**Titanic: Machine Learning from Disaster**

學號：40847016S

姓名：謝尚恆

組別：5

**資料集特徵資料說明、屬性特性說明**

PassengerId：訓練集包含891位乘客的資訊、測試集包含418位乘客

Survived：0表示該乘客沒能存活、1表示該乘客倖存

Pclass：代表船艙等級，其中第1級最高檔、第3級最廉價。

Sex：male代表該乘客為男性、female代表該乘客為女性

Age：代表該乘客的年齡，如果未滿一歲，用小數表示，若不清楚該乘客的真實年齡，則用XX.5表示。

SibSp：代表該乘客有多少兄弟姐妹和伴侶和他一起上船

Parch：代表該乘客有多少父母和兒女和他一起上船

Fare：代表該乘客該趟旅程的船票價錢

Embarked：代表該乘客登上船的港口名稱，C=Cherbourg, Q=Queenstown, S=Southampton。

**對特徵做甚麼樣的分析? 哪些前處理? 採用哪些特徵? 原因?**

**分析**

對性別做存活率分析：女性為74.2%、男性為18.89%

對船艙等級做存活率分析：第一級為62.96%、第二級為47.28%、第三級為24.24%

對年齡做存活率分析：

0~10歲為59.38%

10~20歲為38.26%

20~30歲為36.52%

30~40歲為44.52%

40~50歲為38.37%

50~60歲為40.48%

60~70歲為23.53%

70~80歲為20.00%

對上船港口做存活率分析：C港口為55.36%、Q港口為38.96%、S港口為33.90%

對票價做存活率分析：

10元以下為19.94%

10~25元為42.08%

25~50元為41.95%

50~100元為65.42%

100元以上為73.58%

**前處理**

在對年齡做存活率分析時，因為訓練集中有177位乘客的年齡為空值，所以我忽略這177位乘客，並對剩下的資料做分析。

實作方面我並沒有考慮dropna()函式，而是跑一次迴圈檢查，只要某列年齡為空，就將該列清除，雖然這個函式很方便，但若使用dropna()函式，只要某列存在一個空值，該列就會被清除。我要刪掉的資料只是年齡為空的所有資料，而Cabin欄位大部分都是空值，如果用該函式，就會刪掉很多含有年齡資訊的資料，而這不是我想要的。

我先將訓練集的csv檔讀進來並轉成DataFrame後用變數dfTrain儲存，用一個名稱為nanAge的list記錄所有年齡為空的index，將nanAge餵給drop()函式移除所有年齡為空的列後，便產生一個新的DataFrame，我用變數noNanAgeDfTrain儲存，原本打算用noNanAgeDfTrain來分析各個年齡層的存活率，但這時我遇到一個問題。

仔細看訓練資料可以發現，index為5（也就是PassengerId為6的乘客）的那一列年齡資訊為空，這時我先試著輸出前10筆資料（noNanAgeDfTrain[:10] 或 noNanAgeDfTrain.head(10)），可以正常輸出，但是當我輸出noNanAgeDfTrain[5]時，編譯器卻報出KeyError這個錯，這代表著該index的資訊雖然有被清除，但該index還是存在，只是變成沒有資料，一存取就報錯。所以我決定直接刪除noNanAgeDfTrain這個DataFrame，直接利用dfTrain，然後只要遇到年齡資訊為空的index，就略過。

再來是上船港口的部分，這891位乘客中，只有2位乘客的Embarked為空，將處理年齡空值的方法套用在處理Embarked我覺得沒有必要。我的做法是先找出這三個港口上船的人數，然後將這兩位乘客都分給人數最多的Southampton港口，因為光是由Southampton港口上船的就有644位，再多塞兩位對於資料的影響並不大，至於分配的意思也就是直接在訓練集的csv檔中填入S港口。

