機器學習課堂競賽

TEAM_09: 李承暘, 鄭兆瑋, 葉庭渝

前言—本次競賽資料為鐵達尼號資料集，Training Data有891筆資料，共有12個特徵；Testing Data有418筆資料，共有11個特徵，缺少Survived這項特徵。最後要測試預測模型對於Survived特徵預測的準確度，測試平台為Kaggle。

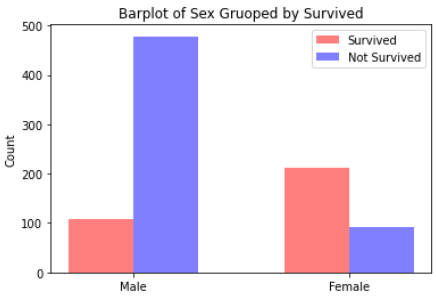
I. 方法介紹

先進行探索式資料分析作為後續模型選用的特徵參考，再開始資料前處理，再進入模型建立的部分。本次競賽所使用模型為決策樹(Decision Tree)和貝氏分類器(Naïve Bayesian Classifier)進行預測分析。

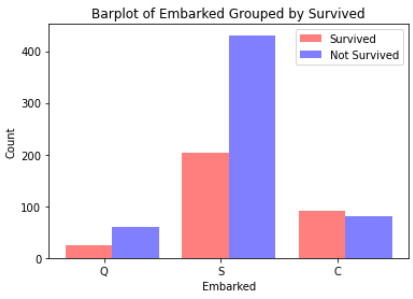
A 探索式資料分析 Exploratory Data Analysis

起初我們認為 Name(姓名)和 Ticket(船票編號)這兩個特徵的資料組成太多元，而 Cabin 的缺失值高達 77%，因而不考慮這些特徵與存活的相關性。接著開始分析沒有缺失值的特徵所對應的生存狀況，作為後續模型選用特徵的參考，可以在 Sex(性別)中發現 Female(女性)的生存率明顯高出許多(圖一)；而在可以在 Embarked(登船口岸)的特徵中發現港口代號為 C 的港口生存率有高於 50%(圖一)。

圖一、特徵為 Sex 的生存狀況

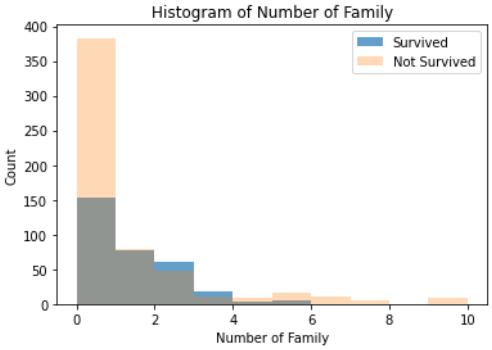


圖二、特徵為 Embarked 的生存狀況

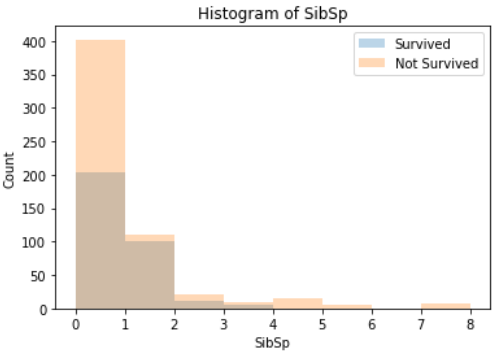


創造新特徵 NumFamily(家人總數)，此特徵為 SibSp (旁系親屬數)和 Parch (直系親屬數)這兩個特徵的總和，發現三個特徵在生存情況的分布相似(圖三、四、五)。

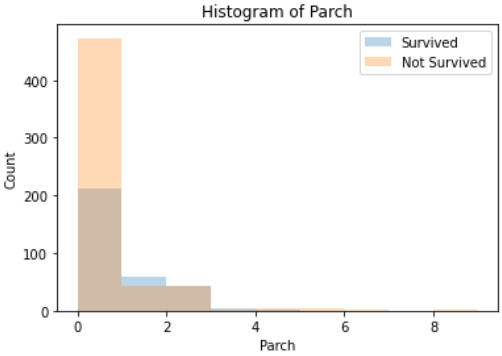
圖三、特徵為 NumFamily 的生存狀況



圖四、特徵為 SibSp 的生存狀況



圖五、特徵為 Parch 的生存狀況



B 資料前處理 Data Preprocessing

- Sex 和 EmbValue 轉換成數值

由於後續流程是決策樹，會需要將資料轉換成數字型態，於是建立 SexValue，Male 轉換成 0，Female 轉換成 1；也建立 EmbValue，將分別將 Q、S、C 轉換成 0、1、2。

