Machine Learning Techniques Final Project Report

組員：B07705025 黃珮雯 B07705047 林昱辰 B07705053 劉品君

一、動機：

對於任何產業，Revenue 可說皆是生計第一考量。而在飯店業中，每日住房率則是影響獲利的最大變數。因此，準確預測飯店的每日住房率、獲利便成為一項非常重要的議題。透過這次收集到的飯店資料[1] 各個欄位，我們試圖使用不同模型來預測出每日獲利（以標籤分類），並評估不同模型的特性、準確率等。

二、訓練方法分析：

定義資料中各個欄位的意義：

X: train.csv 中前處理過後剩下的所有欄位
（不包含 is_canceled / adr，包含 interval variable 與 categorical variable）

y_1: 中間變數 is_canceled（binary variable）

y_2: 中間變數 adr（interval variable）

y_{label}: 最終變數 label（ordinal variable）

預測方式：

$X \rightarrow y_1$:
以訓練資料 X 預測出每筆交易是否被取消 y_1

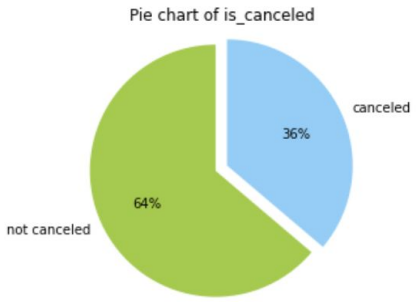
$X \rightarrow y_2$:
以訓練資料預測出每筆交易當日住房率 y_2

$R = (1 - y_1) \times y_2 \times stays$:
已成交交易 $1 - y_1$ 乘上當日住房率 y_2 及住房天數 $stays$
（stays_in_weekend_nights / stays_in_week_nights）得到獲利數量
（interval variable）

$R \rightarrow y_{label}$:
以獲利量 R 預測最終獲利類型 y_{label}

三、探索資料集：

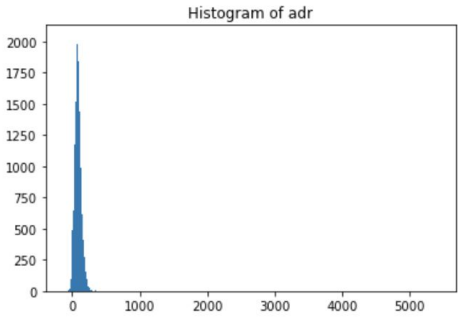
1. is_canceled 圓餅圖：



[1] Hotel Booking Demand, Jesse Mostipak,
https://www.kaggle.com/jessemostipak/hotel-booking-demand

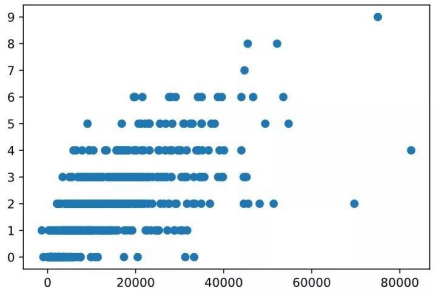
有超過六成的交易是有成交的，而剩餘 36% 為已取消交易。因此，此份資料集算是平衡的結構，較不會有不對稱的問題。

2. adr 直方圖：



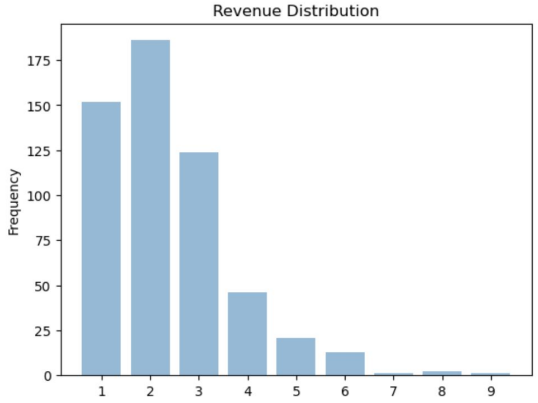
可以看到大部分的資料數值都非常集中分佈在 0~300 左右之間，然而他極端最大值 5399.424061287761 與最小值 -118.23836167072805 使得 histogram 分佈變形，偏向左端。

3. 當日獲利量 revenue 與最終預測目標 label 散佈圖：



縱軸為當日獲利標籤，而橫軸為當日計算出的 revenue。可以看出其實每個標籤分佈的標準差其實頗大，而且還有一些 outlier。這樣包含 noise 的資料可能會影響訓練的困難度。

4. label 長條圖：



從長條圖可以看出在九種標籤中，前 1~3 個佔了大多數，而後面幾個標籤的頻率則是偏小，整體呈現 right-skewed 的情況。

四、訓練資料前處理：

由上所述，為了預測 is_canceled、adr，訓練資料中必須去除掉 is_canceled、adr、reservation_status_date 以及 reservation_status 的欄位，再將此資料與 test data 資料合併一起進行前處理。

