2019秋季 巨量資料分析課程 期末報告企劃書

Expedia -「訂房預測分析」

國立成功大學 TNCBD隊

隊長
工資管系 蘇威誠

隊員
工資管系 黃瀚陞
統計學系 郭士銘

指導教授
統計系 李政德 老師

目錄

一、題目----------3

二、前言與摘要 ----------3

三、大數據分析流程及結果 ----------3

- 分析流程與了解領域現況----------3
- 了解此次專案數據 ----------4
- 敘述統計 ----------7
- 資料預處理 ----------8
- 機器學習嘗試 ----------9
- 以特定feature預測 ----------10

四、結論 ----------11

一、題目

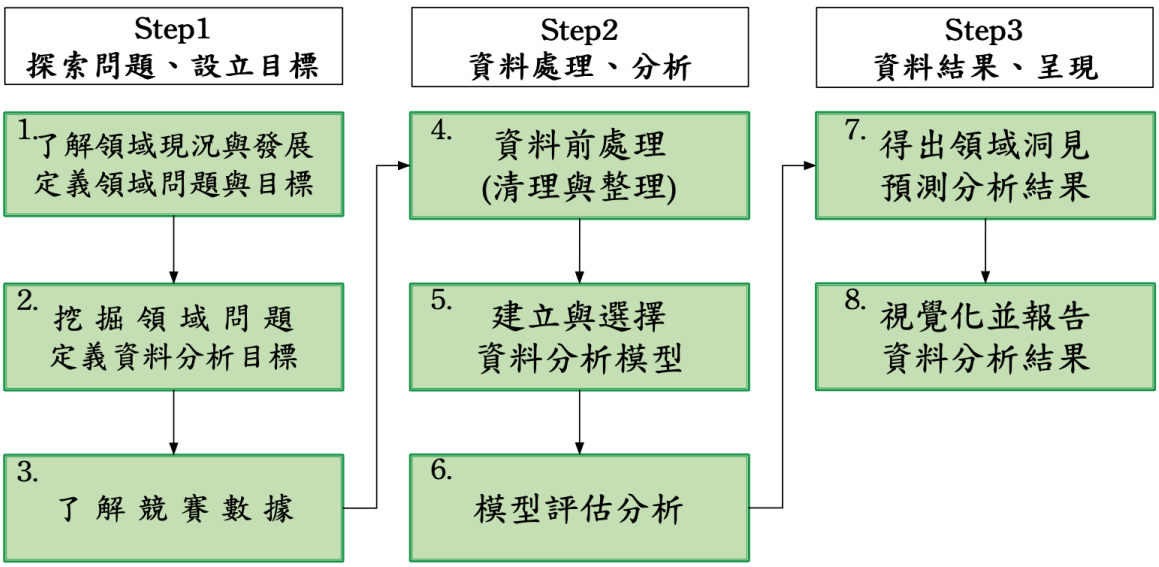
Expedia 訂房預測分析

二、前言與摘要

Expedia是源於美國的綜合旅遊服務網站，提供機票預訂、酒店預計、汽車出租、遊船等多元化的旅遊服務。Expedia希望向用戶提供客制化的酒店推薦，消除酒店搜索的麻煩。Expedia一個每月約有數億訪問者，這不是一件容易的事！目前，Expedia使用搜索參數來調整他們的酒店推薦，但沒有足夠的客戶使特定數據為每個用戶客制化它們。在這研究中，Expedia向Kagglers挑戰了將客戶數據關聯起來並預測用戶將入住100個不同酒店集團的可能性。

根據使用本身基本一些資料(顧客的所在國家、區域、洲別等)以及他下的搜尋指令(從甚麼管道)，來去預測使用者最後會訂哪一群飯店種類(hotel cluster)。而hotel cluster的分類是根據飯店的房間價格和之前的訂房人數所定義出來的飯店種類群，總共有1-100種。

