馬拉松運動博覽會參訪動線類別預測

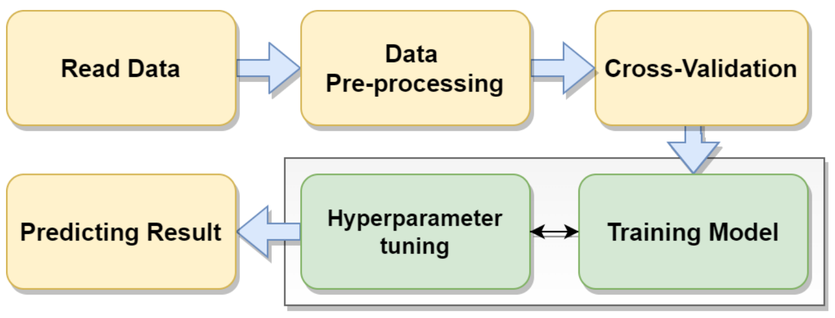
Machine Learning Final Project

Team member: 常安彥、楊晶宇、李秉澤、張書瑜

Introduction

根據該議題資料收集來自2018馬拉松博覽會展場的部份抽樣數據，必須根據人潮動線的收集與分析，瞭解出參訪群眾的喜好與停留情況，訓練與建立出一個良好的決策模型，來整理規劃出五種參訪群眾動線類型。

Framework



1. Pre-Processing

(1) Features extraction

(a) 每位參與者各個攤位總逗留時長

(b) 攤位類別(以官方地圖分類)計次

(c) 攤位參觀順序(one-hot encode)

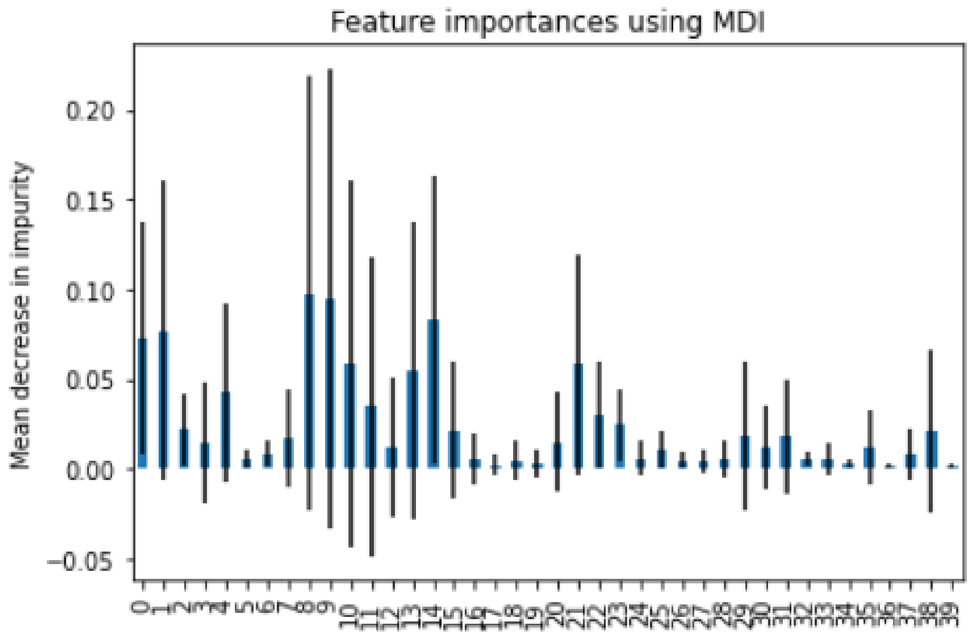
(d) 當天逗留總時長(共三天)

(e) 參與日期

(2) Outlier Cleaning

因為資料裡的time stamp有些資料明顯不正確，如：攤位與攤位之間的時間差過小(小於30秒)，或是攤位與攤位間的時間差過長(大於30分鐘)，我們將過長的補平均，過短的補至30。

(3) Feature Importance

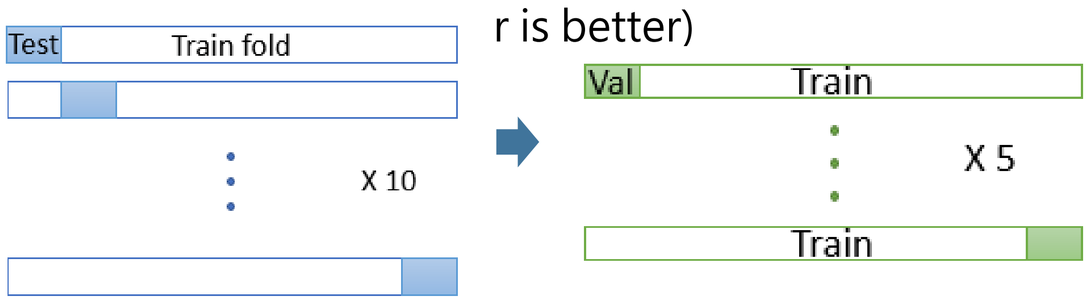


把一些不重要的features 刪掉，避免model overfitting

3. Nested Cross Validation 10 * 5

(1) outer cv 1/10, inner cv 1/5

(2) Grid Search



r is better)

4. Model

(1) LSTM

(2) LightGBM

hyperparameters tuning by cross validation

```
Pipeline(steps=[('sc', StandardScaler()),
   ['clf',
    LGBMClassifier(fraction=0.41, max_depth=4, n_estimators=150)]]))
```

(3) Xgboost

hyperparameters tuning by cross validation

```
Pipeline(steps=[('sc', StandardScaler()),
   ['clf',
    XGBClassifier(colsample_bytree=0.4, learning_rate=0.175,
                  max_depth=5, n_estimators=275,
                  objective='multi:softprob')]]))
```

Results

LSTM	submit_Soft_whol...csv stanleyanyan	2022-05-29 23:13:55	0.7545643
LightGBM	submit_06101330_...csv stanleyanyan	2022-06-10 14:44:50	0.0624151
XGboost	submit_06100304_...csv stanleyanyan	2022-06-11 21:08:36	0.0600909

Summary

這個比賽以log loss作為評斷標準，所以準確率即使很高，loss也不一定能達到好的結果，可能因為少數過於肯定的錯誤答案，影響整體表現。這個現象在我們的RNN神經網路中就相當明顯，準確率即使超過98%，模型對於大部分的決策都過於自信。所以這次我們比賽的重點都集中在如何做features extraction 和 cleaning outlier