機器學習技法期末報告 組別: JSY

一、組員

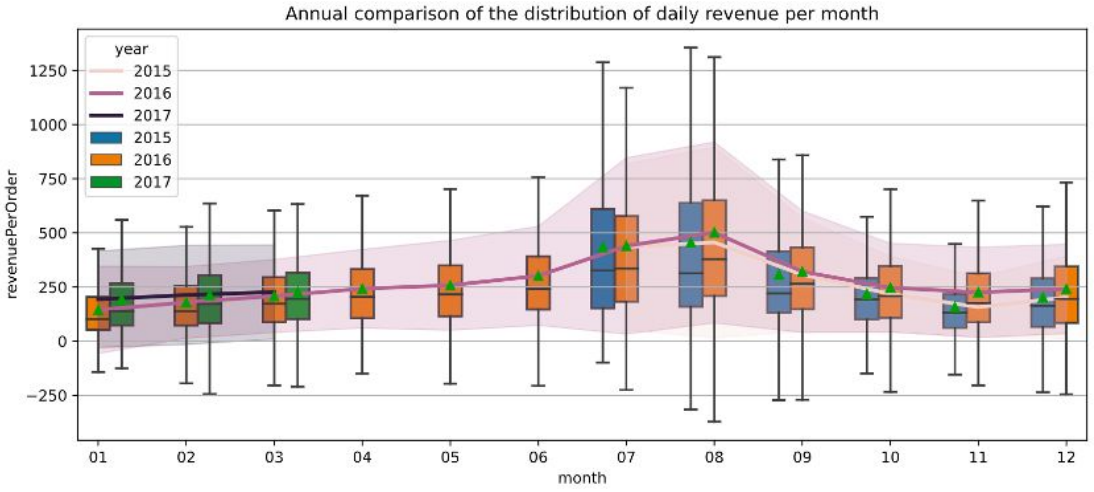
學號	姓名	負責內容
R09922a04	黃品硯	模型搭建，模型優化，報告撰寫
D08522022	簡信堂	嘗試各種不同類型模型，報告撰寫
M10815802	周圓	資料分析處理，報告撰寫

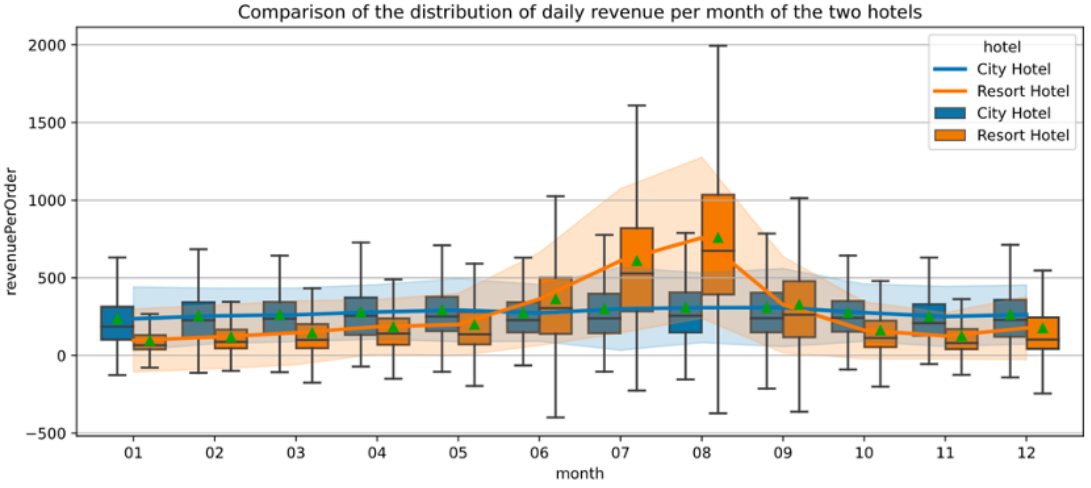
二、資料分析 & 資料處理

該資料集是關於兩間旅館在 2015年7月-2017年8月 間的訂單記錄，我們的目標是預測訂單是否會被取消 (is_canceled)，以及該筆訂單的 Average Daily Rate (adr)。在開始實作之前，我們對該資料集做了一系列分析，結果如下。

