

Final Project

GitHub link :










<https://github.com/Amanda-Tai/TPS-AUG-2022>

Reference :

1. <https://www.kaggle.com/code/heyspaceturtle/feature-selection-is-all-u-need-2>
2. <https://www.kaggle.com/code/samuelcortinhas/tps-aug-22-failure-prediction/notebook?scriptVersionId=102980881>
3. <https://www.kaggle.com/code/qw1zzard/tps-aug-2022/notebook#Data-Science>

分數截圖：

因 Leaderboard 沒有顯示排名，故在 Filename 加上學號加以表示。

Submission and Description	Private Score	Public Score	Selected
 0816189 ML Final project - Version 19 Complete (after deadline) · 1s ago · Notebook 0816189 ML Final project Version 19	0.57738	0.58552	<input type="checkbox"/>
 0816189 ML Final project - Version 18 Complete (after deadline) · 2m ago · Notebook 0816189 ML Final project Version 18	0.58751	0.58546	<input type="checkbox"/>
 0816189 ML Final project - Version 17 Complete (after deadline) · 4m ago · Notebook 0816189 ML Final project Version 17	0.57906	0.58551	<input type="checkbox"/>
 0816189 ML Final project - Version 16 Complete (after deadline) · 7m ago · Notebook 0816189 ML Final project Version 16	0.58019	0.58551	<input type="checkbox"/>
 0816189 ML Final project - Version 15 Complete (after deadline) · 14m ago · Notebook 0816189 ML Final project Version 15	0.57881	0.58551	<input type="checkbox"/>
 0816189 ML Final project - Version 14 Complete (after deadline) · 16h ago · Notebook 0816189 ML Final project Version 14	0.58738	0.58547	<input type="checkbox"/>
 0816189 ML Final project - Version 13 Complete (after deadline) · 17h ago · Notebook 0816189 ML Final project Version 13	0.58581	0.58548	<input type="checkbox"/>
 0816189 ML Final project - Version 12 Complete (after deadline) · 17h ago · Notebook 0816189 ML Final project Version 12	0.58751	0.58546	<input type="checkbox"/>
 0816189 ML Final project - Version 11 Complete (after deadline) · 1d ago · Notebook 0816189 ML Final project Version 11	0.57448	0.58695	<input type="checkbox"/>

程式說明：

預測方式為使用 Fisher Score 篩選出特徵，再用 LogisticRegression 和 GroupKFold 進行預測，在 0816189.ipynb 中將會輸出兩個檔案 my_model.pkl

及 `submission.csv`，分別裝著訓練模型及預測結果，下面將分區介紹程式。

1. 執行環境
 - `kaggle`
2. 填上缺失的 Data
 - 利用 `pd.concat` 將沒有 `failure` 這一 column 的 `train.csv` 與 `test.csv` 的行(row)連起來放入 `data`。
 - 利用 `Scikit-learn` 中的 `IterativeImputer` 來補足缺失的資料，
3. 資料預處理
 - 使用 `get_dummies` 對 `'attribute_0'` 及 `'attribute_1'` 兩個 columns 的變量進行虛擬變量轉換。
 - 再用 `merge` 合併放入 `data`，並把 `'attribute_0'` 及 `'attribute_1'` 兩個 columns 從 `data` 中刪掉。
 - 把資料進行 `combination`、`aggregation` 及 `ratio` 的處理，增加特徵數量。
4. Feature Selection
 - 利用 `Fisher Score` 來進行特徵篩選，其中找到的值越大，說明這個特徵在分類中起到的作用越大。
5. RUN
 - 定義一個 `LogisticRegression` 模型和 `GroupKFold` 驗證器。
 - 使用驗證器分割訓練數據集，並使用模型對訓練數據進行訓練及預測，計算出精確值。
 - 放入 `submission.csv`。

備註：在程式碼中有針對行的詳細註解。