題目：

阿竹是一個旅館銷售人員，負責某間旅館的銷售業務，他想針對旅館網頁的訪客進行廣告優惠，但是行銷經費有限，只能針對較有可能消費的訪客進行廣告。阿竹提供了旅館網站的Google Analytics流量統計資料，請資料工程師幫助阿竹找出可能進行消費的訪客。

1. 基本題：預測訪客是否會訂房，請依照流量統計資料train.csv，訓練一個分類模型或回歸模型，預測test.csv中每位訪客是否會進行消費的機率，並依照sample\_submission.csv的格式提交預測結果。請記錄資料分析、清理、建模、模型評估的過程，將完整過程、清理後資料與相關程式碼上傳至個人github並提供連結或是寄至 adrian.chen@gmail.com

\* 數據包含10個數字屬性和8個分類屬性

\* 'Revenue' 屬性將作為標籤 1 代表會訂房 0 代表不會

\* 使用訓練資料 train.csv 建立模型，預測 test.csv 是否會訂房

\* 屬性說明：

- Administrative：管理頁面瀏覽數量

- Administrative\_Duration：管理頁面瀏覽時間

- Informational：資訊頁面瀏覽數量

- Informational\_Duration：資訊頁面瀏覽時間

- ProductRelated：產品頁面瀏覽數量

- ProductRelated\_Duration：產品頁面瀏覽時間

- BounceRates：跳出率

- ExitRates：退出率

- PageValues：頁面價值 ----

- SpecialDay：特殊日期（例如聖誕節，情人節）接近程度

- Month：訪問月份

- OperatingSystems：作業系統

- Browser：瀏覽器

- Region：地區  
 - TrafficType：流量來源

- VisitorType：訪客類型

- Weekend：訪問時間是否為週末

- Revenue：是否訂房

1. 資料前處理(沒有特別針對遺漏值、outlier進行處理，因最後採用Xgboost)
2. 把PageValue轉成dummy和revene做混淆矩陣(F1-0.671) 沒做任何模型下，似乎也是個辦法。

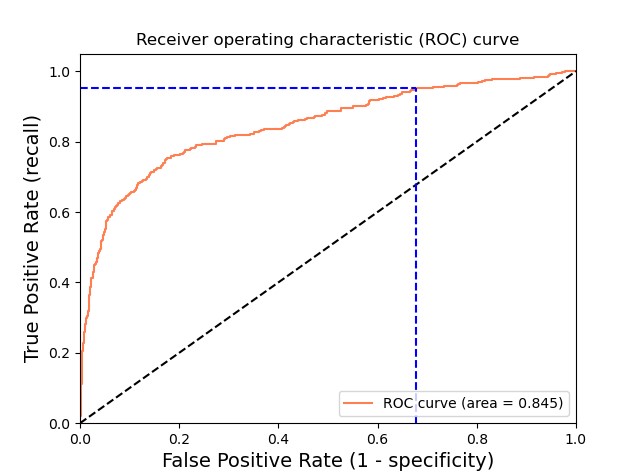
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Accuracy 87.53%  Recall 57.58%  Precision 80.26%  F1 67.06% | |  |  |  |  | | --- | --- | --- | --- | |  | | Revenue | | | 1 | 0 | | PageValue | 1 | 1025 | 755 | | 0 | 252 | 6048 | |

1. 採用xgboost分析 Grid CV 調參的結果似乎比**不調**、**random CV**的結果好

以下為Train.csv資料結果

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Accuracy Score: 91.72%  Precision Score: 78.48%  Recall Score: 65.39%  F1 score: 71.34% | |  |  |  | | --- | --- | --- | |  | 1 | 0 | | 1 | 835 | 442 | | 0 | 229 | 6594 | |

1. 嘗試過logistic 回歸，效果沒有xgboost出色



1. 後續建議：

a.可加入天氣、疫情、訂房個資….等外部變數預測進行預測

b.可區分區域(訂房遊樂景點區分，四季旅遊區域差異)

c.其他ensemble、RF、ANN模型進行預測

d.此次分析沒有特別針對敘述性統計、變數背景細部了解🡪要提升模型這步驟或許較為重要。

2. 加分題：延續上述案例，請依照以下二個混亂矩陣，評估何者較適合用於本案例的情境，將分析過程上傳至個人的github並提供連結或是寄至 adrian.chen@gmail.com

Model A

Accuracy (853+7230)/(9000)=0.89

Precision (853)/(853+341)=0.71 –可能造成房間準備過剩 (成本增加)

Recall (853)/(853+576)=0.60–可能造成房間 (庫存不足)

F1 = (2\*0.71\*0.60)/(0.71+0.60)=0.650

Model B

Accuracy (846+7255)/(9000)=0.90

Precision (853)/(853+316)=0.73 –可能造成房間準備過剩 (成本增加)

Recall (853)/(853+583)=0.59–可能造成房間準備(庫存不足)

F1 = (2\*0.73\*0.59)/(0.73+0.59)=0.652

* 整體模型來說 model B F1 score較佳
* 若很在乎預測準確度選model B (不要過度準備)
* 若很在乎房間庫存量model A (庫存不足，預測沒訂，最後下訂!)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model A | | | Model B | | |
|  | 預測為  訂房 | 預測為  沒訂房 |  | 預測為真 | 預測為假 |
| 實際訂房 | 853 | 576 | 實際為真 | 846 | 583 |
| 實際  沒訂房 | 341 | 7230 | 實際為假 | 316 | 7255 |