1. 訂單數量和取消率的關係

旅館行業全年的營收並不均勻，受時間影響可能存在很大的波動，我們對所有訂單的總收入與時間的變化關係做了分析，發現在每年的夏天（6月- 9月）是旅館營收增長的高峰期，且這種變化趨勢在不同年份中都是相似的。通過對比不同旅館間營收的變化趨勢，我們還發現，這種關係主要體現在Resort hotel上，而對於City hotel，它全年的營收都是相對平緩的。



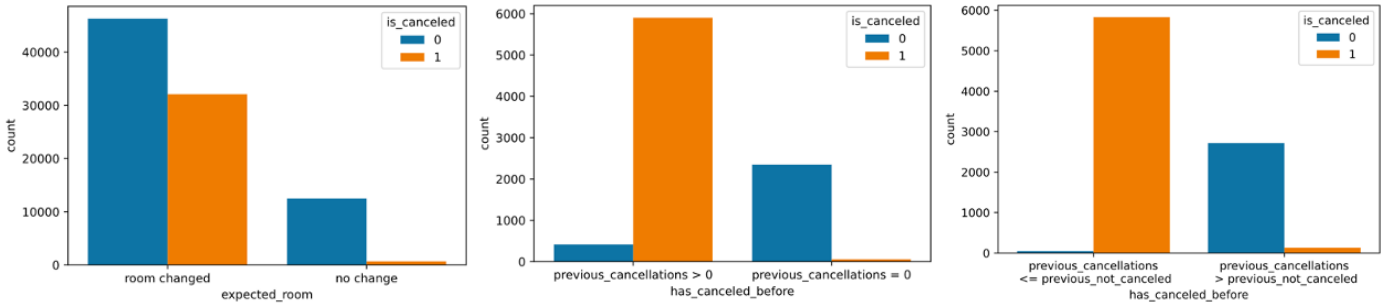


2. 訂單是否更換房型和是否取消之間的關係

我們新增了一個特徵，用於標記當前訂單的預定房型（reserved_room_type）和入住房型（assigned_room_type）是否一致，通過觀察該特徵與訂單是否取消之間的關係發現，被取消的訂單中出現預訂房型和入住房型不一致情況的比例遠少於未被取消的訂單，因此我們認為此特徵會是一個判斷訂單是否被取消的重要依據。

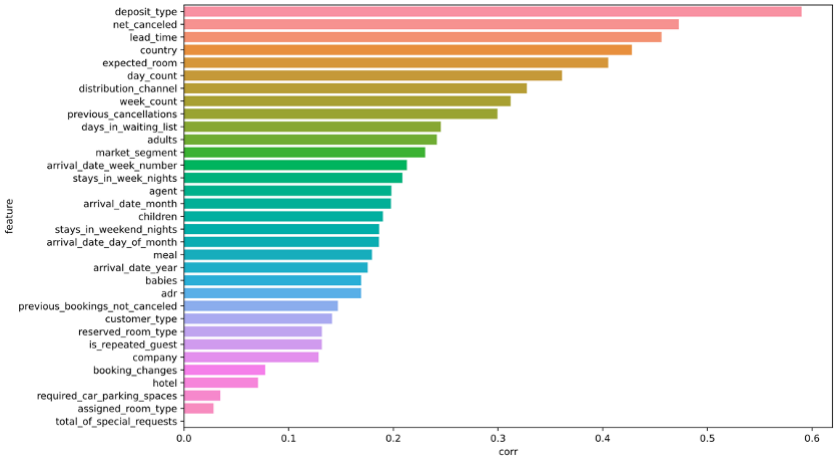
3. 顧客取消訂單的頻率

訂單是否會被取消跟顧客以往的消費習慣有很大關係，那些在之前有過取消訂單行為的客戶有較高的機率也會取消本次訂單，特別的是，我們發現之前取消訂單次數大於未取消訂單數目的顧客，幾乎一定會取消本次訂單，因此我們將歷史訂單取消數是否大於未取消訂單數作為一個新的特徵值加入到我們的數據中。

