

# Final project report - Group 10

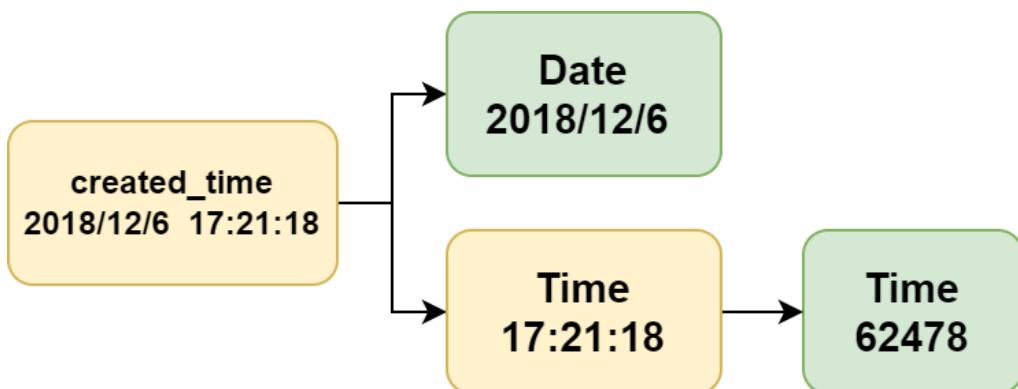
成員：李秉澤、常安彥、楊晶宇、張書瑜  
馬拉松運動博覽會參訪動線類別預測

## 一、介紹 (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。

## 二、資料準備 (Data Pre-processing)

這次的資料我們首先做的處理是將日期和時間做分割，並且把原本24時制的時間轉換成秒，這樣的目的是方便我們後面將 raw data 轉換成我們有興趣分析的 features，譬如這次的 model 我們最終使用的 features 有每位參加者每個攤位的逗留時間、各類別攤位的總逗留時間、訪問攤位的順序(前14個，不足補0)、每日參觀總時長、有參觀的日期以及前面 features 統計到訪攤位逗留時間中屬於outlier(大於1800秒和小於30秒)分別的資料總數。以上選中的幾個 features 是我們透過 feature importance 分析後，決定留下的幾個對模型影響較大的特徵。



### 三、特徵重要性評估 (Feature Importance)

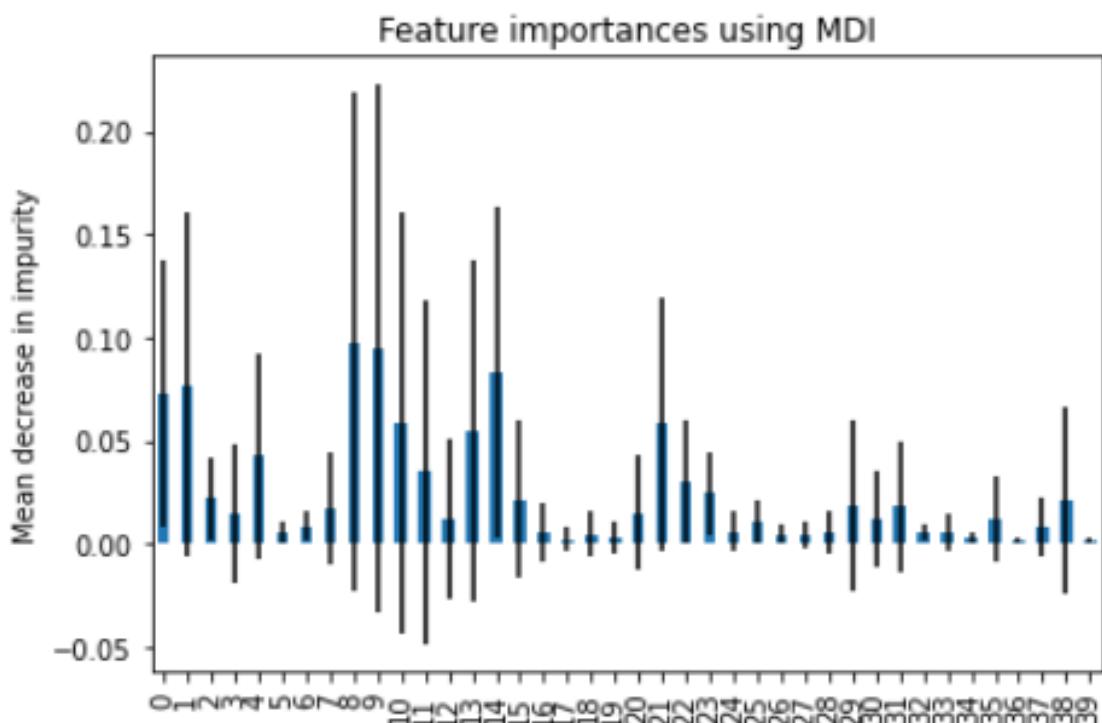
特徵重要性評分是一種為輸入特徵評分的手段，其依據是輸入特徵在預測目標變量過程中的有用程度，並且根據許多評估類型和來源，比如說統計相關性得分，線性模型的部分係數，基於決策樹的特徵重要性和經過隨機排序得到重要性得分等等，將該特徵給予一個特定值。