三、大數據分析-(1)分析流程



1. 了解領域現況與發展定義領域問題與目標

Expedia提供客戶行為的數據，包括客戶搜索的內容、他們與搜索結果（點擊/預訂）的互動方式與搜索結果是否是旅行套餐。本研究中的數據是Expedia的隨機選擇，並不代表整體統計數據。

Expedia對於預測用戶將預訂哪個酒店集團有興趣。Expedia使用內部算法來形成酒店群，將相似的搜索酒店（如：基於歷史價格，顧客星級，相對於市中心的地理位置等）組合在一起。這些酒店群可以有效地識別客戶要預訂的酒店類型，同時避免出現離群值，例如：沒有歷史數據的新酒店。

而這份資料與研究的目標是根據用戶事件的搜索和與該用戶事件相關的其他屬性，進而預測該用戶事件的預訂結果（也就是酒店群）。訓練和測試數據集是根據時間劃分的：2013年和2014年的訓練數據，而測試數據則是2015年的數據。公共/私人排行榜數據也按時間劃分。培訓數據包括日誌中的所有用戶，包括點擊事件和預訂事件。測試數據僅包括預訂事件。

2. 了解競賽數據

- train.csv

  資料描述：主要為資料訓練集，用以訓練模型

  筆數：37,670,293 筆

  變數個數：24個

	date_time	site_name	posa_continent	user_location_country	user_location_r
0	2014-08-11 07:46:59	2	3	66	348
1	2014-08-11 08:22:12	2	3	66	348
2	2014-08-11 08:24:33	2	3	66	348
3	2014-08-09 18:05:16	2	3	66	442
4	2014-08-09 18:08:18	2	3	66	442

- test.csv

  資料描述：主要為資料測試集，用以評估模型的實際性能

  筆數：2,528,243 筆

  變數個數：22個

	id	date_time	site_name	posa_continent	user_location_country	user_locati
0	0	2015-09-03 17:09:54	2	3	66	174
1	1	2015-09-24 17:38:35	2	3	66	174
2	2	2015-06-07 15:53:02	2	3	66	142
3	3	2015-09-14 14:49:10	2	3	66	258
4	4	2015-07-17 09:32:04	2	3	66	467

- **變數名稱與變數說明**

Column name	Description	Data type
date_time 資料的時間	Timestamp	string
site_name 登入網站的ID(依國家分)	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)	int
posa_continent 那個國家的州別	ID of continent associated with site_name	int
user_location_country 使用者的國家	The ID of the country the customer is located	int
user_location_region 使用者所在區域	The ID of the region the customer is located	int
user_location_city 使用者所在城市	The ID of the city the customer is located	int
orig_destination_distance 使用者在搜尋時刻 所在位置與飯店距離	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated	double
user_id 使用者的ID	ID of user	int
is_mobile是否用手機連結 1 是用手機；0 其他裝置	1 when a user connected from a mobile device, 0 otherwise	tinyint
is_package此預訂 是否為套裝行程。 1是套裝行程；0 其他	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise	int
Channel 行銷管道的ID	ID of a marketing channel	int
srch_ci check in 日期	Checkin date	string
srch_co check out日期	Checkout date	string
srch_adults_cnt 搜尋幾位大人	The number of adults specified in the hotel room	int
srch_children_cnt 搜尋幾位小孩	The number of (extra occupancy) children specified in the hotel room	int
srch_rm_cnt 搜尋幾間房間	The number of hotel rooms specified in the search	int
srch_destination_id 搜尋的地址	ID of the destination where the hotel search was performed	int
srch_destination_type_id 目的地的種類	Type of destination	int
hotel_continent 飯店所在洲別	Hotel continent	int
hotel_country 飯店所在國家	Hotel country	int
hotel_market 飯店的市場 *	Hotel market	int
is_booking * 是否有預訂1 有0 只是參閱	1 if a booking, 0 if a click	tinyint