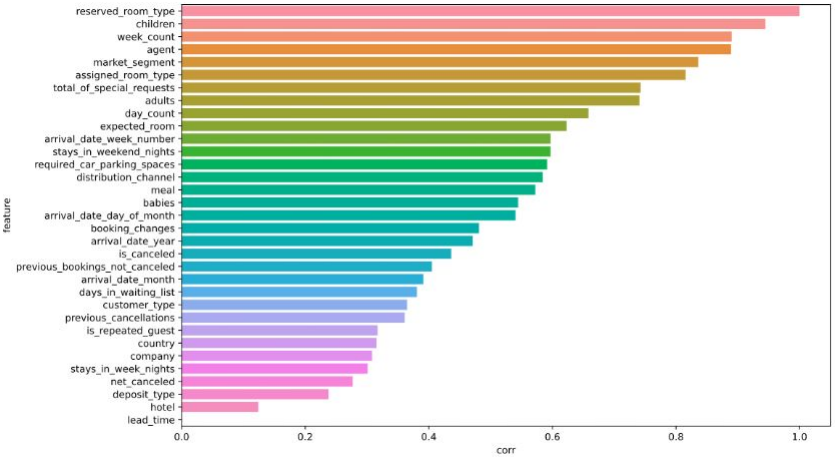


4. 特徵重要性分析

①各特徵與 is_canceled 的相關性如下圖所示，對其影響最大的特徵是 deposit_type，我們新加入的一些特徵，包括當天總訂單數目 day_count、是否更換房間 expected_room 和淨取消數 net_cancel與目標屬性 is_canceled 的相關性都非常高，而關聯度不高的特徵 total_of_special_requests 則可考慮在之後的訓練過程中移除。



②各特徵與 adr 的相關性如下，與 is_canceled 不同，與 adr 最相關的特徵為訂單的房型 reserved_room_type，而與 is_canceled 關聯較大的 lead_time 卻對 adr 影響輕微，因此在處理相關變量的增刪時，兩個目標屬性需分開處理。



三、方法

Hist-Gradient Boosting

Hist-Gradient Boosting 是 scikit-learn 近幾年新加入的功能，它是由微軟 LightBGM 發展而來的，因此可以透過 LightBGM 來大致了解 Hist-Gradient Boosting 的原理以及實作方式。LightBGM 基本上就是 Gradient-Based Decision Tree，但是在分類時，並不用將該特徵內數值排序再做分類，而是將其分成數個 bin，主要目的是要解決當訓練資料過多時 Gradient-Boosting 速度會變慢的問題。根據 LightBGM 的原論文[1]得知它有以下這些優勢:

1. **更快的訓練速度:**
   使用 bin 來取代連續的資料，加速訓練的速度。
2. **減少記憶體使用量:**
   使用 bin 來取代連續的資料，以減少記憶體使用量。
3. **比其它 boosting algorithm 有更好的 performance:**
   用 leaf-wise 而不是 level-wise 的方式，因此能產出更複雜的樹，但相對也更容易 overfitting，但可由限制 max_depth 來避免。

方法一、分開預測 is_canceled 跟 adr

用兩個不同的模型分別預測 is_canceled 跟 adr，並用預測結果計算 label，公式如下:

$$label = daily\ revenue \mid 10000$$

$$revenue = (1 - is\_canceled) \times adr \times (stays\_in\_weekend\_nights + stays\_in\_week\_nights)$$

我們嘗試了幾種常見的模型，結果如下：

- **Is_canceled**

模型（scikit-learn 套件，預設參數）	**Accuracy (30% validation data)**
Random Forest Classifier	**0.88**
Hist-Gradient Boosting Classifier	0.87
Linear Support Vector Machine	0.76
MLP Classifier (Neural Network)	0.87

由實驗結果發現 Random Forest Classifier 的表現最好，因此使用它來預測 is_canceled。在稍微調過參數後 accuracy 進步到 0.89，設定參數如下: max_depth=35, n_estimators=100。

- **adr**

模型（scikit-learn 套件，預設參數）	**R^2 score (30% validation data)**	**RMSE score (30% validation data)**
Random Forest Regressor	**0.616**	**154.9**
Hist-Gradient Boosting Regressor	0.605	157.2
Support Vector Machine for Regressor	0.312	215.1
MLP Regressor (Neural Network)	0.496	177.4
Linear Regression	0.288	210.9

一樣發現 Random Forest 表現最好，因此用它來預測 adr。參數設定跟 Random Forest Classifier 用一樣，R2 score 小進步到 0.62，RMSE 進步到 154.1。

Random Forest 在預測 is_canceled 跟 adr 都比其他方法來的好，因此最後用 Random Forest 分別預測的 is_canceled 跟 adr，丟到網站上面的 public score 為 0.60。

