

ML Final Project

GitHub link: https://github.com/ChicoChen/109550110-ML_Final_Project

109550110 陳尚奇

● Introduction :

這次的 project 的目標是參與競賽 Tabular Playground Series - Aug 2022，並訓練出適合這次競賽的 model。

有鑑於這次提供的 dataset 有別於上次 HW5，是一維的數據，model 架構並未採用深度學習的方式，而是使用 Logistic Regression 作為基底，降低對 model 進行訓練所需的硬體與時間成本。

● Methodology:

➤ Data pre-process:

這次提供的 dataset 為維度 26570* 26 的 .CSV 檔，捨去 ID 與 failure label 後，每份資料有 24 個 feature 可供訓練，其中 3 個 feature 是不可量化的。整份檔案有零散的資料遺失。

對於不可量化的 3 個 feature:


product code 從 [A, Z] 對應到 [0, 25]

attribute_0 & attribute_1，只留下其代號


ex: material_7 □ 7

而因為缺失的資料數量不多，與其捨去有缺失的資料，我選擇將空缺值填上該 feature 的中位數，其他還有試過平均值跟眾數，但中位數能得到最好的結果。


Median:

 submission.csv Complete (after deadline) · 5d ago · UN-normalized, Impute_median	0.59019	0.58464	<input type="checkbox"/>
--	----------------	----------------	--------------------------

Mean:


 submission_prob (1).csv Complete (after deadline) · 5d ago	0.59004	0.58602	<input type="checkbox"/>
--	----------------	----------------	--------------------------

Most_frequent:


 submission_prob (3).csv Complete (after deadline) · 5d ago	0.59006	0.58267	<input type="checkbox"/>
---	----------------	----------------	--------------------------

原本有對可量化的 21 個 features 進行 normalized，但在偶然中發現不進行 normalized 反而效果更好。

Normalized:

 submission_prob (1).csv Complete (after deadline) · 5d ago · impute mean, unnormalized	0.59004	0.58602	
--	----------------	----------------	--

Unnormalized:

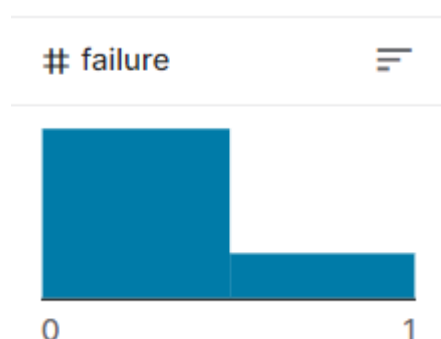
 submission_prob.csv Complete (after deadline) · 5d ago · impute mean, normalized	0.58547	0.58601	
--	----------------	----------------	--

➤ Model architecture & training:

我使用 scikit-learn 套件中提供的 Logistic Regression model 作為這次的模型。

藉由 GridSearchCV function，對 training dataset 進行 5-fold Cross-Validation 與 Grid Search，再用找出的最佳參數與整份 training dataset 對 model 進行訓練。

另外因為 dataset 中 failure label 的數量分布大概呈現 Failure(0): Failure(1) = 4: 1 的分布，我將參數 class_weight 設定為 "balanced"，藉此平衡分布失衡的情況。



➤ Hyperparameters:


最終找出的最佳參數如下：penalty = 'l2'，C = 1e-5

Grid Search 的詳細結果請見 Appendix

● Summary:

這次的 project 我主要是著重在 data pre-process 跟 grid search 上，在 model 架構上沒花太多心思，畢竟上次 HW5 的時候為了找出一個好的架構耗費近乎一周結果還是一無所獲，這次不敢再皮了乖乖用 package 裡面提供的 model，模型架構什麼的等以後有空再研究。

Final Result:

	109550110_submission.csv Complete (after deadline) · 12m ago	0.59035	0.58568
---	---	---------	---------


● Appendix:

(1) Grid Search result:


C	1	0.1	0.01	1e-3	1e-4	1e-5	1e-6
Private score	0.5886	0.58877	0.58839	0.58916	0.59019	0.59035	0.59019
Public score	0.58196	0.58224	0.58187	0.58273	0.58464	0.58565	0.58439

(2)有根據 correlation coefficient 選出十個 feature 對進行訓練過，但成果不甚理想就沒特地在上面的部分中提到，以下是 feature selection 的結果。

feature selection & normalized:

	submission (2).csv Complete (after deadline) · 2d ago · normalized, feature-selected	0.58803	0.58492
---	---	---------	---------

only feature selection:

	submission (1).csv Complete (after deadline) · 2d ago · UN_normalized, feature-selected	0.58774	0.58268
---	--	---------	---------

(3)Downloadable model link:

https://drive.google.com/file/d/1co8qtvfzecxFkX_Spcemd1HgAG29bC00/view?usp=share_link

or download from github