＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

總表類檔案：

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

檔名：Correlation\_heatmap.png

利用斯皮爾曼係數定義模型與模型間的相關性

使用各模型的預測結果與正確標籤去做Correlation

係數範圍 介於-1與1之間，當值大於0時稱兩變數為正相關，等於1稱為完全正相關；小於0時稱兩變數為負相關，等於-1時稱為完全負相關；等於0時則稱兩變數為無相關。

檔名：Data\_info.json

紀錄輸入資料(包含features與label)的類別與筆數資訊

檔名：Drop\_features.json(自定義模式中會有)

在insert\_random\_feature步驟中：

隨機特徵被添加到原始數據內。（新列random\_feature將出現在數據中）從前面的步驟中選擇最佳模型，其超參數用於訓練具有新插入特徵的模型。訓練後，計算特徵重要性並排序，重要性低於random\_feature的特徵保存到drop\_features.json文件中

檔名：Features\_heatmap.png

將最重要的25個特徵在各模型的表現用heatmap圖呈現

檔名：Features\_scores\_threshold\_2.5.csv(自定義模式中會有)

數字表示為 by CV fold 的各特徵重要性，counter就是基於random features 的特徵重要性去做統計，當各特徵重要性低於random features時就會被記+1，而當數字累計大於CV fold數的一半時，該特徵就會被列入到Drop features.json清單中

檔名：Golden\_features.json(自定義模式中會有)

將現有特徵兩兩配對做四則運算組成新特徵，並計算其重要性，並建立新特徵欄位(score值越小表示該特徵重要性越高)

檔名：Leaderboard.csv

紀錄每一個演算法模型的個別評估結果與訓練時間等資料

檔名：ldb\_performance\_boxplot.png

紀錄各演算法的評價指標表現，並以箱型圖呈現

檔名：ldb\_performance.png

將leaderboard裏頭的metric\_value按順序作圖到如右圖的點狀分佈

檔名：Params.json

紀錄本次程式的參數設定以及訓練花費時間與結果等資料

檔名：Progress.json

紀錄本次程式執行的流程與動作按步驟做紀錄，包含每個模型的前處理，訓練時間，訓練結果…等資料

檔名：README.md

將leaderboard.csv的模型表現總表與前述的部分圖片自動整合到此報告中

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

模型資料夾內的檔案：

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

檔名：confusion\_matrix.png ＆ 檔名：confusion\_matrix\_normalized.png

顯示預測結果與實際label的比對數量混淆矩陣，normalized就是標準化數字到0-1之間的混淆矩陣

檔名：calibration\_curve\_curve.png

校正曲線，解讀時若虛線與實線較為貼近越理想

可以參考https://axk51013.medium.com/%E6%A8%A1%E5%9E%8B%E4%BF%A1%E5%BF%83%E7%9A%84%E6%9C%AC%E8%B3%AA-probability-calibration-cbc680a44efa

檔名：Cumulative\_gains\_curve.png

累積增益曲線評估模型效能之一的表現方式。理想的Gains圖應在前期快速達到較高的累計Gains，很快趨於100％並平穩

可參考下列解說

https://www.ibm.com/docs/zh-tw/spss-statistics/29.0.0?topic=overtraining-cumulative-gains-lift-charts

檔名：Lift\_curve.png

增益曲線，評估模型效能之一的表現方式。理想的Lift圖應在較高的累計Lift上保持較長一段，然後迅速下降到1，可參考下列解說

https://www.ibm.com/docs/zh-tw/spss-statistics/29.0.0?topic=overtraining-cumulative-gains-lift-charts

檔名：Ensemble.json

將level-0的模型中挑選部分表現較好的模型做權重配比，repeat值為權重比例

檔名：Framework.json

類似總表檔案中的Progress.json，基本上會記錄每個CV fold對於資料的前處理狀況，包含對哪些欄位做刪除，對應的標籤類別，模型的超參數，訓練完的結果…等

檔名：Ks\_statistic.png

K-S或Kolmogorov Smirnov圖測量分類模型的性能。更準確地說，K-S是衡量正負例分佈之間分離程度的指標。如果將所有資料劃分為兩個獨立的組，其中一組包含所有正例而另一組包含所有負例，則K-S值為1。另一方面，如果模型不能區分正負例，那麼模型從總體中隨機選擇案例。 K-S值將為0。在大多數分類模型中，K-S將介於0和1之間，並且值越高，模型在區分正負例情況時越好

Learner\_fold系列檔案

(注意：每一個learner代表一個CV fold，若選用split模式則learner為0)

檔名：learner\_fold\_0\_importance.csv

紀錄各特徵的平均重要性

檔名：learner\_fold\_0\_training.log

紀錄訓練過程中每次迭代在train與validation的評價指標表現

檔名：learner\_fold\_0\_coefs.csv(線性模型專屬)

紀錄最後線性模型的截距與各特徵的權重

檔名：learning\_curves.png

紀錄不同CV fold的訓練資料與驗證資料表現(依據選擇的評價指標顯示)

檔名：permutation\_importance.png

依據不同CV fold的訓練結果確認特徵的重要性

檔名：precision\_recall\_curve.png

依據不同label的精確率與召回率作圖(area越大表現越好)

檔名：predictions\_out\_of\_folds.csv (驗證模式為)

將各類別的預測機率結果列出，可與實際label比較看正確性(target為實際標籤值，prediction為預測的或然率)

檔名：predictions\_validation.csv

將各類別的預測機率結果列出，可與實際label比較看正確性(target為實際標籤值，prediction為預測的或然率)

檔名：roc\_curve.png

紀錄不同類別的真陽性率與偽陽性率的圖表(area越大越好)

檔名：README.md

將上述的”部分”資料與圖片彙總成Markdown報告

檔名：Status.txt

紀錄該模型訓練結果是否OK