- **網頁操作與資料連結**

- destination.csv

  資料描述：針對srch_destination_id這個變數進行補充，並且分為149個特徵

  筆數：62,016筆

  變數個數：149個（srch_destination_id & di, i = 1,2,,,149 ）

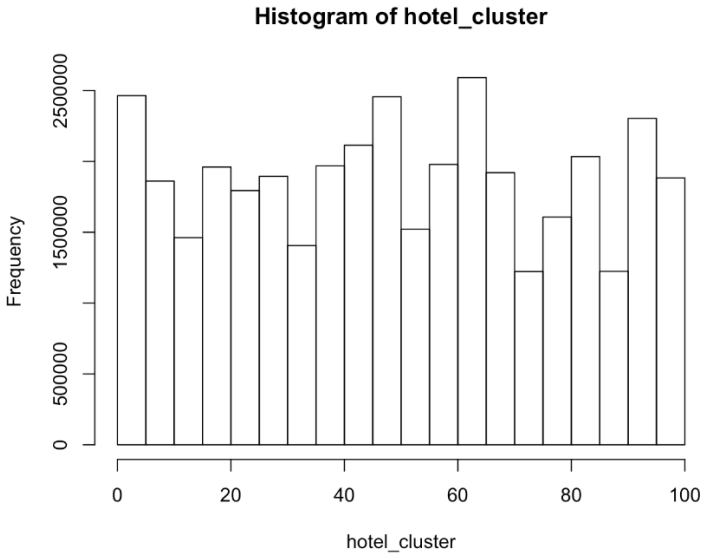
  變數名稱：(1) srch_destination_id 為使用者搜尋要去的地點

  (2) di, i = 1,2,,,149

Feature Name	Feature Description	Feature Data Type
srch_destination_id	ID of the destination where the hotel search was performed	int
d1-d149	latent description of search regions	double

三、大數據分析-(2)敘述統計

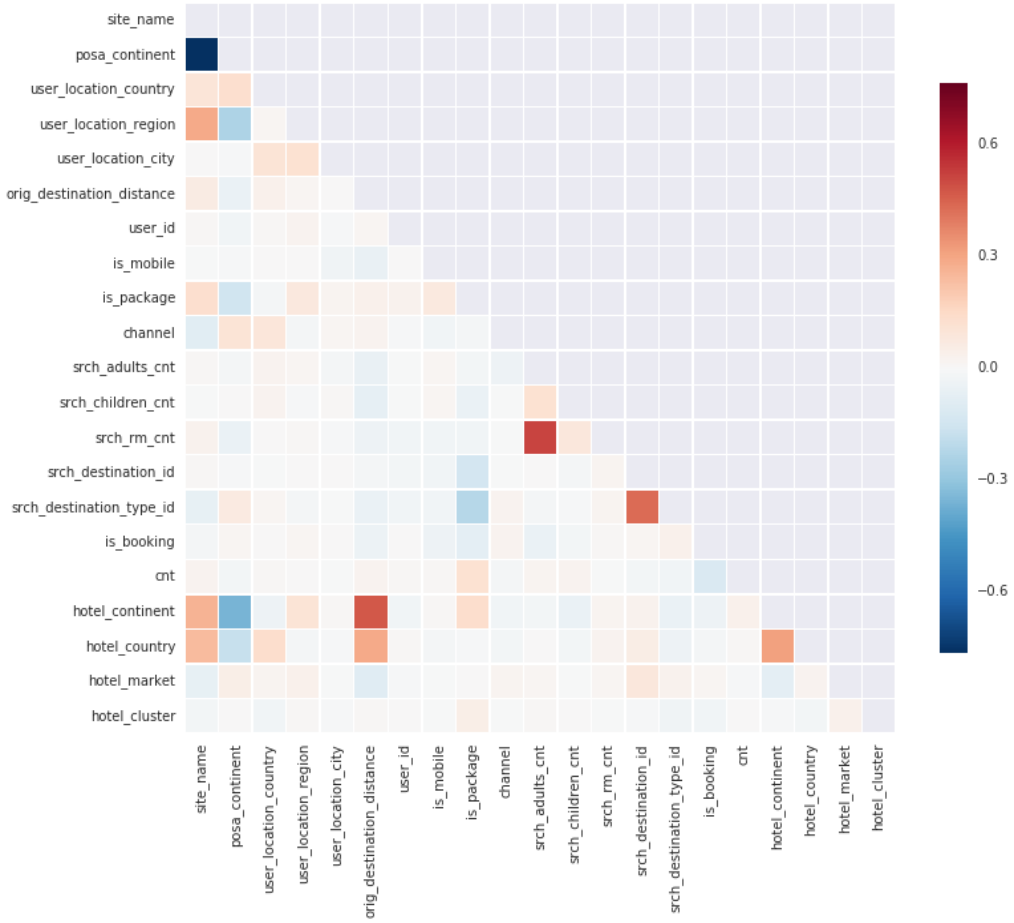
- hotel_cluster




```
91    1043720
41     772743
48     754033
64     704734
65     670960
       ...
35     139122
53     134812
88     107784
27     105040
74      48355
```

