國立臺北商業大學 資訊管理系

機器人與深度學習

鐵達尼號生存預測 期末報告

班級：日二技資二甲

組員：10636001 曾意淳

10636002 吳昱萱

10636025 曾佳怡

指導老師：許晉龍 老師

目錄

摘要

鐵達尼號生存預測這個比賽，你會拿到許多關於乘客的資訊像是乘客的性別、姓名、出發港口、住的艙等、房間號碼、年齡、兄弟姊妹 + 老婆丈夫數量 (Sibsp)、父母小孩的數量（ parch ）、票的費用、票的號碼這些去預估這個乘客是否會在鐵達尼號沈船的意外中生存下來。

介紹

鐵達尼號的沉沒是歷史上最昭著的沉船事件之一。1912 年 4 月 15 日，在其處女航中，鐵達尼號在與冰山相撞後沉沒，在 2224 名乘客和機組人員中造成 1502 人死亡。這場聳人聽聞的悲劇震驚了國際社會，也導致後續更好的船舶安全規定。

造成海難失事的原因之一是乘客和機組人員沒有足夠的救生艇。儘管倖存下來有一些運氣因素，但有些人比其他人更容易生存，比如女人，孩子和上流社會。

資料集介紹及資料集來源

- **資料集**

Variable	Definition	Key
PassengerId	乘客編號	
survival	生存情況	0 = No, 1 = Yes
pclass	客艙等級	1 = 1st, 2 = 2nd, 3 = 3rd
Name	姓名	
sex	性別	
Age	年齡	
sibsp	同代直系親屬數	
parch	不同代直系親屬數	
ticket	船票編號	
fare	船票價格	

Variable	Definition	Key
cabin	客艙號	
embarked	登船港口	C = Cherbourg, Q = Queenstown, S = Southampton

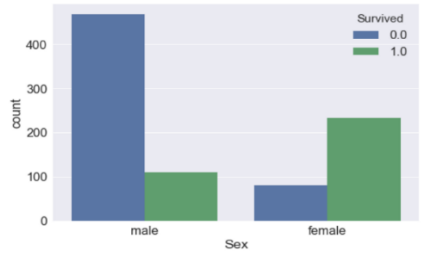
- **資料特徵**

➢ 性別

```
sns.countplot(df_data['Sex'], hue=df_data['Survived'])
display(df_data[["Sex", "Survived"]].groupby(['Sex'], as_index=False).mean().round(3))
```

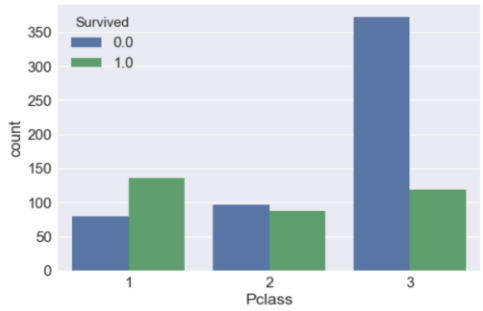
started 21:19:04 2018-06-16, finished in 302ms

	Sex	Survived
0	female	0.742
1	male	0.189

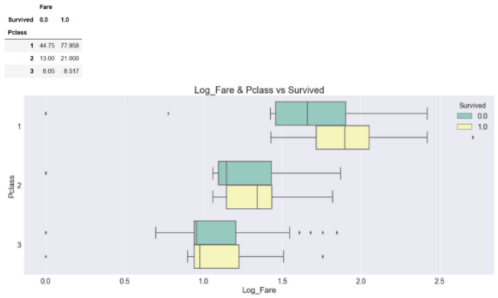


➢ 艙等

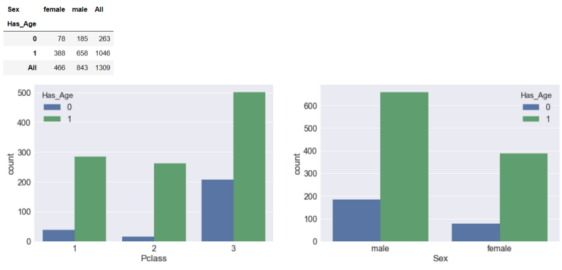
	Pclass	Survived
0	1	0.630
1	2	0.473
2	3	0.242