最後是票價的部分，肉眼觀察後，我決定將區間設成[-1,10,25,50,100,513]，幸好票價這個欄位並沒有空值，我只需要直接逐個區間分析即可。

**原因**

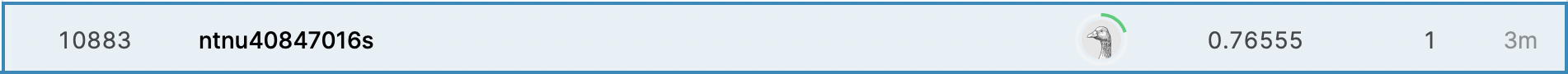
性別、年齡在分析方面一直都是很重要的指標，一定有這兩項。然後船艙等級較好的，其安全措施應該也會比較好，這也有可能提升存活率，值得分析。雖然從哪個港口登上船，對於是生是死並沒有關係，但也可以分析，看看從哪個港口上船的乘客比較幸運。最後是票價，價格越高，通常能享受更多服務、位置也會是比較安全的，是生是死和位置是否安全也有一定的關係，值得分析。

**基於什麼理由選擇哪個分類器?**

我選擇RandomForestClassifier，因為它的學習速度很快，而且在未調整參數的情況下，準確度也很高，它應付噪音的能力也不錯。

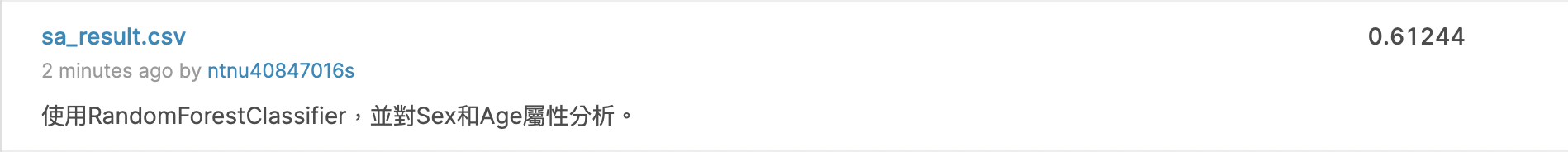
**採用的評估指標結果與觀察**

這是僅對性別分析產生的結果：

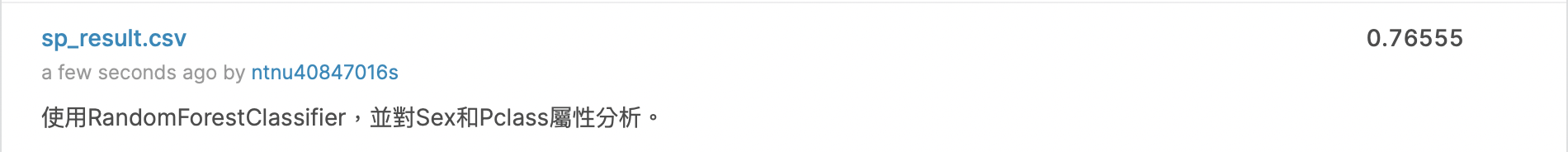


這是對年齡、性別分析產生的結果：

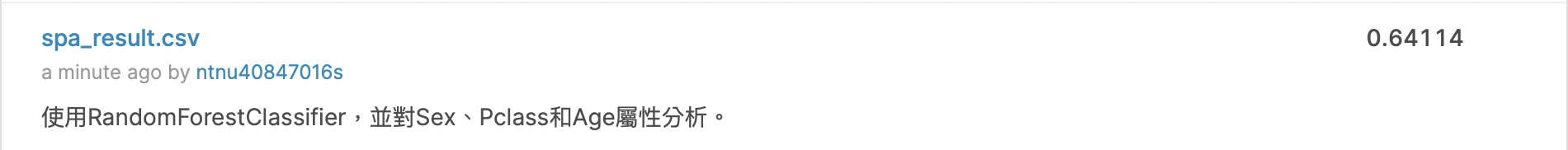
雖然第一頁的年齡分析是拿掉空值後，剩下的714位乘客的數據。但年齡也是一個影響生死很重要的條件，所以在做完分析後，我分別以「訓練集中的所有乘客年齡的中位數」填入訓練集中的所有年齡空值、「測試集中的所有乘客年齡的中位數」填入測試集中的所有年齡空值。



這是對性別、船艙等級分析產生的結果：



這是對年齡、船艙等級、性別分析產生的結果：



這是僅對船艙等級分析產生的結果：



這是僅對票價分析產生的結果：



**總結**

上面我總共試了7種不同的情況，從上面的點數可以發現，若僅對性別進行分析，表現是最好的，然後當學習的特徵越來越多，表現會越來越差，這可能是因為特徵變多，使得干擾也變多，進而影響到辨識表現。