資料描述：每一個種類的hotel_cluster大致上呈現均勻分布，會呈現如此均勻分布推測應該是因為這些飯店已經由Expedia針對飯店的各種屬性進行分群（其中應該也已包含了訂房的人數狀況），故每種飯店群都有其客群。

- Correlation



從上方繪製的各變數相關係數圖表中可以發現雖然某些變數之間有呈現顯著，但大多變數間並沒有太多的相關。而且各變數與我們想要預測的hotel_cluster皆呈現低度的相關。下表為所有變數與hotel_cluster前三高之正相關與負相關係數與變數，可以看到皆呈現低度相關：

變數名稱	相關係數
is_package	0.038733
hotel_market	0.034205
srch_children_cnt	0.016261
srch_destination_type_id	-0.032850
hotel_country	-0.024289
site_name	-0.022408

三、大數據分析-(3)資料預處理

- 從date_time切出year, month, day, hour, minute, dayofweek, quarter
- 透過srch_ci和srch_co算出停留時間(stay span)
- test 拿掉 id 這個欄位
- test 新增 is_booking 這個欄位，並且全部設為1
- 隨機從train抽取10000個不同的id，並把這些id的資料從train裡面抽出來，形成一個較小的train (down sampling)

	month	day	hour	minute	dayofweek	quarter	site_name	posa_continent	user_location_country	user_location_region
12	1	17	6	24	4	1	2	3	66	318
13	1	18	14	33	5	1	2	3	66	318
14	1	21	6	39	1	1	2	3	66	318
15	1	21	6	40	1	1	2	3	66	318
16	1	22	6	10	2	1	2	3	66	318

- 針對destination資料表裡的d1-d149這149個欄位做PCA，降維後留下三個欄位既保留足夠資訊量，又可以節省模型運算時間

	srch_destination_id	d1	d2	d3	d4	d5	d6
0	0	-2.198657	-2.198657	-2.198657	-2.198657	-2.198657	-1.897627
1	1	-2.181690	-2.181690	-2.181690	-2.082564	-2.181690	-2.165028
2	2	-2.183490	-2.224164	-2.224164	-2.189562	-2.105819	-2.075407
3	3	-2.177409	-2.177409	-2.177409	-2.177409	-2.177409	-2.115485
4	4	-2.189562	-2.187783	-2.194008	-2.171153	-2.152303	-2.056618

(此圖為destination資料表的截圖，每個destination_id都會有d1~d149共149個特性)

ci_day	co_day	ci_dayofweek	co_dayofweek	ci_quarter	co_quarter	stay_span	0	1	2
17.0	20.0	3.0	6.0	2.0	2.0	72.0	-0.14173	-0.140009	-0.072771
16.0	19.0	2.0	5.0	2.0	2.0	72.0	-0.14173	-0.140009	-0.072771
17.0	20.0	3.0	6.0	2.0	2.0	72.0	-0.14173	-0.140009	-0.072771
18.0	20.0	4.0	6.0	2.0	2.0	48.0	-0.14173	-0.140009	-0.072771
18.0	20.0	4.0	6.0	2.0	2.0	48.0	-0.14173	-0.140009	-0.072771

