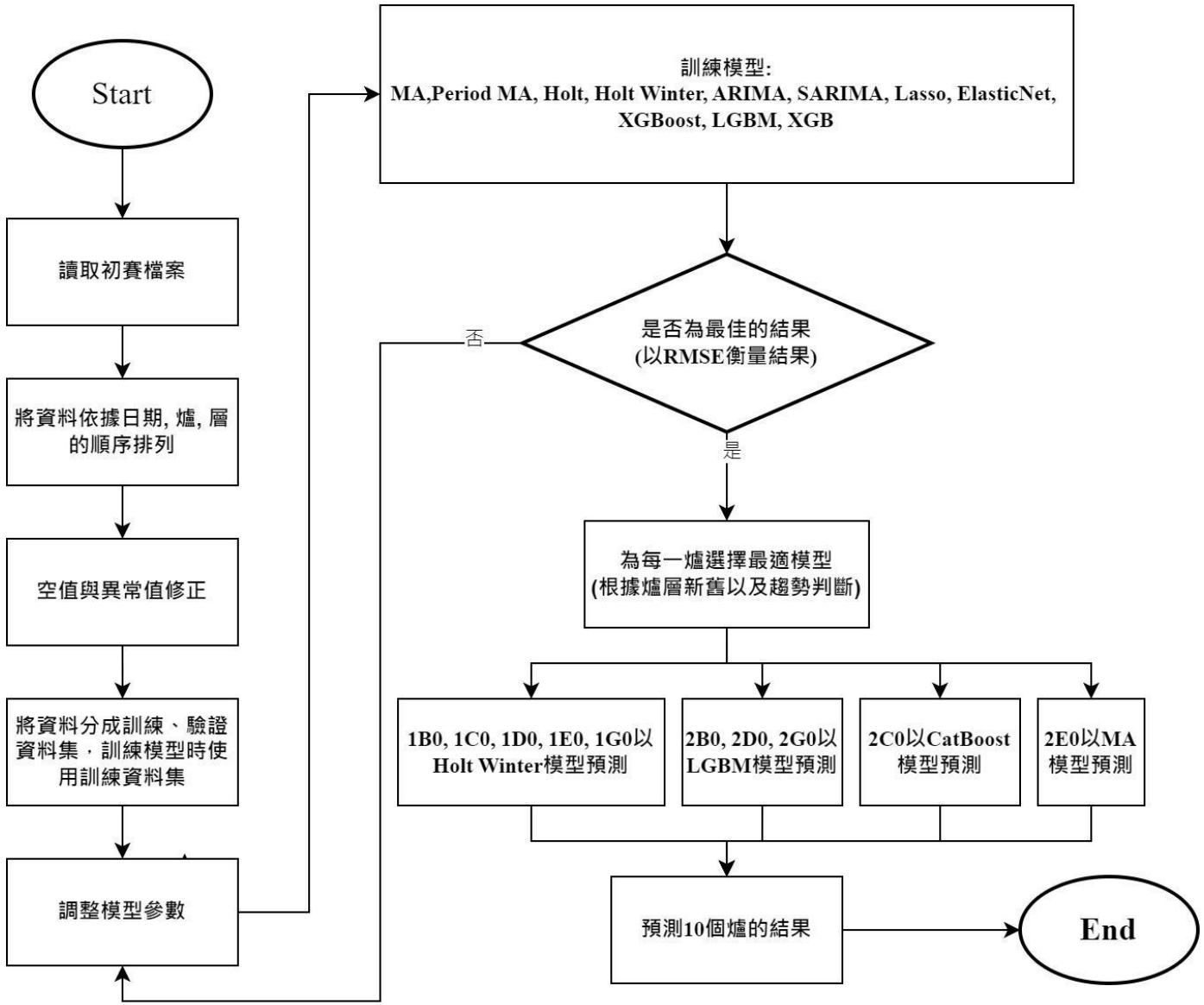


Fig. UV oven燈管異常預測流程圖