團隊測驗報告

報名序號：111096

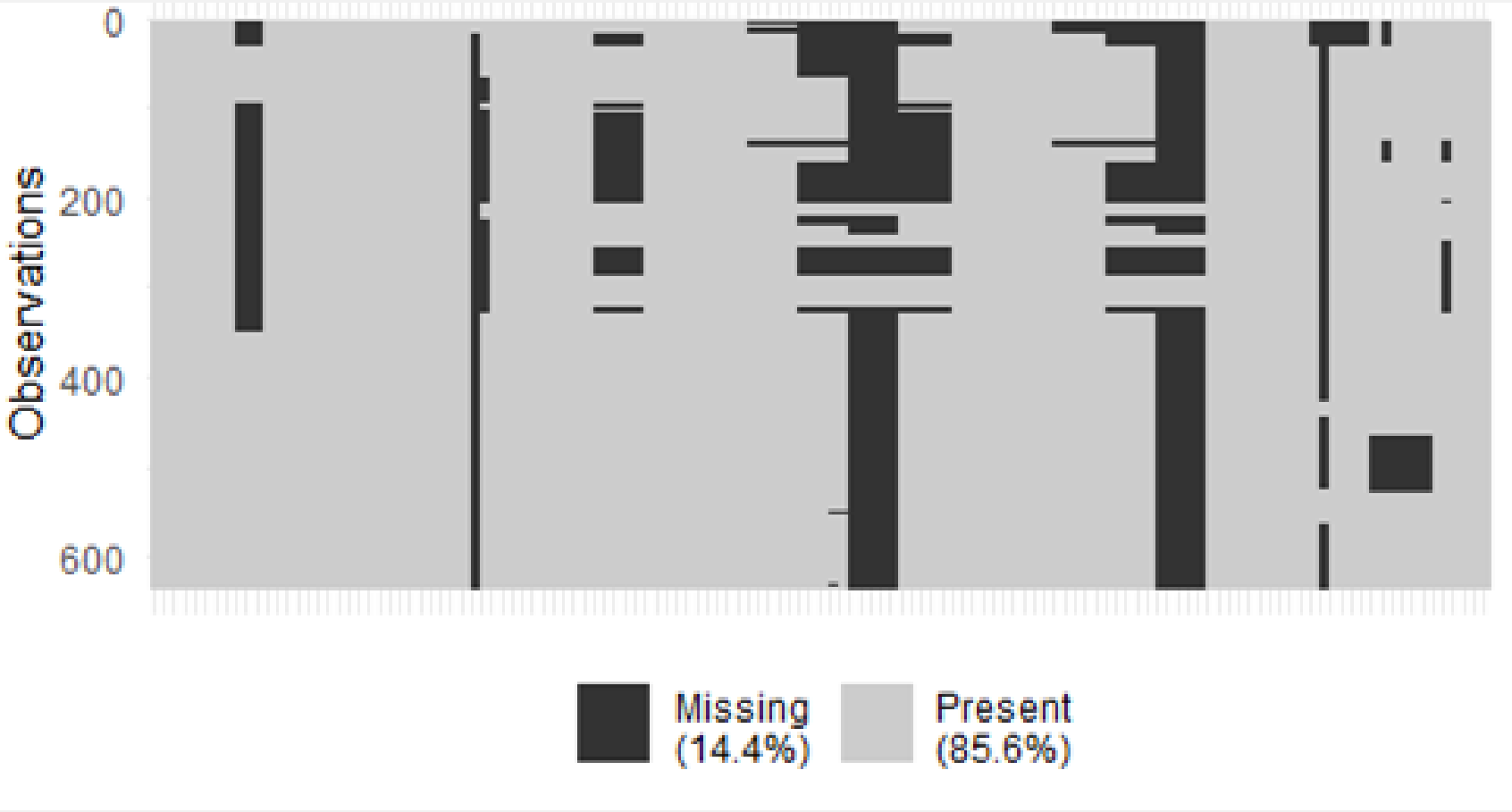
團隊名稱：兩把刷子

1

註1：請用本PowerPoint 文件撰寫團隊程式說明，**請轉成PDF檔案繳交**。
註2：依據競賽須知第七條，第4項規定：
測試報告之簡報資料不得出現企業、學校系所標誌、提及企業名稱、學校系所、教授姓名及任何可供辨識參賽團隊組織或個人身分的資料或資訊，違者取消參賽資格或由評審會議決議處理方式。

