**T\_Brain趨勢科技設計文件及檢討**

**設計流程:**

1. 訂單(order.csv)
   1. 把給定的source\_1、source\_2、unit、people\_amount當作Category
   2. order\_date提取出年月日當特徵。
2. 團體(group.csv)
3. 把給定的sub\_line、area、days當作類別特徵。
4. price直接當成連續特徵。
5. begin\_date提出年月日當特徵。
6. 把begin\_date加上days算出回台灣的日期，並提出年月日特徵。
7. 判斷旅客是否是在國訂連續假日出遊，當成特徵。
8. 航班(airline.csv)
   1. 透過go\_back次數，分別算出來回轉機次數和總共轉機次數。
   2. 把src\_airport和dst\_airport透過map轉成各個國家當成特徵，再次map成各大洲當特徵。
   3. 把fly\_time和arrive\_time分別轉成抵達的時段，ex: 早中晚。

**遇到問題**

問題1:

在做join的時候，出現問題。

問題2:

一開始做EDA，發現各項特徵都不明顯，所以跑出的模型都很糟，後來發現是Class Balance問題。

問題3:

太過依賴for迴圈，由於之前主要寫C#和Java，習慣用For，這次的數以

練習的數據都大很多，導致跑很慢。

解決辦法1:

1. 應該事先在讀檔案前就先做型態的處理，這樣讀進來比較好處理。

解決辦法2:

1. 透過Sampling解決，嘗試過SMOTE，發現太容易Overfitting，很大的原因是因為SMOTE造出來的數據是有問題，所以後來決定透過Undersampling，分數就有明顯的提升。
2. 後來看了別人的文件發現也可以從algorithm中做調整，也可以設定scale\_pos\_weight。

解決辦法3:

改用pandas的內建判斷或是用Apply、Transform，提升速度。

**總結及檢討**

看完其它前幾名的文件，發現自己特徵做的太少，大部分都好幾百個特徵甚至上千，然後做到七成五左右，自己大致只做了接近20個而已，所以只有6成初，自己太注重EDA的呈現才決定是否加入特徵，應該要先利用Groupby類別特徵，直接做出大量的統計特徵，包含Mean、Median…等，然後再搭配EDA來解釋；但發現自己有些特徵蠻特別，把這些特徵加到前幾名的程式碼，在驗證資料中是有提升一些分數的，包括是否為國定假日和出發抵達時間為早中晚，亦或是哪個國家，雖然，因為其實原本的area特徵是被雄獅碼掉的，所以大部分參賽者都是直接丟這個進去。前幾名都有做兩個特徵，整個團的總人數、每團出發日期減去每團最後一人報名時間，個人認為這兩點不算特徵，在出發之前是無法得知整團人數的總和，也無法得知最後一人報名的時間。

這次的比賽由於是用下班時間和自己組隊參加，所以大部分地時間都在處理遇到的問題，在特徵工程和條參數上面相對花了較少時間，雖然的確排名沒有很靠前，但也因為全部都是自己找答案和寫程式，對於整個架構的流程，我學習到了很多，也很謝謝這次的主辦單位，提供了這麼好的機會。

|  |  |  |  |
| --- | --- | --- | --- |
| 訂單(order.csv, 主資料表) | | 團體(group.csv) | |
| 欄位名稱 | **欄位說明** | **欄位名稱** | **欄位說明** |
| group\_id | 團號 | group\_id | 團號 |
| order\_id | 訂單編號 | sub\_line | 副線別(碼掉) |
| order\_date | 下單日期 | area | 地區(碼掉) |
| source\_1 | 訂單來源1(碼掉) | days | 天數 |
| source\_2 | 訂單來源2(碼掉) | begin\_date | 出發日 |
| unit | 負責單位(碼掉) | price | 價格 |
| deal\_or\_not | 是否成行 | product\_name | 網上銷售團名 |
| people\_amount | 訂單人數 | promotion\_prog | 行程特色 |
| 航班(airline.csv) | | **日行程(day\_schedule.csv)** | |
| 欄位名稱 | **欄位說明** | **欄位名稱** | **欄位說明** |
| group\_id | 團號 | group\_id | 團號 |
| go\_back | 去/回程 | day | 第幾天 |
| fly\_time | 起飛時間 | title | 日行程主題 |
| src\_airport | 起飛機場 |  |  |
| arrive\_time | 抵達時間 |  |  |
| dst\_airport | 抵達機場 |  |  |