

Kaggle Report

104502518 資工4A 劉冠聲

Team: Just Monika

結果截圖

Public Leaderboard

Private Leaderboard

The private leaderboard is calculated with approximately 50% of the test data.

This competition has completed. This leaderboard reflects the final standings.

Refresh

#	Δ pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	LAI,WEI-TING			0.88707	104	2d
2	—	Kuan Yu Lin			0.88380	186	5d
3	—	gazer099			0.88326	100	2d
4	—	Tony Kuo			0.88071	22	2d
5	—	Just Monika			0.88010	152	2d
6	—	martha			0.87725	34	4d
7	1	yui_			0.87699	54	2d
8	1	Hannah			0.87632	24	2d
9	2	Jui-An Wang			0.87383	8	1d
10	1	juhuitseng			0.87353	9	2d

程式報告

第一版

一開始對資料該怎麼處理毫無頭緒，決定先只拿event_time當feature，透過datetime把資料時間歸類至我設定的Class Week的instance，裡面含有7個Class Day的instance，分別代表星期一至星期日，再裡面又含有每天的4個time slot，總共就是7(天) * 4(slots)= 28個變數，規類的方法為離8/14一週內的權重為1/2的1次方，一週到兩週間的權重為1/2的2次方...以此類推，將每個user_id特定星期的特定slot經權重計算完後儲存起來，希望透過這種篩選方式讓較近期的資料對結果影響比較大，將此處理後的資料寫成第一版答案的csv提交後，得分為0.80xxx，不斷調整權重後，此種方法最後得分為0.82。

第二版

後來從同學那得知直接套用lightgbm比較方便後，研究了一下xgboost並將第一版答案(0.80xxx)作為前處理資料餵進去train出模型後，將模型的預測提交得分為0.86xxx，將訓練模型的參數調整後，此種方法最後得分為0.872xx。

第三版

抱著嘗試的心態改用lightgbm去訓練模型，模型預測提交得分為0.876xxx，將訓練模型的參數調整後，此種方法最後得分為0.87778，private data得分為0.88010，即為最後結果。

程式簡介

reference_calculator.ipynb: 產生前處理資料(data01~data45)檔案reference.csv

kaggle_v2.ipynb: 產出前處理資料(data46~data75)檔案answer_v2.csv

kaggle_v7.ipynb: 產出提交檔案answer_v7.csv

parameter_tuning.ipynb: 簡易篩選模型參數程式

以上程式碼檔案，皆未好好將code重新整理乾淨，如有傷眼，深感抱歉。