資料前處理

Missing data & Zero data 處理



我們發現原始資料中存在0值及缺失值，詢問主辦單位後證實兩者皆代表在製作工件的當下並沒有使用該機器。

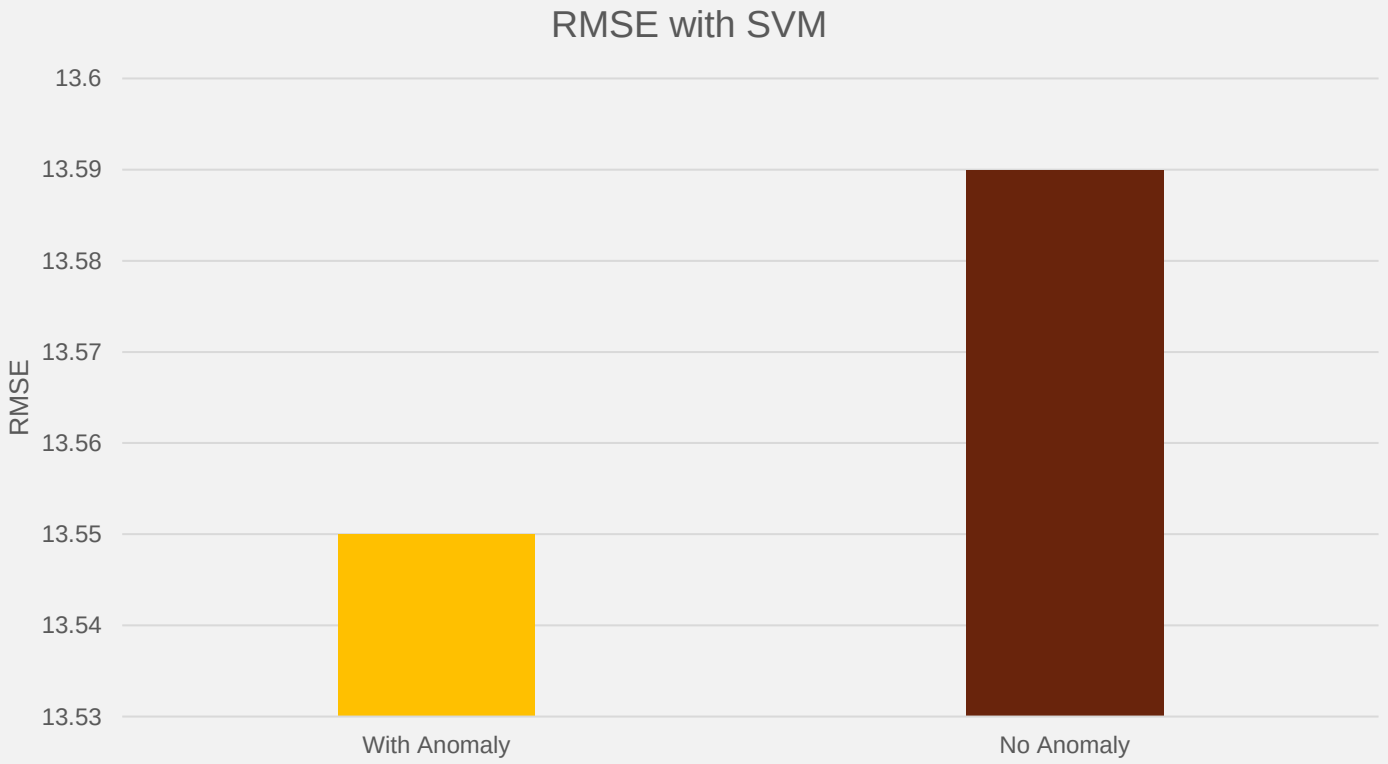
為了讓模型能正常訓練及準確判別是否有使用機器，我們將**zero & missing data以negative value (-99)**填補。

圖一 **Missing & Zero** 資料的分布

Anomaly Detection

異常值對模型學習或多或少產生影響，因此我們使用**IForest**、**LOF**、**SVM**進行異常值探測，並且額外新增 **one-hot encoding**欄位。
(令異常值為1、非異常值為0)

