計算生物學原理與應用期中報告

第五組組員：
公衛三 b07801003 梁嫚芳　公衛三 b07801054 劉昀真
公衛三 b07801013 劉德駿　流預所 博三 d07849008 邱偉珉

研究主題

預測乘客對航空公司的搭機滿意度

背景介紹

此資料來源為 kaggle，內容主要是某航空公司對乘客進行滿意度調查，看看乘客對於搭乘本公司飛機是否滿意，將利用此筆資料找出哪些變數主要會影響乘客的滿意度，並配適一個模型來預測乘客對於航空公司的滿意度，最後與原始資料的滿意度進行比較，判斷預測結果與實際情形是否吻合。

預期使用的研究方法與材料

預期研究材料

我們期末報告使用的資料裡，總共有 129,880 筆乘客，並且詳細記載了每一位乘客填寫的 23 個寶貴的問題 (即解釋變數)，關於乘客對於每一個問題可能的回答，我們詳細的統整在表 1 裡，可以發現有 5 個問題屬於名目尺度資料； 14 個問題屬於順序尺度資料； 4 個問題屬於比例尺度資料。在初步的資料整理中，我們排除了一些資料可能會有的問題 (例如 : 遺失值、不正確的回答，以及乘客可能會有重複填寫問卷等等問題)，最後，剩下了 129,487 筆完整的資料。

在初步的探索性資料分析中，129,487 筆資料裡有 6,3784 位男性乘客和 6,5703 位女性乘客，不同性別的比例大致相同。藉由圖 1 可知，男性乘客與女性乘客的年齡範圍都是落在 7 歲到 85 歲之間，並且都一樣顯示在 20 至 30 歲和在 40 歲左右的乘客搭乘該航空公司的人數最多。忠實與不忠實的乘客分別是 105,773 位與 23,714 位；乘客乘坐商務艙與經濟艙的人數，分別為 61,990 位與 67,497 位。

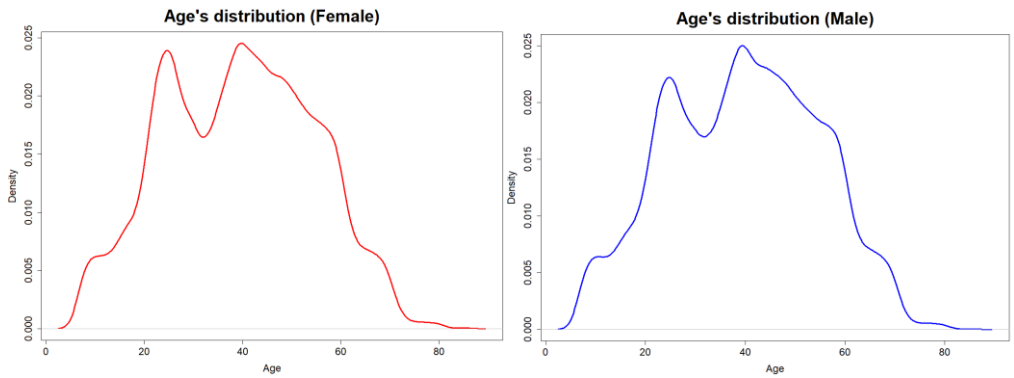

圖 1 : 男性乘客與女性乘客的年齡分佈圖

表 1 : 某航空公司對乘客提問 23 個問題與每一個問題乘客可能回答

No.	Covariate	Data presentation	Scale of data
1	Gender	Male (default = 1); Female (default = 0)	Nominal
2	Customer Type	Loyal Customer (default = 1); disloyal Customer (default = 0)	Nominal
3	Age	7 - 80 (numeric)	Ratio
4	Type of Travel	Business travel (default = 1); Personal Travel (default = 0)	Nominal
5	Class	Business; Eco; Eco Plus	Nominal
6	Flight Distance	31 - 4983 (numeric)	Ratio
7	Inflight wifi service	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
8	Departure/Arrival time convenient	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
9	Ease of Online booking	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
10	Gate location	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
11	Food and drink	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
12	Online boarding	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
13	Seat comfort	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
14	Inflight entertainment	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
15	On-board service	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
16	Leg room service	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
17	Baggage handling	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
18	Checkin service	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
19	Inflight service	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
20	Cleanliness	Satisfaction level (0-5) (0 : Not Applicable; 5 : Applicable)	Ordinal
21	Departure Delay in Minutes	0 - 1592 (min) (numeric)	Ratio
22	Arrival Delay in Minutes	0 - 1584 (min) (numeric)	Ratio
23	satisfaction	satisfied (default = 1); neutral or dissatisfied (default = 0)	Nominal

預期研究方法

基因演算法

我們期末報告使用感興趣的反應變數是第 23 個問題 (即 satisfaction)，其餘的問題都皆表示為解釋變數。我們預計會用 Sensitivity、Specificity 來評估預測出來的 satisfaction 與真實觀察到的 satisfaction 之間的準確性，作為我們的目標函數 (如果滿意度只有 1-5 分，可能會用 cross-entrophy loss 作為目標函數) 。最後結論將會放上描述性統計或是資料視覺化。

預期結果與結論

目標：我們想透過演算法得知乘客對於航空公司的服務滿意度主要是受到哪些解釋變數的影響，接著選出有代表性的解釋變數，來進一步最佳化參數，使得預期的滿意度最大化。

結論： 透過知道哪些解釋變數會影響滿意度，以及這些解釋變數適合的最佳參數，除了可以討論什麼類型的顧客會較容易滿意，以及顧客對機場服務的態度是外歸因(服務本身不夠周到)還是內歸因(乘客本身的個人因素，和服務無關)？