310706034 資管碩一 吳啓玄 資料探勘研究與實務 HW0- titanic 小練習

1. 觀察下圖訓練集資料變數遺失值比例，發現 Age、Cabin、Embarked 有遺失值，其中 Age、Cabin 有較大比例遺失值，新增 Has_Age、Has_Cabin 變數，若 Age、Cabin 是遺失值，則 Has_Age、Has_Cabin 為 0，反之為 1，另外，Embarked 補值 ”No”，Age 補值 0。

```
PassengerId     0.000000
Survived        0.000000
Pclass          0.000000
Name            0.000000
Sex             0.000000
Age             0.198653
SibSp           0.000000
Parch           0.000000
Ticket          0.000000
Fare            0.000000
Cabin           0.771044
Embarked        0.002245
dtype: float64
----------------------------
```

處理後之訓練集資料變數遺失值比例如下:

```
PassengerId     0.000000
Survived        0.000000
Pclass          0.000000
Name            0.000000
Sex             0.000000
Age             0.000000
SibSp           0.000000
Parch           0.000000
Ticket          0.000000
Fare            0.000000
Cabin           0.771044
Embarked        0.000000
Has_Age         0.000000
Has_Cabin       0.000000
dtype: float64
----------------------------
```

2. 針對有興趣之變數進行資料探索性分析(EDA)，觀察變數存活率，如下:

	Pclass	Survived
0	1	0.630
1	2	0.473
2	3	0.242

	Sex	Survived
0	female	0.742
1	male	0.189

Survived	0	1
Age	23.653005	24.034123

	Has_Age	Survived
0	0	0.294
1	1	0.406

Survived	0	1
SibSp	0.553734	0.473684

Survived	0	1
Parch	0.32969	0.464912

Survived	0	1
Fare	22.117887	48.395408

	Has_Cabin	Survived
0	0	0.300
1	1	0.667

	Embarked	Survived
0	C	0.554
1	No	1.000
2	Q	0.390
3	S	0.337

3. 把訓練集資料轉換為 dummy variable 並做 min-MAX scale，轉換後如下:

	Sex	Age	SibSp	Parch	Fare	Has_Age	Has_Cabin	Pclass_1	Pclass_2	Pclass_3	Embarked_C	Embarked_No	Embarked_Q	Embarked_S
0	1.0	0.2750	0.125	0.000000	0.014151	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
1	0.0	0.4750	0.125	0.000000	0.139136	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0
2	0.0	0.3250	0.000	0.000000	0.015469	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
3	0.0	0.4375	0.125	0.000000	0.103644	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
4	1.0	0.4375	0.000	0.000000	0.015713	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	1.0	0.3375	0.000	0.000000	0.025374	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
887	0.0	0.2375	0.000	0.000000	0.058556	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
888	0.0	0.0000	0.125	0.333333	0.045771	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
889	1.0	0.3250	0.000	0.000000	0.058556	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0
890	1.0	0.4000	0.000	0.000000	0.015127	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0

891 rows × 14 columns

4. 建構 Random Forest、XGBoost、KNN 三個模型，根據 5-fold Cross Validation 進行 tuning，Random Forest 調參樹木數量、樹木深度，XGBoost 調參樹木數量、樹木深度，KNN 調參觀看鄰居數，得到以下結果並挑選準確率最高模型

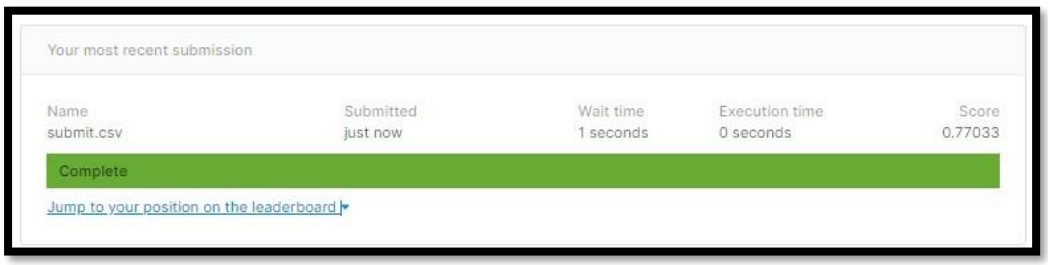
```
After Tuning Best Accuracy
--------------------------------------
random forest: 0.8238340342728016
knn: 0.8047768501663425
xgboost: 0.8372857949908982
```

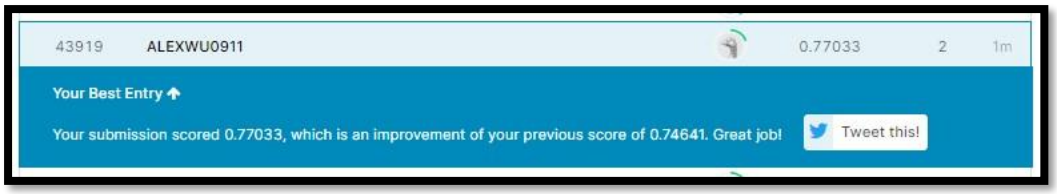
5. 測試集資料沒有 Embarked=”No”所以補一欄，另外，Fare 有遺失值所以採取平均插值，其餘皆和訓練集資料處理方式相同，測試集資料處理後如下:

	Sex	Age	SibSp	Parch	Fare	Has_Age	Has_Cabin	Pclass_1	Pclass_2	Pclass_3	Embarked_C	Embarked_Q	Embarked_S	Embarked_No
0	1.0	0.453947	0.000	0.000000	0.015282	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0
1	0.0	0.618421	0.125	0.000000	0.013663	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0
2	1.0	0.815789	0.000	0.000000	0.018909	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0
3	1.0	0.355263	0.000	0.000000	0.016908	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0
4	0.0	0.289474	0.125	0.111111	0.023984	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
413	1.0	0.000000	0.000	0.000000	0.015713	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0
414	0.0	0.513158	0.000	0.000000	0.212559	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0
415	1.0	0.506579	0.000	0.000000	0.014151	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0
416	1.0	0.000000	0.000	0.000000	0.015713	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0
417	1.0	0.000000	0.125	0.111111	0.043640	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0

418 rows × 14 columns

6. 使用 XGBoost 模型，對所有訓練集資料依照 cross validation tuning 得到的最佳參數進行建模，最後，用該模型對測試集做預測後，上傳 Kaggle 的準確率是 0.77033:




43919
ALEXWU0911
0.77033
2
1m
Your Best Entry
Your submission scored 0.77033, which is an improvement of your previous score of 0.74641. Great job!
Tweet this!