而經過多次嘗試過後，我們發現**SVM**相對其他方法較能夠精準預測，降低些許誤差。

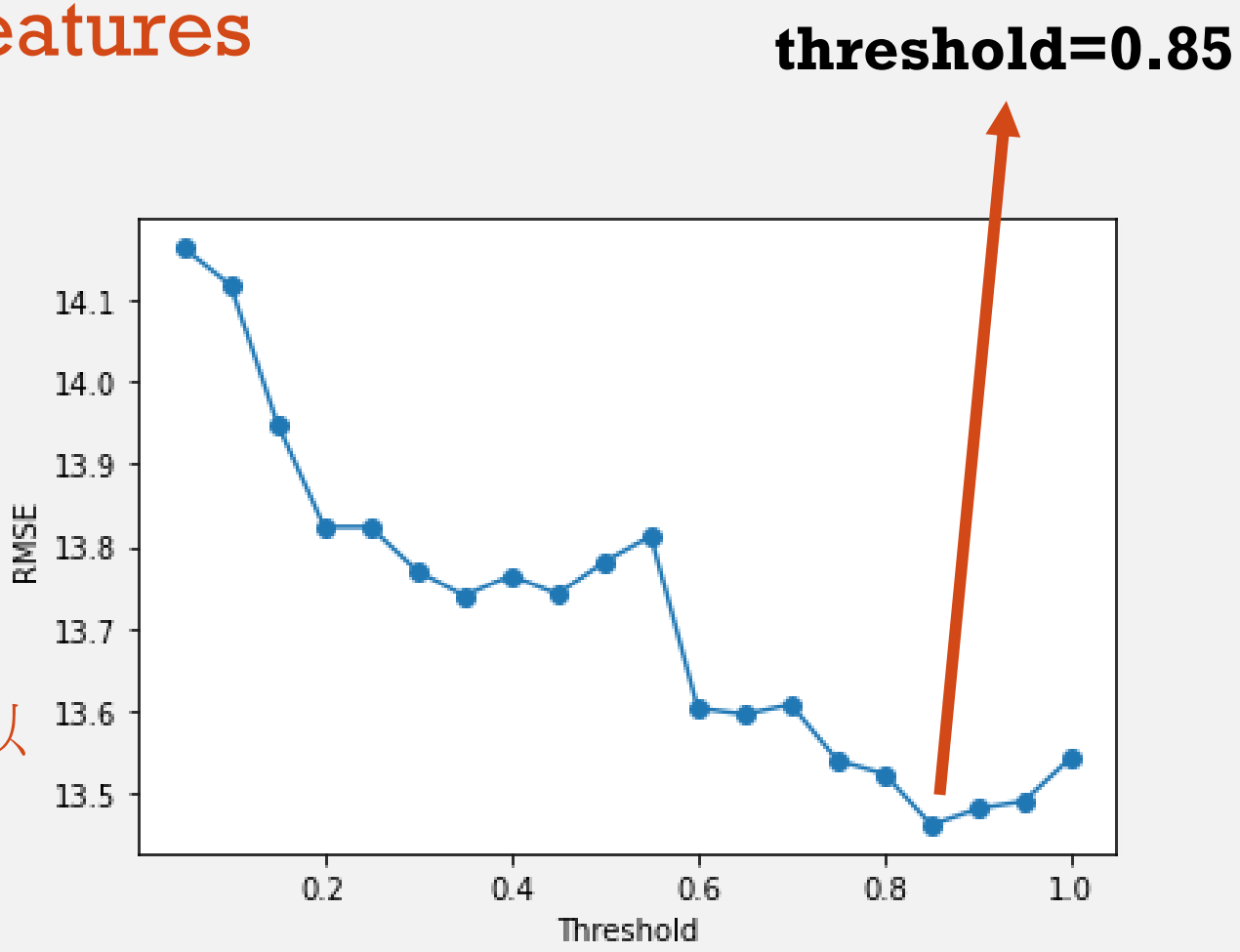


圖二 在**CV=10**下是否檢測異常值的平均**RMSE**差異

Remove High-Correlative Features

我們發現相同製程當中，有許多自變數間相關係數非常高。雖然相較迴歸**Tree-Based Model**受到高度貢獻性的影響較小，但我們發現移除高度共線性的自變數仍舊能夠提升模型預測的準確度。

而經過測試後我們發現移除相關係數0.85以上的自變數能最有效提升模型表現。



圖三 在不同**threshold**下**RMSE**的差異

演算法和模型介紹

模型介紹

我們使用XGBoost + MultioutputRegressor作為預測的模型，創造6個models同時預測工件的6面膜厚度。

超參數調整

1. 我們固定了learning rate及n_estimators，事先用了GridsearchCV對超參數進行了粗調，調整順序如下：
    1) max_depth & min_child_weight
    2) subsample & colsample_bytree
    3) gamma
    4) reg_alpha & reg_lambda

2. 手動粗調完超參數後，我們使用OptunaSearchCV，在timeout=7200、n_trials=1500的設置下，讓機器再細調附近的範圍內尋找最適參數。

預測結果

No	sensor_po	sensor_po	sensor_po	sensor_po	sensor_po	sensor_point10_i_value
1	49.12855	68.33193	75.81395	42.97736	65.11358	48.85075
2	60.05249	73.5195	87.43704	38.43504	72.68941	73.24535
3	60.11991	77.15467	85.95113	37.94644	73.29256	73.11927
4	76.94973	55.65463	81.22474	45.37712	62.40887	67.48421
5	76.94973	55.65463	81.22474	45.37712	62.40887	67.48421
6	79.37505	55.88069	80.36057	45.37621	67.42784	73.20262
7	72.18974	56.45359	77.76251	45.07046	60.03341	72.83845
8	82.2793	89.28392	90.56058	66.36386	83.72485	92.86587
9	85.1595	67.35203	107.1703	78.19783	64.30894	78.27393
10	84.25906	67.49738	106.0552	82.51894	64.80626	78.1292
11	86.82323	70.73954	109.2849	80.92484	65.7136	76.9475
12	85.60712	74.24847	106.0552	81.25142	68.80239	78.21368
13	85.32788	75.07906	105.8535	80.39112	72.2538	78.21368
14	82.12114	85.3722	91.57922	75.83218	90.83467	98.68046
15	81.34637	85.3722	91.95042	75.51967	89.38298	98.93913
16	77.7585	84.61961	90.72735	77.67377	90.53172	93.40491

圖四 預測結果示意圖，詳細預測結果請看111096_TestResult.csv

四、補充說明(其他或自行定義項目)