Group : 未命名

成員 : 吳承哲、何明昕、溫皓良

Data Preprocessing

Dataset 為來自Kaggle hotel booking demand data，分為train以及test。Train之中有91531筆資料，有33種特徵，而test中有27859筆資料，有29種特徵。

Package : numpy、pandas。

1. 刪去沒有用或不能使用的features

ID : ID為連續變大的數字，表示資料的順序，與is canceled及adr無關。

Reservation status、reservation status date :在 test中沒有這兩個特徵。

1. 刪除異常值
2. 將stay in week nights、stay in weekend nights同時為0的資料刪除。
3. 將adults、children、babies同時為0的資料刪除。
4. 將adr <= 0的資料刪除。
5. 資料型態轉換

Agent及company的資料型態為integer，但裡面的數字為ID，因此要轉換為object，才能夠做one hot encoding。

1. 刪除極端值

Adr : 去除5000以上的資料，有1筆。

Babied : 去除2以上的資料，有2筆。

Distributed channel : 去除Undefined的資料，有1筆。

1. One hot encoding

由於train跟test在one hot encoding後，特徵數不一樣，因此先將train及test合併後，做one hot encoding，再分開回train跟test。

1. 填補nan

將資料中nan的值，全部用0取代。

1. 提取label

預測cancel : 將train中is canceled取出來並刪除is canceled及adr。

預測adr : 將train中adr取出來並刪除is canceled及adr。

最後train跟test的維度分別為 (88974, 931), (27859, 931)。

Training & Prediction

Package : sklearn、time。

1. 使用sklearn的train test split將train分為train及validation，比例為8:2、random state=0。
2. 使用不同模型來訓練，使用預設參數(無調參)，求出train跟validation的準確度，評估模型的優劣，預測is canceled使用accuracy，預測adr使用mean absolute error。
3. 將train跟validation合併為原始完整的資料，然後訓練模型來預測最後答案。

Task 1: predict “is canceled”

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Random Forest Classifier | Nearest Neighbor Classifier |  |
| Train accuracy (%) | 99.46 | 86.51 |  |
| Val accuracy (%) | 90.63 | 80.08 |  |
| Time | 54s | 18s |  |

Task2 : predict “adr”

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Random Forest Regressor | Nearest Neighbor Regressor |  |
| Train MAE | 5.44 | 15.79 |  |
| Val MAE | 11.15 | 19.79 |  |
| Time | 4m54s | 29s |  |

Task3: calculate revenue

以日為單位，先去除掉cancel的訂單，再將當天每組訂單的stay in week nights加上stay in weekend nights乘上adr後做加總，最後除10000就為revenue，寫成以下公式 :

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Random Forest Classifier + Random Forest Regressor | Nearest Neighbor Classifier + Nearest Neighbor Regressor |  |
| Public score | 0.539474 | 0.671053 |  |
| Private score | 0.415584 | 0.649351 |  |

Best model : Random Forest