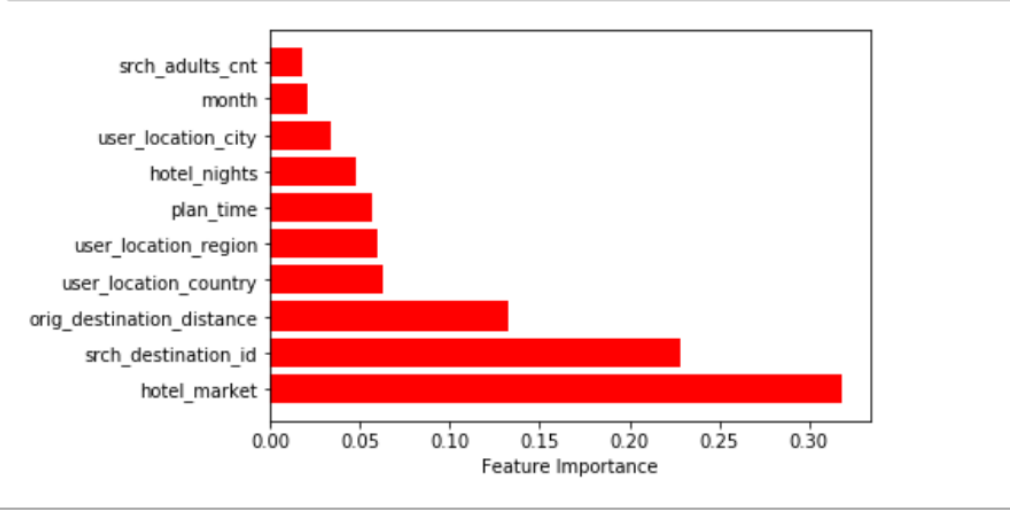
（針對destination_id做主成分分析過後，留下三個可以解釋destinaion_id的主成分，並且加回train、test資料表當中）

三、大數據分析-(4)機器學習嘗試

針對此份已經清洗與預處理完成過後的資料資料，我們根據我們過往的習慣，先進行了Random Forest、Logistic Regression、CatBoost三種機器學習嘗試。

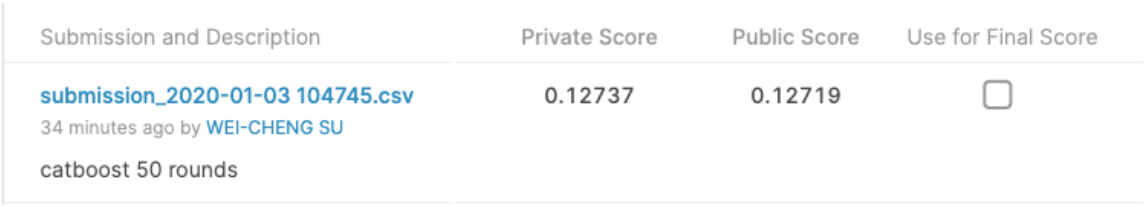
在預測上，因為可以輸出五個結果，因此我們使用了predict_proba這個函式，讓預測後的輸出結果是各種結果的機率，接著再取機率前五大的當作輸出結果。

我們亦針對各個模型挑選並繪製出模型中重要的變數（如下圖），並且調整放入模型中預測的變數項目，希望藉此提升預測準確度。



最後，根據測試結果分數顯示，Random Forest、Logistic Regression的預測效果非常差，CatBoost的表現相對較好，但也不盡理想。推測應該是因為我們要預測的hotel cluster總共有一百種，種類太多，所以才會造成預測不準。

方法嘗試	分數
Most Common Clusters （直接填入出現頻率最高的飯店群）	0.5997
Random Forest	0.041~0.043
Logistic Regression	0.032~0.033
CatBoost	0.10~0.13