由於特徵重要性在預測建模項目中起著重要作用，包括提供對數據、模型的見解，以及如何降維和選擇特徵，從而提高預測模型的效率和有效性，因而我們將此步驟放入我們的模型建構中。

#### 1. Random Forest with MDI

對於難以用人腦分析的高維度資料 (各攤位逗留時間、各攤位類別逗留時間、訪問攤位順序……)，我們運用 mean decrease in impurity 來分析各個特徵在降低 classifiers 不確定性時的影響力，這裡的 ensemble decision tree method 是代入random forest。

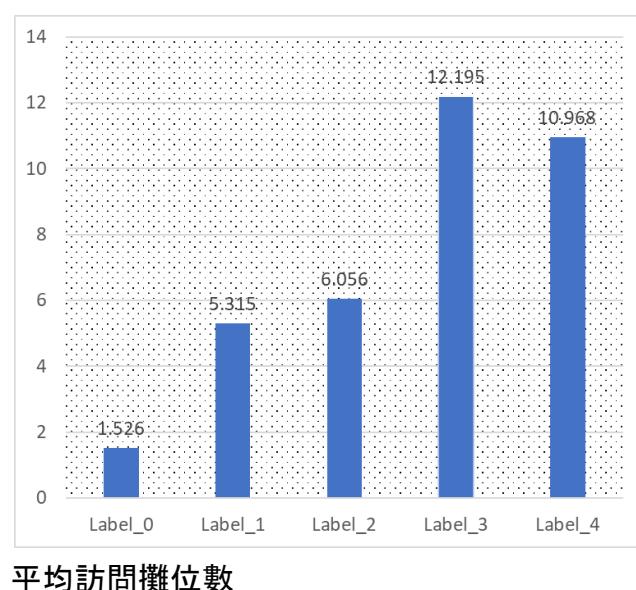
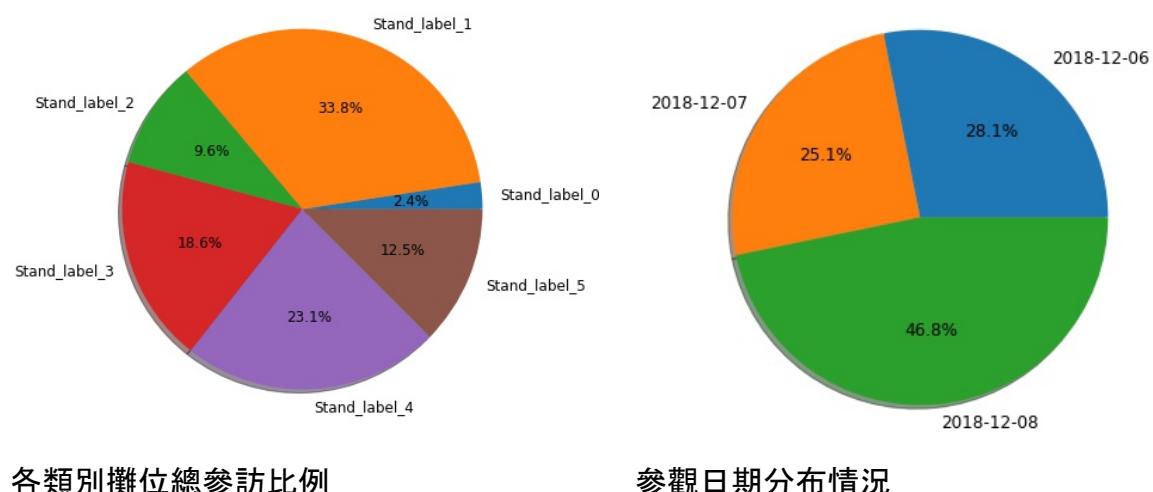
**Random Forest with MDI 評估結果**