對於訓練資料，先對各個欄位做基本 NA 值分析：

Attribute	# of NAs	Attribute	# of NAs
ID	0	distribution_channel	0
Hotel	0	is_repeated_guest	0
lead_time	0	previous_cancellations	0
arrival_date_year	0	previous_bookings_not_canceled	0
arrival_date_month	0	reserved_room_type	0
arrival_date_week_number	0	assigned_room_type	0
arrival_date_day_of_month	0	booking_changes	0
stays_in_weekend_nights	0	deposit_type	0
stays_in_week_nights	0	agent	16340
adults	0	company	112593
children	4	days_in_waiting_list	0
babies	0	customer_type	0
meal	0	required_car_parking_spaces	0
country	488	total_of_special_requests	0
market_segment	0	-	-

透過以上表格，發現 children、country、agent 以及 company 皆有 NA 值，其中 children 欄位只有 4 個空值，因此我們直接將整份資料取中位數填入空格。然而，另外三個欄位皆有較大量的 NA，而且又屬於 nominal data，決定直接刪除欄位。

剩下的欄位中，在此假設當天的星期幾（如：週一、週二等）可能也會影響 adr 和 is_canceled 的因素。因此，利用 arrival_date_year、arrival_date_month、arrival_date_day_of_month 計算並加入這項因子。其餘許多 nominal data 欄位皆被轉為數字型態，一個數字分別代表原先一個類別。

最後，由於考量到最後要預測的資料時間為 2017/04/01~2017/08/31，時間比整份訓練資料晚，也就是說訓練資料與測試資料可能是不同的分佈型態。因此，我們選擇以 8:2 的比例切出前面十六個月（2015/01/01~2016/11/30）為實際做訓練的資料，後面四個月（2016/12/1~2017/3/30）為 validation 資料，期望能因此使 E_{val} 貼近 E_{out}，以利模型選擇。然而，這樣做有一個淺在影響是 validation 資料中沒有四到十二月的部分，在選擇模型時可能會因為沒有考慮到這些月份而造成一些 bias。不過後續使用 feature selection 篩選時，發現 arrival_date_month 欄位的顯著性並不高，因此即合理推測此 bias 影響不大，仍然使用此方式以模擬出時間對於資料的效果。

五、模型選擇：

資料集特性：

- 欄位數量多（原始資料 29 欄位，處理過後 23 欄位）
- 許多非連續性變數（categorical variable）
- 資料筆數多（74,843 筆訓練資料）

- 資料可能包含雜訊

各個模型特性分析：

以下將針對預測 is_canceled、adr 和對 revenue 預測 label 的三部分，我們將針對不同問題提出適合的模型並進行探討並分析優缺點。

1. 預測 is_canceled（二元分類）：
    a. kNN Classification：
        - 優點：
          模型可解釋性高，不僅是數學上或是直觀思考上都有很好的解釋，且非線性模型通常會有較低的 bias。
        - 缺點：
          運行時間較長，因 testing time 與訓練資料大小有直接線性關係；此外，kNN 預測較容易 overfit noise。
    b. Random Forest Classification：
        - 優點：
          融合 bagging 與 decision tree 的方式，可以有效降低 overfitting。此外，也能有效處理高維度、多欄位的資料集；在分類別時，可以動態判斷各特徵的重要性，並保持較低的時間複雜度 $O(mn\ log\ n)$，而 m 為特徵數量、n 為訓練資料數量。
        - 缺點：
          決策樹類型的模型通常較欠缺理論上的支援。
    c. Logistic Regression Classification：
        - 優點：
          此模型的函數適合用於二元分類，且其解釋性也非常直觀，也就是該筆向量 X 屬於某個類型的機率。
        - 缺點：
          迴歸模型通常較適合連續性變數，且模型結構較簡單所以預測上容易有較高 bias。
2. 預測 adr（迴歸連續變數）：
    a. Random Forest Regression：
        - 優點：
          同前所述，此模型可以減少 overfitting 問題，善於處理高維度資料並能給予不同特徵不同權重確保相對重要性。
        - 缺點：
          即使準確性高，模型解釋性通常較低。
3. 預測 label（多元分類）：
    a. Decision Tree：
        - 優點：
          模型結構類別性資料、且和 Random Forest 一樣有較為直觀的解釋性。此外，時間複雜度（同 Random Forest）較低，因此不容易 overfit noise。
        - 缺點：
          對於線性模型，預測上可能會有 bias。

特徵篩選：

我們曾考慮過是否要使用 feature selection 來改善模型的準確率，因而對於預測 is_canceled、預測 adr 個別有進行特徵篩選。

針對 is_canceled 篩選時，我們使用 χ^2 來提取分數高的前 15 個特徵，其中選出的特徵有：'ID', 'hotel', 'lead_time', 'arrival_date_week_number', 'distribution_channel', 'is_repeated_guest', 'previous_cancellations', 'previous_bookings_not_canceled', 'reserved_room_type', 'assigned_room_type', 'booking_changes', 'deposit_type', 'days_in_waiting_list', 'required_car_parking_spaces', 'total_of_special_requests'。