三、大數據分析-(5)以特定feature預測

因為發現使用機器學習所得到的效果皆不太好，但如果直接填入出現頻率最高的飯店群會有不差的分數，因此我們嘗試直接從特定幾個重要的feature中以某一feature為基準，利用訓練資料集，檢視該feature不同的數值中，各自最受歡迎的五個hotel_cluster。以下舉srch_destination_id來說明：

srch_destination_id	is_booking	hotel_cluster
大阪	1	20
大阪	1	22
大阪	1	20
大阪	1	21
大阪	1	25
大阪	0	22
大阪	0	25

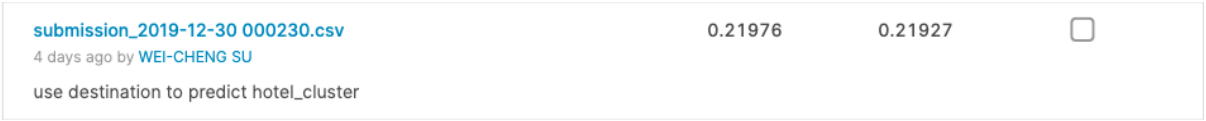
- 先利用訓練資料集，看不同destination最受歡迎的前五個hotel_cluster
- 若以大阪來舉例，我們把所有destination在大阪的資料找出來
- 因為測試資料集中的資料皆有訂房(is_booking=1)，所以我們再根據該筆資料是否有訂房給予不同的權重，有訂房為1，沒訂房為0.15
- 統計完所有hotel_cluster得到的權重，就可以知道大阪大家最常訂的前五個hotel_cluster
- 以此類推，計算所有的destination最受歡迎的前五個hotel_cluster。（如下表）

hotel_cluster	sum
20	1+1=2
21	1
22	1+0.15=1.15
25	1+0.15=1.15

根據我們在機器學習模型中得出的重要變數，我們總共嘗試了幾種不同的欄位來做以上的方法，結果如下：

方法嘗試	分數
use "srch_destination_id" to predict "hotel_cluster"	0.21976
use "hotel_market" to predict "hotel_cluster"	0.21330
use "hotel_country" to predict "hotel_cluster"	0.13615
use "user_location_country" to predict "hotel_cluster"	0.08570

最後結果發現使用srch_destination_id的結果最好，分數為0.22



四、結論

在此次的報告實作當中，一開始我們使用之前習慣的機器學習手法來預測hotel_cluster，結果發現使用機器學習預測的效果皆不如預期，甚至有些機器學習模型還比直接填入出現頻率最高的結果還要來的差。

為了要有更好的結果，我們試著利用其他的方法預測，發現可以直接從資料當中探索出有用的資訊，利用最關鍵的幾項因素來進行預測或是加權，進而去計算出前五名受歡迎的飯店群。此法的預測效果（不論是預測時所需花費的電腦運算時間還是最後的準確率）皆有顯著提升，這都是我們在開始實作這份報告前預想不到的。

不過，使用單一因素來預測出來的結果儘管優於使用機器學習模型，但仍舊有不少進步空間。由於目前我們只有使用一項因素來進行計算前五個受歡迎的hotel_cluster，因此我們認為，若是可以針對多個欄位的因素，加權或交叉配對各因素後再找出前五個最受歡迎的hotel_cluster，可能會有更好的結果。另外，由於一個ID會對應到好幾筆的搜尋資料，若是可以針對個別的每一個ID都進行建模或是分析，應該也會更佳結果。再者，若是可以針對變數以及資料本身再作分析並且留下重要資料，也許可以使用深度模型的方式進行預測，一可能會有更高的準確率

最後在商業應用上，由此次預測可發現，使用者選擇何種飯店群跟「目的地」、「飯店市場」有較大關係，因此除了可以推薦使用者飯店種類，也可以試著跟其他業者合作，連帶推薦目的地相關的套裝行程、目的地景點介紹或是，使Expedia的網站不僅僅只是個訂房網站，而可以有更多與其他行業合作的可能。

附錄、組員分工

- 蘇威誠：資料預處理、嘗試以特定feature預測、投影片製作
- 黃瀚陞：敘述統計、機器學習模型嘗試、投影片製作、報告
- 郭士銘：書面整理