## 2. Label Observing

由於比賽並沒有公布各個label所代表的意義，所以我們要自己透過觀察資料分布去自行判斷。接下來，我們觀察人腦能觀察到的趨勢，例如去的各類別參與者參觀的攤位總數平均、去各個類別的比例。

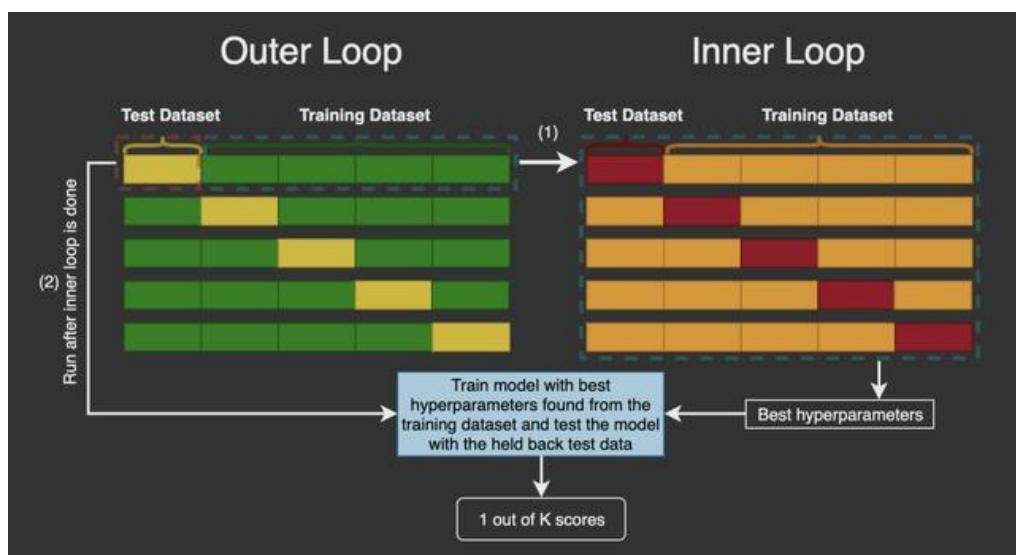
透過下圖結果，可以看到類別0的參與者去的攤位總數非常少、類別1及2去的略多一些、類別3及4就經過非常多的攤位。那對於類別1和2之間的差別就來自於經過的攤位編號，3和4亦然。



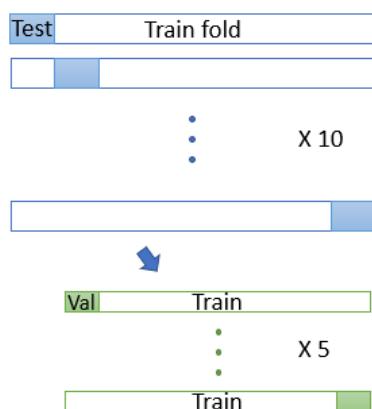
#### 四、嵌套交叉驗證(Nested cross-validation)

嵌套交叉驗證是通過對基礎模型泛化誤差的估計來進行超參數的搜索，透過不斷的尋找以得到模型最佳參數，而其內置的交叉驗證是傳統交叉驗證法（例如K-Fold交叉驗證）的延伸，由於傳統的交叉驗證僅是將數據集拆分為訓練集和測試集，無法解決最優模型的選擇及模型調參問題，因而我們使用到嵌套交叉驗證。

其內置交叉驗證的原理如下圖，其運行流程包含兩個循環-外層循環和內層循環。內層循環是指帶有搜索模型最佳超參數功能的交叉驗證，目的是給外層循環提供模型的最佳超參數。例如，隨機搜索或者網格搜索。而外層循環是給內層循環提供訓練數據，同時保留部分數據，以作對內層循環模型的測試。通過這樣的方式，可以防止數據的信息洩漏，以得到相對較低的模型評分偏差。嵌套交叉驗證設計：



