T\_Brain Travel Document

**Design Process:**

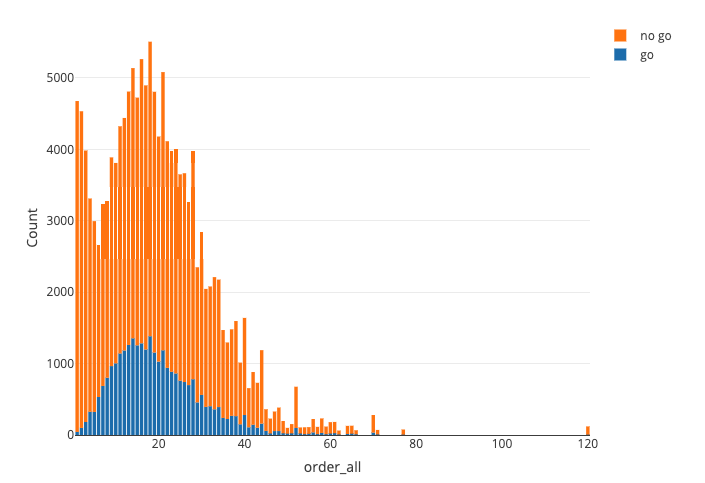
1. **Clean data**
   1. Transform the column to the certain type I want.
   2. Clean the strange data such as duplicate src\_airport, and etc.
2. **EDA**
   1. Observe the distribution of the data.
   2. Check if existing missing-value.
   3. Find relationship between column and column.
3. **Sampling(Class imbalance)**
   1. Oversample
   2. Undersample
   3. Smote
   4. TomekLinks
4. **Feature engineering**
   1. **One-hot-encoding** for categorical features.
   2. **Label encodeing** for ordinal features.
   3. Extract day, month, weekend, and etc from date and map them.
   4. Map airport into state or country due to Area feature covered.
   5. Intersection feature such as count of transform including go or back.
   6. Set hodilday dictionary to chech if they travel on holiday.
   7. Continuous feature without handling because model will automatically find the best binning point.
5. **Apply different models to predict**
   1. Random Forest
   2. Xgboost
   3. Stacking

**Improvement:**

1. Complete transformation of type when I import csv file.
2. Try more methods to solve class imbalance including tuning hyperparameter called **scale\_pos\_weight.**
3. Try **combination feature** to produce more stastistical feature.
4. Apply more models to stack.

**Discussion:**

Actually, there are a feature called ‘amount of group I didn’t adopt. Instead, all of the top 5 competitors adopt it. I think it is **data leakage**. We won’t know the actual amount before register of the group ends.



**Conclusion**

This is my first competition. No matter cleaning data, eda, feature engineering, and modeling, I complete by myself. Although I just rank 13%, I learn many skills including feature engineering, tuning hyper-parameter, and etc. through Top 5 github. To sum up, it is a great experience and thanks for who held this competition.

|  |  |  |  |
| --- | --- | --- | --- |
| 訂單(order.csv, 主資料表) | | 團體(group.csv) | |
| 欄位名稱 | **欄位說明** | **欄位名稱** | **欄位說明** |
| group\_id | 團號 | group\_id | 團號 |
| order\_id | 訂單編號 | sub\_line | 副線別(碼掉) |
| order\_date | 下單日期 | area | 地區(碼掉) |
| source\_1 | 訂單來源1(碼掉) | days | 天數 |
| source\_2 | 訂單來源2(碼掉) | begin\_date | 出發日 |
| unit | 負責單位(碼掉) | price | 價格 |
| deal\_or\_not | 是否成行 | product\_name | 網上銷售團名 |
| people\_amount | 訂單人數 | promotion\_prog | 行程特色 |
| 航班(airline.csv) | | **日行程(day\_schedule.csv)** | |
| 欄位名稱 | **欄位說明** | **欄位名稱** | **欄位說明** |
| group\_id | 團號 | group\_id | 團號 |
| go\_back | 去/回程 | day | 第幾天 |
| fly\_time | 起飛時間 | title | 日行程主題 |
| src\_airport | 起飛機場 |  |  |
| arrive\_time | 抵達時間 |  |  |
| dst\_airport | 抵達機場 |  |  |