方法二、預測 revenue

直接預測單筆訂單的 revenue (飯店實際收到的收入)，revenue 的公式如下

$$revenue = (1 - is\_canceled) \times adr \times (stays\_in\_weekend\_nights + stays\_in\_week\_nights)$$

因為 Random Forest 跟 Hist-Gradient Boosting 不管在 adr 或 is_canceled 的預測效果都相對比其他的方法來的好，因此拿這兩個模型直接預測 revenue 並分析做比較。

以日期來切分成 training 跟 validation 資料集，因此同一天的訂單都會被歸類在同一邊 (training 或 validation)。用此方法會更符合 test data 的情形且可直接計算跟 true label 的 MAE 來得到更適合檢驗模型成效的分數。

結果如下:

模型 （scikit-learn，預設參數）	**Revenue's RMSE (30% validation)**	**Label's MAE (30% validation)**	**Public Score**
Random Forest Regressor	173.6	0.48	0.47
Hist-Gradient Boosting Regressor	**158.1**	**0.41**	**0.38**

我們使用同樣的資料處理方法、同樣的預測目標，但使用 Hist-Gradient Boosting 模型的分數卻比用 Random Forest 模型的分數要來的低很多 (0.47→0.38)，這跟之前分別預測 adr 跟 is_canceled 時得到的結果不一樣，使得我們很訝異。

方法三、預測 masked adr

預測被 is_canceled masked 過的 adr (若被 canceled 則 adr 為 0 其餘的正常)，公式如下:

$$revenue = masked\ adr \times (stays\_in\_weekend\_nights + stays\_in\_week\_nights)$$

$$masked\ adr = (1 - is\_canceled) \times adr$$

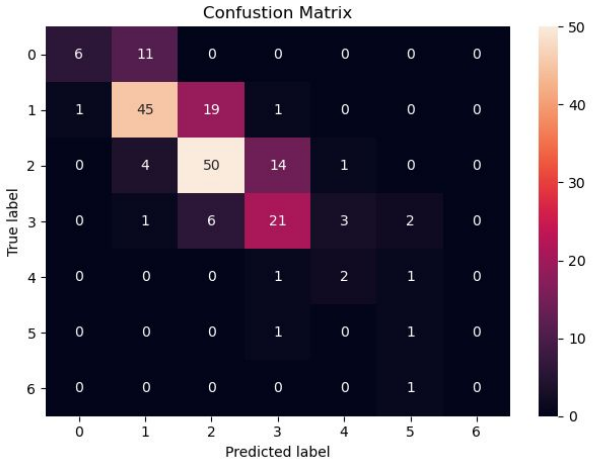
結果如下:

模型 (scikit-learn，預設參數)	**Masked adr's RMSE (30% validation)**	**Label's MAE (30% validation)**	**Public Score**
Hist-Gradient Boosting Regressor	**38.5**	**0.38**	**0.36**

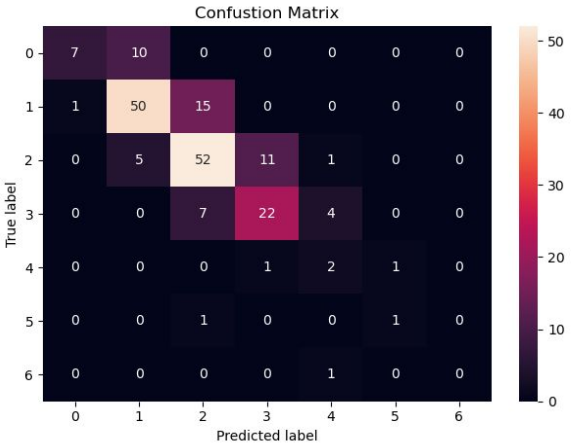
在 public score 上比直接預估 revenue 成效來的好 (0.38 → 0.36)

方法四、預測 masked adr + quantization

分析了前面的方法，發現預測的值普遍會比實際的值來的高，如圖一（方法三的confusion matrix）所示：



圖一: 預測 masked adr 的 confusion matrix



圖二: 預測 masked adr + quantization 的 confusion matrix

再次分析預測的 daily revenue 發現有許多位於邊界附近的值，這些值很可能因為微小的預測誤差卻被分類成錯誤的 label，我們認為把它們預測到另外一邊的 label 也許有機會可以降低模型的預估誤差。因此我們不再使用固定的 10000 作為每一級的閾值，而是設計了一個特殊的 quantization 來降低這種誤判，例如 daily revenue 預測為 19000，照原本的公式是 label = 1，但經過我們特殊設計的 quantization 會變成 label = 2。

