Report

0616329 李京叡

Data Input:

用 panda 的 read_csv 讀取檔案

Data Preprocessing:

1. 移除有 missing value(?)的欄位: 'workclass', 'occupation', 'native-country'，另外因為我覺得 education-level 和 education-num 所代表的資訊是重複的，因此也移除了’education-level’的欄位。
2. 打散所有 data
3. 將 data 以 7:3 分成 train 和 validation 兩組

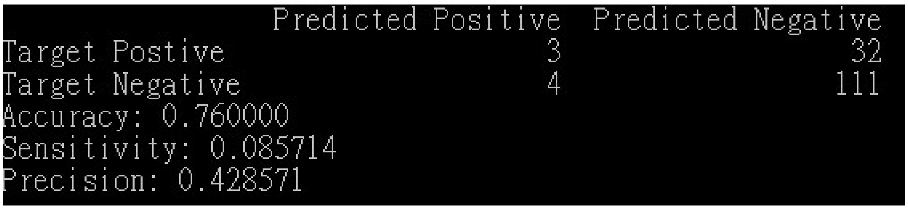
Model construction:

Decision tree:
5 個主要的 function: tree_build, entropy, info_gain, remainder, find_threshold，在 build_tree 時，會去找當下 information gain 最大的屬性來分割資料，計算 infromation gain 需要用到 entropy, remainder，此外若屬性是連續的，則會先用 find_threshold 找出最佳分割點，把 data 分成兩類，大於 threshold 和小於 thershold。

Random forest:
隨機從 data 中取 100 筆資料建 decision tree，總共建了 3 棵，驗證時把資料送進每一棵樹並取得結果，最多數的結果即為最後答案。

Results (number of data=500):

Decision tree:



```
                 Predicted Positive  Predicted Negative
Target Postive                    3                  32
Target Negative                   4                 111
Accuracy: 0.760000
Sensitivity: 0.085714
Precision: 0.428571
```

Confusion matrix:

	Predicted Positive	Predicted Negative

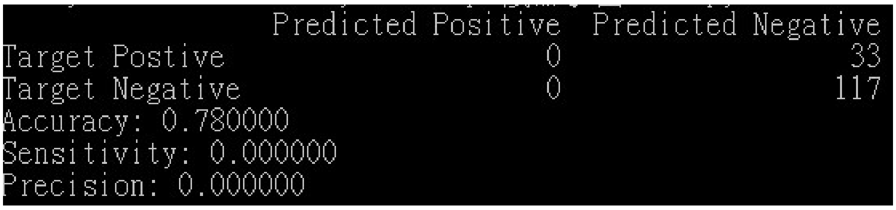
Target Postive	3	32
Target Negative	4	111

Accuracy: 0.76

Sensitivity: 0.085714

Precision: 0.428571

Random forest:



```
                 Predicted Positive  Predicted Negative
Target Postive                    0                  33
Target Negative                   0                 117
Accuracy: 0.780000
Sensitivity: 0.000000
Precision: 0.000000
```

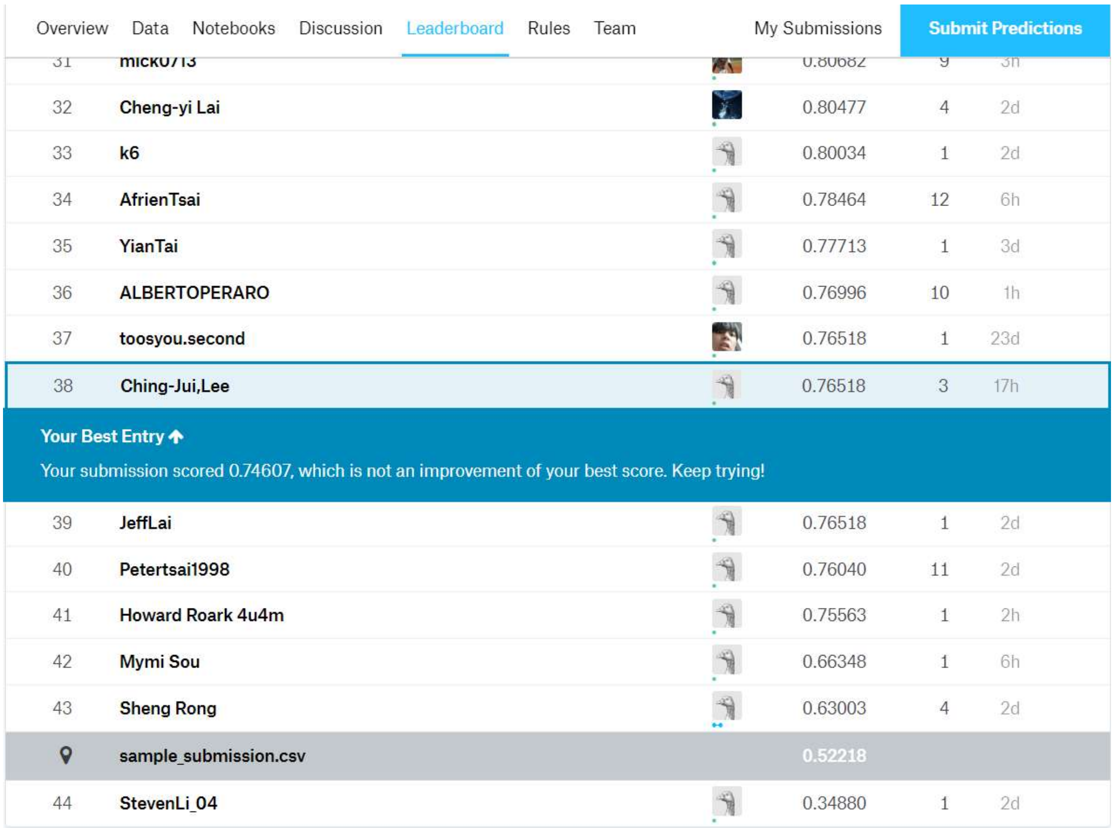
	Predicted Positive	Predicted Negative
Target Postive	0	33
Target Negative	0	117

Accuracy: 0.78

Sensitivity: 0

Precision: 0

Kaggle Submission:



31	mick0713	0.80682	9	3h
32	Cheng-yi Lai	0.80477	4	2d
33	k6	0.80034	1	2d
34	AfrienTsai	0.78464	12	6h
35	YianTai	0.77713	1	3d
36	ALBERTOPERARO	0.76996	10	1h
37	toosyou.second	0.76518	1	23d
38	Ching-Jui,Lee	0.76518	3	17h

Your Best Entry ↑

Your submission scored 0.74607, which is not an improvement of your best score. Keep trying!

39	JeffLai	0.76518	1	2d
40	Petertsai1998	0.76040	11	2d
41	Howard Roark 4u4m	0.75563	1	2h
42	Mymi Sou	0.66348	1	6h
43	Sheng Rong	0.63003	4	2d
	sample_submission.csv	0.52218		
44	StevenLi_04	0.34880	1	2d

上面顯示的最高成績，是預測結果全部皆為 0 的，因為我一開始程式有一些錯誤，導致預測結果皆為 0，但後來改好後的正確率又沒辦法超越，因此 0.74607 應該才是比較接近我的 model 的正確率。

Comparison & Conclusion:

我覺得正確率沒辦法提高的原因是為 train 的 data 太少，但是因為算 threshold 非常花時間，要先把 data 排序，再算出連續兩個值的中點，再去算 information gain，我試過把所有 data 都餵進去，但是跑了好幾個小時，還是跑不出來，為了方便測試與得到結果，我只好把 data 量都縮小，因此正確率沒辦法提高。