針對 adr 進行篩選時，使用 ANOVA F-value 選取前 15 個特徵，選出了 'ID', 'hotel', 'lead_time', 'stays_in_weekend_nights', 'stays_in_week_nights', 'adults', 'children', 'babies', 'meal', 'previous_cancellations', 'previous_bookings_not_canceled', 'booking_changes', 'deposit_type', 'days_in_waiting_list', 'customer_type'。

驗證結果：

1. 預測 is_canceled：

is_canceled	kNN (k = 3)	Random Forest (random_state = 0, oob = True, min_sample_leaf = 1)	Logistic Regression (random_state=0, penalty='l2', solver='newton-cg', multi_class='multinomial')
No feature selection	0.33215484180249	0.23382070949185	0.23400047938638
With feature selection	0.33215484180249	0.23382070949185	0.25239693192713

在使用 Random Forest 時，我們使用的參數如表格內說明，而 random_state 設為 0 是方便固定結果，方便模型之間的比較。表格呈現出三種模型對於預測 is_canceled 的表現 E_{val}（以 0/1 error 計算）。首先，特徵篩選對於結果的誤差不大，但是在 Logistic Regression 有較明顯的錯誤率上升。可以看出 kNN 的錯誤率相較其他兩種模型略高，而 Random Forest 與 Logistic Regression 都大約在 0.234 左右。在考量兩個模型的優缺點後，我們認為 Random Forest 有效動態判斷各特徵重要性這項特點，在這份欄位繁多的資料集會相較於 Logistic Regression 更有可能有較高的適應力，於是我們選擇使用 Random Forest 做後續的訓練。

2. 預測 adr：

adr	Random Forest (random_state = 0, oob = True, min_sample_leaf = 1)
No feature selection	0.6056049624454187
With feature selection	0.500844494392595

延續前面的結果，我們在預測連續變數 adr 時使用 Random Forest Regression 計算。在這部分使用的錯誤衡量標準是 R^2 來評估模型合理性，而 0.6056049624454187 表示大約 60.56% 的誤差可以被模型解釋，我們認為是足夠好的指標。相較之下，使用特徵篩選後的模型的合理性大幅下降了 10%，顯示減少特徵可能是一種負面影響。

3. 預測 label：

label	Decision Tree
E_{val} (0/1 error)	0.008264462809917

計算出前面兩項特徵後，以 $R = (1 - y_1) \times y_2 \times stays$ 計算出當日的 revenue。使用 Decision Tree 訓練出結果後，結果會有 0.008 左右的 validation error。

最終模型選擇：

使用這些特徵分別對以上提到的模型去做訓練時，最後計算出來各個模型的結果如上所示。然而，在 E_{val} 的表現上，無特徵篩選的版本大致上比篩選過的資料有較高的準確率，因此有可能表示此預測問題的目標函數有很多變量共同影響，此外，因全部特徵皆使用時，電腦計算效能也可以負荷，於是我們故最終便不篩選特徵。

綜合以上各個錯誤率的評估後，我們的最終模型選擇為同樣利用 Random Forest 設定參數為 (random_state = 0, oob = True, min_sample_leaf = 1) 預測 is_canceled 和 adr，並用 Decision Stump 設定參數為 (max_depth = 10, min_sample_split = 2) 分類 revenue 和其相應的 label。此模型的好處是他可以對於 is_canceled 和 adr 分別針對不同特徵給予不同權重、且不會有太大的計算量負荷，有效減少運算時間同時確保高準確率。

六、驗證結果：

我們選擇使用的模型最後以測試資料算出的錯誤率（Mean absolute error）為 0.55842，不過單看一個數據較難比較出模型的有效性。因此另外也使用了只以「月分、日期、星期」來預測的 naïve prediction 來作為模型的分數 baseline，而其得到的測試分數為 0.948052。相較之下，我們目前使用 Random Forest + Decision Tree 的結果，無論是 E_{val} l 或是 public socre，比起 baseline都好上許多，可見目前模型對於原始資料是有展現出學習效果的。

七、組員分工：

黃珮雯：kNN 模型訓練、報告撰寫
林昱辰：資料處理、 Random Forest/Logistic Regression 模型訓練、報告討論
劉品君：模型參數調整、報告討論

八、參考資料：

Hotel Booking Demand, Jesse Mostipak, Kaggle,
https://www.kaggle.com/jessemostipak/hotel-booking-demand
sklearn.neighbors.KNeighborsClassifier, scikit-learn 0.24.0 documentation,
https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
sklearn.emsemble.RandomForestClassifier, scikit-learn 0.24.0 documentation,
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
sklearn.linear_model,LogisticRegression, scikit-learn 0.24.0 documentation,
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
sklearn.emsemble.RandomForestRegressor, scikit-learn 0.24.0 documentation,
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
隨機森林（Random Forest, RF）的生成方法以及優缺點，IT 閱讀,
https://www.itread01.com/content/1547100921.html
Random Forest 算法參數解釋及調優, CSDN 博客,
https://blog.csdn.net/qq_16633405/article/details/61200502
Why is the runtime to construct a decision tree mnlog(n)?, Stack Overflow,
https://stackoverflow.com/questions/34212610/why-is-the-runtime-to-construct-a-decision-tree-mnlogn