➢ 票價



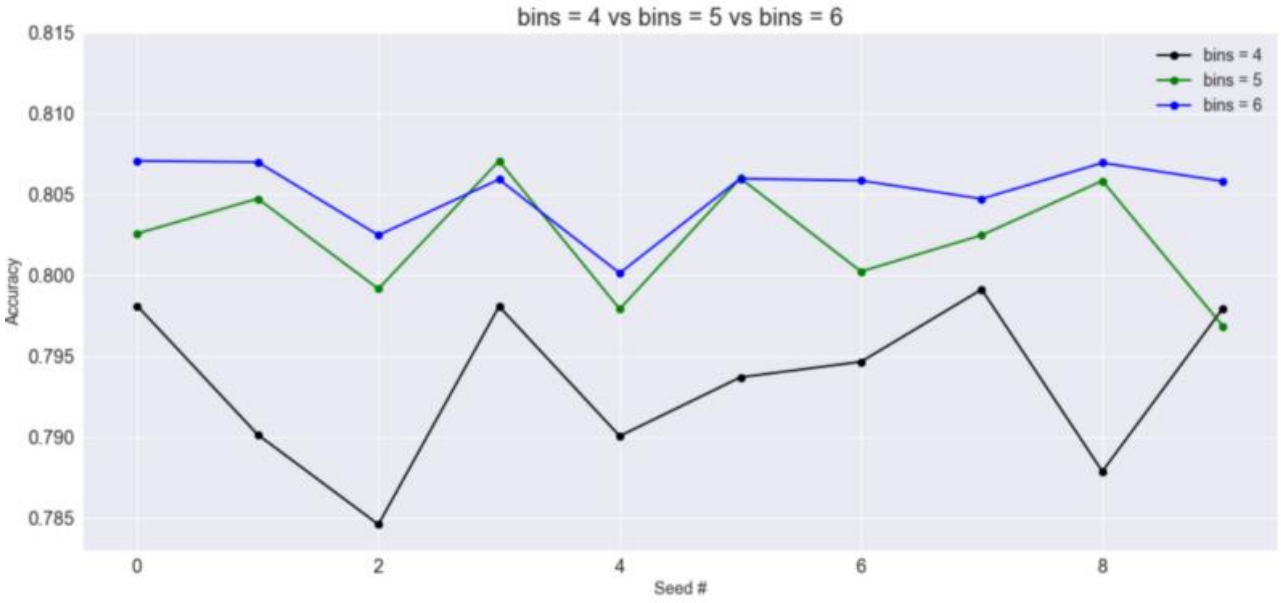
➢ 年齡



機器學習或深度學習方法

先挑選特徵（性別、艙等），再將資料轉換（女性為 0，男性為 1）。我們選用隨機森林模型來進行，因為隨機森林其平行化計算的特質在資料集小或是大時的運算效能都不錯。且隨機森林是以不純度函數來切分樣本，因此不需要標準化，簡化了建模時的步驟。

研究結果與結論

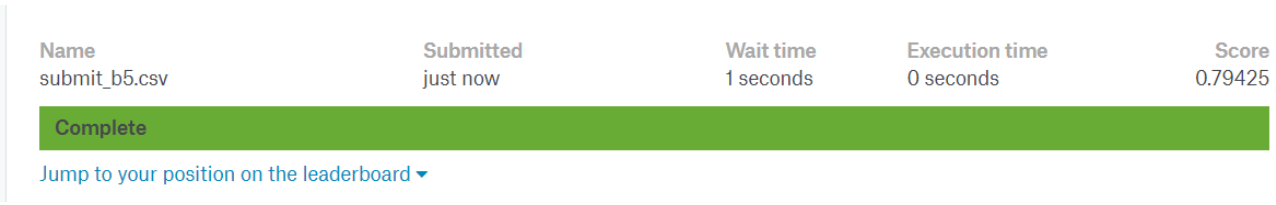



```
b4, b5, b6 = ['Sex_Code', 'Pclass','FareBin_Code_4'], ['Sex_Code','Pclass','FareBin_Code_5'],\
['Sex_Code','Pclass','FareBin_Code_6']
b4_Model = RandomForestClassifier(random_state=2,n_estimators=250,min_samples_split=20,oob_score=True)
b4_Model.fit(X[b4], Y)
b5_Model = RandomForestClassifier(random_state=2,n_estimators=250,min_samples_split=20,oob_score=True)
b5_Model.fit(X[b5], Y)
b6_Model = RandomForestClassifier(random_state=2,n_estimators=250,min_samples_split=20,oob_score=True)
b6_Model.fit(X[b6], Y)
print('b4 oob score :%.5f' %(b4_Model.oob_score_),'   LB_Public : 0.7790')
print('b5 oob score :%.5f '%(b5_Model.oob_score_),' LB_Public : 0.79425')
print('b6 oob score : %.5f' %(b6_Model.oob_score_), '  LB_Public : 0.77033')
```

started 21:40:15 2018-06-16, finished in 1.55s

```
b4 oob score :0.80584    LB_Public : 0.7790
b5 oob score :0.81033   LB_Public : 0.79425
b6 oob score : 0.80135   LB_Public : 0.77033
```

當切分的區間太少時，區間內的資料太多一起平均，這樣沒有辦法看出差異性，使得特徵失真；當切分區間太多時，一點點票價的不同，都影響了生存率的高低，如此一來很明顯地會 overfitting，並且，切分區間趨近於無限大時，就回到了原本的數值特徵，所以，由上圖得知，我們將票價特徵切分成 4 份的準確率較低，6 份比 5 份稍微好一點，最後我們將此結果上傳至 Kaggle，得到了以下結果。



參考文獻

Recursive Forward Elimination Workflow to 0.82296

3 Strategies Analyzing Age and Their Impact

How am I doing with my score?

[資料分析&機器學習] 第 4.1 講 : Kaggle 競賽-鐵達尼號生存預測-（前 16%排名）

[機器學習專案] Kaggle 競賽-鐵達尼號生存預測（Top 3%）