Behavior Classification of Exposition Visitors

Machine Learning Final Project
Team member: 周伯宇, 呂宸漢, and 林子鵑

Introduction

這次的主題是探討一場馬拉松博覽會群眾行為和人潮動線的關係與預測，從人潮的動線收集與分析，駐留情況，民眾對產品的喜好等等特徵項來做人群分類之分析。

本議題在 Aidea 上提供人潮 dataset，由參與者對參訪動線類別判斷機率值，再由 logloss 計算實際正確類別的預測機率並做對數計算加總取得最後平均。

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij}\log(p_{ij})$$

Framework

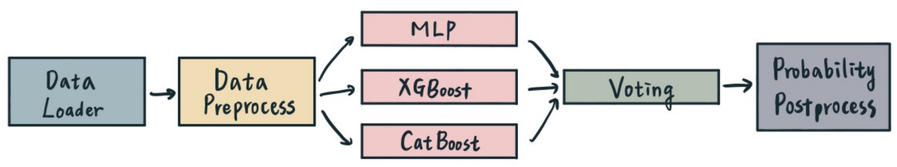


Fig. 1 The proposed method

Description

題目總共給了4個csv檔案以及一張展場配置圖。train.csv 包含了含有3種特徵共41640筆的樣本資料；training_label.csv 包含了訓練集的資料標記；test.csv 則含有無標記的20419筆樣本資料；最後是 submit_samples.csv，為上傳範例格式檔。

根據展場配置圖及 training data 資訊，我們可以觀察到每一個樣本根據他屬於不同類別有不同的行為，因此我們透過統計數據得到 sniffer_loc 的數量與label有一定的關係，且經過同色塊區域的 sniffer_loc 屬於某些類別也有關係。

Feature統計數據&Preprocessing

1. 駐留攤位數量統計：

Label	C0	C1	C2	C3	C4
Station Cnt	1.52583	5.31506	6.05625	12.29480	10.96780

2. 類別通過攤位種類（loc顏色）統計：

	P1(1)	P2(2, 4, 5, 6)	P3(3)	P4(8)	P5(7, 13)	P6(9, 10, 11)	P7(12, 14)
C0	1.93	31.97	17.24	4.23	10.62	12.89	21.11
C1	3.06	51.2	17.11	5.05	4.76	11.68	7.13
C2	0.19	20.17	3.49	9.27	12.57	39.46	14.86
C3	7.75	31.12	7.69	7.76	13.24	22.79	9.64
C4	0	33.58	8.08	8.13	12.05	25.01	13.16

因此我們將 data set 做了一些 preprocessing，將 feature 數量擴展成 22 項，1-14項為是否通過 sniffer_loc 編號，第15項為通過攤位數量，16-19項分別為通過 p2,p5,p6,p7 group 的數量，20-22項為哪幾天進行參訪。

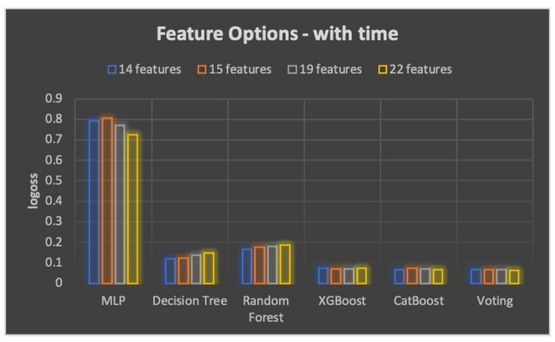
而在 feature 的調整上我們忽略了駐留時間資訊，因為我們觀察到有些人在一個 loc 停留時間不到一秒，且加入 timing 資訊後對結果並沒有正向的影響，關於 feature 組合與實驗結果會在 Results 部分展示。

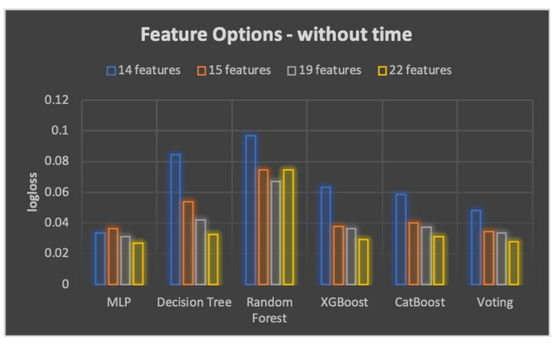
我們所嘗試使用的 model 主要有五種：MLP, Decision Tree, Random Forest, Extreme Gradient Boosting (XGBoost) 以及 CatBoost。而透過觀察五種 model 的結果，loss 較低的總共有三種分別為 MLP, XGBoost以及CatBoost，最後會依據這三種 model predict 的結果進行投票並決定最終類別機率。

Results

使用不同 feature 的組合來進行實驗，四種組合分別如下：1-14項、1-15項、1-19項以及最終使用的 feature1-22項。下圖是不同組合下在五種 model 以及 voting 結果的 logloss，最終 accuracy 達到 99.2%。

以下左圖為考量時間的 feature：1-14項代表在每個 sniffer 的駐留時間，15項為時間加總，16-19項也分別以類別時間加總計算，20-22項則為參訪日期；右圖為不考量時間的 feature（本次使用 feature）。





可以從上圖中觀察到，加入時間的因素考量後，logloss 上升很多，使得 prediction 的結果不穩定，accuracy 表現也較差。因此我們最後選擇不使用時間資訊來調整 feature。

結果顯示，我們將每個 model 使用不考量時間的 feature 來預測，得到最好的 logloss 為0.0407222。

Summary

我們在展場人潮預測中使用了有效率的 feature 調整方式，使用多種 model 做預測，最後在三種較好的 model 之間使用 voting 機制得到最後結果上傳 Aidea 系統，logloss 達到0.0407222。