結果如下:

模型（scikit-learn，預設參數）	**Masked adr's RMSE (30% validation)**	**Label's MAE (30% validation)**	**Public Score**
Hist-Gradient Boosting Regressor	**38.0**	**0.39**	**0.35**

confusion matrix 如圖二所示，高估的情形有稍微被改善，且 public score 也有稍微降低 (0.36→0.35)

方法五、用 data augmentation 後的資料預測 adr & is_canceled

Test Data 的日期是從 2017年 4月 1號到 2017年 8月 31號，為了加強模型預測 4月到 8月這段期間的成效，從過去的年份（2015、2016年）多複製一次同月份的資料，並使用預測 masked adr 的方法（方法三）。

比較了複製兩種不同日期區間的資料，結果如下:

複製的資料日期	**Masked adr's RMSE (30% validation)**	**Label's MAE (30% validation)**	**Public Score**
2015/6/1 - 2016/3/31	**37.6**	**0.35**	**0.32**
2016/4/1 - 2016/8/31（同 test data 區間）	43.8	0.57	0.42

結果出乎意料用 test data 同樣月份複製得到的結果變差，反而是複製幾乎沒有重疊的月份得到比較好的 performance。我們的猜測認為可能這段期間剛好跟 test data 的分布很像，成效才會比較好。

四、比較

方法	**Public Score**	**Private Score**
1. RF 預測 is_canceled & adr	0.60	0.53

2-1. RF 預測 revenue	0.47	0.49
2-2. HGBR 預測 revenue	0.38	0.48
3. HGBR 預測 masked adr	0.36	0.44
4. HGBR 預測 masked adr + quantization	0.35	**0.33**
5. Data Augmentation + HGBR 預測 is_canceled & adr	**0.32**	0.38

- **Random Forest（方法 2-1）vs Hist-Gradient Boosting（方法 2-2）：**

  方法2-1 跟方法2-2 兩個方法的差別在一個模型是用 Random Forest 另外一個是用 Hist-Gradient Boosting，使用 Hist-Gradient Boosting 方法的 public score 是 0.38 遠比 Random Forest 的 public score 0.47 還要來得小，且在 training data 的 validation 上也得到相似的結果 (0.48 vs 0.40)，讓我們原先以為 Hist-Gradient Boosting 比 Random Forest 在此資料集上的成效還要來的好，但從 private score 看到兩者的分數差異不大 (0.49 vs 0.48)，Hist-Gradient Boosting 或 Random Forest 以成效來說並沒有哪個比較好。

五、結論與推薦

在此次的Final Project，我們試過許多種 ML 的方法，有些成績較好，有些差強人意，我們綜合以上討論，推薦以下方法。

為了避免 adr 預測誤差與 is_canceled 預測誤差在相乘後導致 daily revenue 誤差加劇。直接將 is_canceled 為 1 的訂單的 adr 先設為 0，稱為 masked adr。使用 Random Forest 或 Hist-Gradient Boosting 預測 masked adr 並根據 validation 分析的結果做適當的後處理（方法四）， 兩種模型的優缺點分別如下：

1. **Random Forest：**

   <u>優點：</u>
   a. 由多個 Decision Tree 組合而成，可以避免單個 Decision Tree 容易 overfit 的缺點。
   b. 不用擔心資料是否為線性。
   c. 不需要擔心特徵之間的關聯性，事實上我們曾經嘗試過增加一些認為更相關的特徵，或刪除一些看似較不相關的特徵，其結果顯示影響並不顯著，應該是由於它的計算原理，較不相關的特徵本來就不容易成為分支的條件，保不保留沒有甚麼差別。

   <u>缺點：</u>
   a. 當資料很大時，訓練速度會變慢。

2. **Histogram-Based Gradient Boosting：**

   <u>優點：</u>
   a. 在學習時，會將之前的學習錯誤放大，以達到更好的學習效果，這是前個方法沒有的特色，事實上也有不錯的結果。
   b. 更快的訓練速度。

   <u>缺點：</u>
   a. 相較容易發生 overfitting。

六、參考資料

[1] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146-3154).