- 缺失值(missing value)

看到資料中缺失值的部分，在測試集(training data)中有缺失值的特徵為 Pclass、Age、Fare、Cabin，缺失值狀況如下表一，Fare 缺一筆以平均值替代。

表一、測試集的缺失值數量

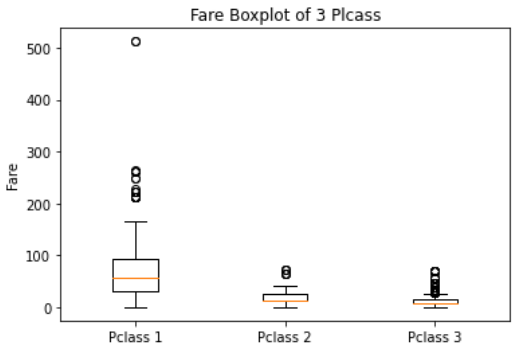
特徵名稱	Pclass	Age	Fare	Cabin
缺失值數量	133	189	1	690
資料佔比(%)	15	21	0	77

- 分析 Pclass 和 LogFare

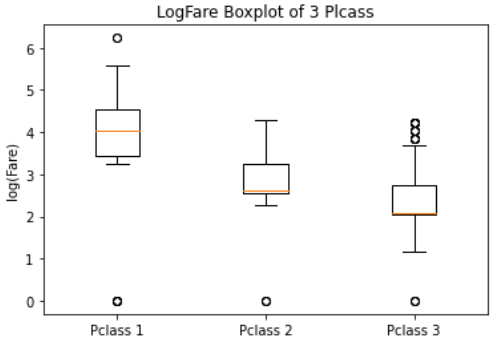
在觀察資料的過程中，我們發現 Pclass 與 Fare 之間可能存在著相關性，但兩者在 boxplot 的資料呈現上較難觀察相關性情況(圖六)，所以將 Fare 指數化縮短差異，建立新特徵 LogFare，與 Pclass 的關係圖如圖三。進一步對 Pclass 和 LogFare 執行 Shapiro 檢定、ANOVA(變異數分析)和 t 檢定(表二)，發現三個 Pclass 在 LogFare 分布上有顯著差異。

下面以 Pclass(1)代表 Pclass 為 1 的資料集合，以此類推。

圖六、3 個 Pclass 的 Fare 分布圖



圖七、3 個 Pclass 的 LogFare 分布圖



表二、ANOVA 和 t 檢定結果(Pclass 對應的 LogFare)

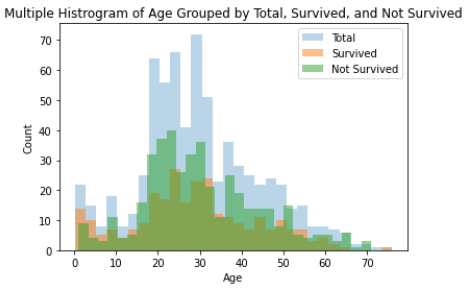
Testing Method	p-value
One-way ANOVA	1.515e-107
t-Test: Pclass(1)&(2)	2.269e-31
t-Test: Pclass(2)&(3)	1.531e-13
t-Test: Pclass(1)&(3)	6.820e-100

接著以 Pclass(2)的上界作為 Pclass(1)和 Pclass(2)的分界，以 Pclass(2)移除數值為 0 的下界作為 Pclass(2)和 Pclass(3)的分界，以這樣的規則去填補 Pclass 的缺失值。

- Age 特徵處理

由於決策樹的特性，需要將 Age 連續性資料切割成離散型，經過分布圖(圖四)觀察到某些年齡區間生存率是大於 50%，最終決定以新特徵 AgeLabel 表示特定年齡區間，以數字 0 表示 15 歲以下的區間，數字 11 表示 66 歲以上的區間，16 歲以上至 65 歲以下則是每 5 歲為一個區間，分別填入數字 1 到 10。