上圖是 nested cross validation 的示意圖，我們的應用在outer loop分割十次的K-Fold，在inner loop做分割五次的K-Fold，如右圖所示。意即，每次取十分之一作為testing data，剩下十分之九中取五分之一做為validation data，而剩餘即為該次iteration的training set。



## 五、模型設計 ( Model Design)

### (a) LSTM

LSTM是我們第一個測試的model, 主要是把每一個參加者每一天經過的攤位順序做成14維的陣列(Zero padding), 共三天, 所以每筆資料(一位參加者一筆)會是一個42維的陣列, 送進LSTM得到最後一層的hidden state後丟進2層fully connected 的NN做training, 然而結果準確率雖然可達95%以上, 但Loss降得不夠低, 所以後來就沒有繼續往LSTM做嘗試。

### (b) XGboost

XGboost 是很有名的ML演算法, 我們將data extract出我們認為有用的Feature, 透過 nested cross validation去挑選XGboost可以tune的參數(n\_estimators、max\_depth、lr、colsample\_bytree)

### (c) LightGBM

LightGBM也是知名的ML演算法, 我們如(b)一樣取出data feature, 並透過nested cross validation去挑選LightGBM可以tune 的參數(n\_estimators、max\_depth、lr、fraction)

## 六、結果分析 ( Result Analysis )

### (a) LSTM

LSTM的結果並不理想，雖然在我們自己切的validation set有95%的準確率，但是在自己切的validation set loss也表現得跟private data差不多(0.77)，我們認為是LSTM的模型太容易overfit on train data，所以決定暫時往ML的方式做嘗試。

---

submit_Soft_whol...csv	2022-05-29 23:13:55	0.7545643
stanleyanyan		

### (b) Xgboost

Xgboost 在我們透過importance analyzed 挑選features和nested cross validation挑選參數之後，上傳了幾次結果，最好的結果落在0.060

submit_06100304_...csv	2022-06-11 21:08:36	0.0600909
stanleyanyan		

### (c) LightGBM

LightGBM 在我們透過importance analyzed 挑選features和nested cross validation挑選參數之後，上傳了幾次結果，最好的結果落在0.062

submit_06101330_...csv	2022-06-10 14:44:50	0.0624151
stanleyanyan		

---

通過多次的結果分析，我們認為最後進步loss的方式就是阻止model overfitting 以及仔細的處理data，如果test data中任一筆資料和整體資料的bias太大都會造成loss的上升，因此盡量去掉會overfit 的 features 以及處理outlier變成本次競賽的重點，如同在final 報告時，許多組別沒有取很多複雜的feature，反而獲得更好的效果，更是應證了我們的觀點，過多的feature反而會造成overfitting 導致loss的上升，所以feature要盡量簡單、在空間中對於類別有良好的分辨性。

## 七、總結(Summary)

一開始著手時，由於此次比賽給予的特徵值較少，只給了參觀者的編號、參觀紀錄點與紀錄時間而已。由於學習模型若給予的特徵值過少，會導致學期成效較差。因此我們便將重點著重於特徵值的建構，並且透過多次討論，評估我們選擇的特徵是否會出現問題，同時透過特徵重要性評估去證實我們的想法，這大大地幫助了我們在後續建構模型的部分。

到比賽中期，我們發現由於這個比賽以 log loss 作為評斷標準，所以即使我們建構的模型準確率很高，loss 也不一定能達到好的結果，我們便猜想這可能是因為少數過於肯定的錯誤答案，影響整體表現，而這個現象在我們的RNN神經網路中就相當明顯，期模型準確率雖然超過98%，但模型對於大部分的決策都過於自信，因而在錯誤的部分會得到較大的loss。

另外，這次的比賽我們認為必較可惜的部分在於我們設計模型和抓取特徵前沒有將各類別參觀群眾有參觀的攤位先做統計。因為在報告當天我們發現許多別組直接將有參觀的攤位編號透過NN模型分析即可獲得相當好的表現結果，可見這次的資料中參訪過哪些攤位與其所對應的類別群

是有高度正相關的。其他譬如與時間日期相關的特徵整體而言對模型的準確度提升則較小。