圖八、年齡分布圖



而 Age(年紀)的缺失值則是選擇與整體生存比例相近的區間(表三)以亂數的方式去填補。目的是想在決策樹演算過程中，讓原本是缺失值的資料不會過度影響結果判斷。

表三、特定年齡區間對應之生存率

AgeLabel	Intervals	Survival Rate
-	Total	0.359
3	21~25	0.364
4	26~30	0.348
8	46~50	0.370

- 改良填補 Age 的缺失值的方法，引入 Name 特徵

在競賽結束後，我們發現資料前處理可能不夠完善，因此我們開始考慮 Name 的影響，發現 Name 皆有稱謂的存在，共有 14 種稱謂，且 Name 沒有缺失值。因此建立 Title(稱謂)特徵，取特定稱謂的 Age 平均數或眾數(表四)，可以使用 Title 特徵填補 Age 特徵的缺失值。

表四、特定 Title 的 Age 平均數與眾數

Title	'Mr'	'Miss'	'Mrs'	…	'Ms'
Age Mean	33	24	37	…	26
Age Median	31	23	35	…	26

C 決策樹介紹 Introduction to Decision Tree

- 各項專有名詞定義

1.訊息量：最早由哈萊特(RVL Hartley)在1928年首先提出，希望能將一段文字所攜帶的信息量用科學方式量化，並將此定義為消息可能數m取以2為底的對數。至於取2為底數的原因是對於一必然事件其訊息量為1，而其組成為「此事會發生」和「此事不發生」兩個事件，故以此作為定義，但是在使用時不論以2抑或是以10都不會對於結果產生差異。

$$I(x_k) = \log\left(\frac{1}{P_k}\right) = -\log\left(P_k\right)$$

2.資訊熵：考慮到每件事發生之機率不同，因此此集合資訊雜亂程度乃為各子事件發生機率與訊息量相乘之總和。

$$H(Y) = -\sum_{i=1}^{n} p_i \log\left(p_i\right)$$

3.條件熵：在特定條件下所算出的資訊熵

$$H(Y|X) = -\sum_{i=1}^{n} p_i\, H(Y|X = x_i)$$

4.訊息增益：選擇該分類方式所降低之資料雜亂程度，為資訊熵減去條件熵

$$g(D,A) = H(D) - H(D|A)$$

- ID3 算法生成決策樹

在每個節點上選擇會產生最大信息增益的分枝，分為以下三個部分。

- 開始：在節點上計算所有可能特徵的信息增益，選擇會產生最大信息增益的特徵作為分類依據
- 遞迴：不斷的在子節點上重複上述步驟，不斷分支製造出決策樹
- 停止：每一節點在符合停止條件後即會停止分枝，所有節點均符合停止條件時則停止分枝
- 停止條件: 1.所有類別均相同 2.所有特徵均分類完畢 3.信息增益小於設定值

D 貝氏分類器 Naïve Bayes Classifier

- 模型介紹

Naïve Bayes Classifier 是假設資料裡面的特徵都是相對獨立的情況，運用貝氏定理為基礎的分類器。貝氏定理為已知一些條件之下，某事件發生的機率。所以在所有特徵都是獨立的情況，可以將所有的特徵獨立處理，已知的條件為一個一個的特徵，所以某事件的機率則可以被處理成一個一個獨立的條件機率，再去做相乘，即可以得到某事件的機率。如下面的算式所表示的：

$$P(C|F_1, \ldots, F_n) \propto p(C) \times \sum_{i=1}^{n} p(F_i \mid C)$$

- 實驗過程

因為不知道什麼特徵對結果的影響最大，在希望取到最好的特徵組合的情況之下，我們將所有的特徵組合都測試過一次，我們總共選取了6個特徵，分別是Pclass, Age, NumFamily, LogFare, Sex, Embarked

實驗方式：將Train資料以 Random 8:2 Cross Validation 做100次，取平均的正確率去看這一個組合是否對於結果是更具有影響力的。

實驗結果：可以明顯觀察到具有Sex這一個特徵的正確率明顯較高

表五、有無Sex的特徵組合平均正確率比較表

	有Sex這一個特徵的特徵組合	沒有Sex這一個特徵的特徵組合
平均正確率	0.76254	0.65559

為了避免這是一次剛好出現的結果，所以我在將這一個實驗重複進行了5次，得到的結論皆為相同的，有Sex的特徵組合表現比起沒有的，正確率高了大約10%上下，而且根據這5次的實驗結果，我們選出了6組較為出色，我們認為對結果較具有影響力的組合：

表六、說明6組特徵組合所包含的特徵

	組合內所有的特徵
組合1	SexValue
組合2	AgeLabel, NumFamily, SexValue
組合3	AgeLabel, NumFamily, SexValue, EmbValue
組合4	SexValue, EmbValue
組合5	SexValue, EmbValue, LogFare
組合6	Pclass, SexValue

最後的結果就是用這6個特徵組合，去預測Test Data的結果，為了避免Overfitting，我們有使用Cross Validation去確定結果是否有Overfitting的問題，而結果比起原本的平均正確率都高了大概1-2%。

II. 結果與討論

A 決策樹結果 Results of Decision Tree

1.

表七、分類特徵與最佳模型準確度一覽表

分類特徵	準確度
SexValue	0.791874
SexValue, NumFamily	0.803826
SexValue,NumFamily, AgeLabel	0.806226
SexValue, EmbValue	0.791874

2.原因分析：根據前頁所述，由於男性和女性生存率相差懸殊，因此表現較為良好的模型均採用性別做為首要的分類依據，在僅僅性別作為分類依據的模型下，其準確度依然高達79.1%。

而在加入其他特徵作為分類依據後，所分枝出的決策樹並不會將男性節點繼續分枝。原因是男性的死亡率高達81.6%，導致分枝信息增益過小而不進行分枝，**即便進行分枝，也會因為死亡比率仍大於50%而導致相同結果**。而女性的部分比率差距相對不懸殊，因此可以透過分枝增加準確度。

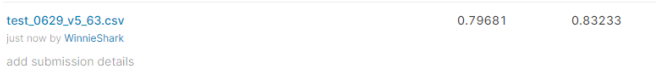
首先選擇加入家庭人數進行分枝，決策樹顯示**家庭人數為4人以上之女性死亡率極高**，全數改判定為死亡。而從訓練資料推測，準確度上升的原因推測為高人數家庭的家庭數極其稀少，導致4人以上的家庭數和是否為家人具有高相關性，而在大家庭當中家人間存活率相關性也很高，因此產生了類似數學上鏈鎖率的效應，導致家庭人數和存活率具有高相關性。

接下來在加入年齡作為判斷依據，結果顯示準確度再次升高。根據決策樹顯示它將大家庭判定為死亡，單身漢判定為存活，並在其他資料中做更為細緻的年齡區分，由於分支已過於複雜因此無法找出規律，不過依然顯示出**年齡對於存活率是具有決定性的因素**。

B 貝氏分類器結果 Results of Bayes Classifier

這是最後調整之後的結果，可以看見正確率在經過我們重新調整前處理得方法之後，有著顯著的提升(圖九)：

圖九、貝氏分類器預測結果(Private: 0.79681; Public: 0.83233)



在最終調整下，我們想觀察出會決定性的影響存活與否結果的特徵，其中與決策樹那邊的結果十分一致的是，Sex與NumFamily都是決定結果的重要關鍵，在表現好的特徵組合裡面，都有Sex與NumFamily的存在。其他的特徵，在最終選出的6個組合中，與這兩個重要特徵搭配的還有Pclass, Age, Embarked，我們認為這是次重要的特徵，可以再加強這一個預測模型的準度。而新加入的Title，在我們的意料之外的，原本以為會成為我們正確率調高的關鍵，但卻在表現好的組合之中，都沒有出現，不過還是讓我們可以更好的填補Age的缺失值，我想一定程度上還是有處理這一個特徵的必要性。

III. 結論

從上述可知，對於預測鐵達尼號資料集的生存率，貝式分類器優於決策樹分類器。而原因推測為於此決策樹所接收的年齡、票價資料均已進行離散化處理喪失了某些資訊，導致分類上的準確度出現差距。因此我們最終送交貝式分類器所預測之結果，其正確率為81.10%。

參考資料

[1] 決策樹各項專有名詞定義: https://arbu00.blogspot.com/2018/01/10.html

[2] 鐵達尼生存預測手把手資料分析實戰教學: https://aifreeblog.herokuapp.com/posts/64/Data_Analytics_in_Practice_Titanic/

[3] 預測鐵達尼號旅客生存機率: http://ielab.ie.nthu.edu.tw/108_IIE_project/2/108IIE_proj2_4_word.pdf