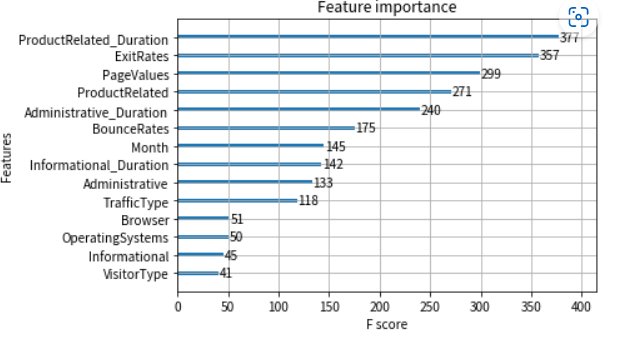
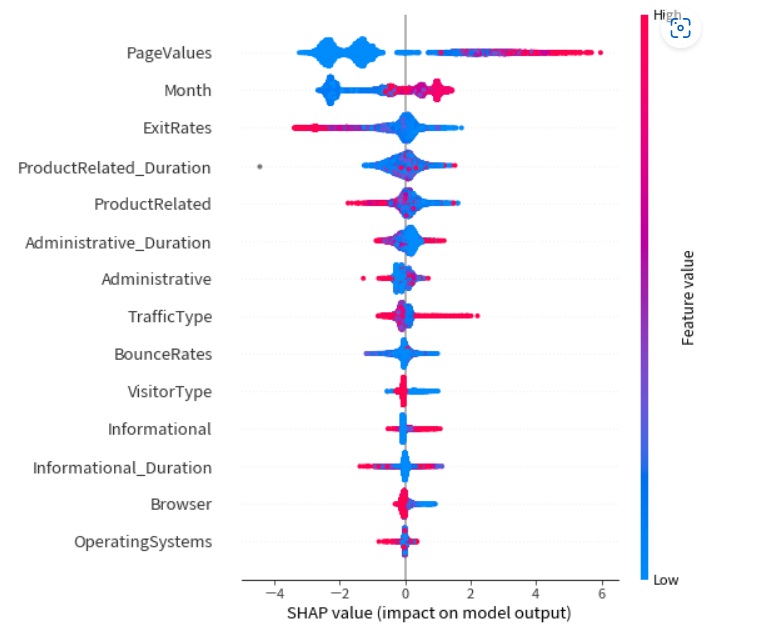
敦謙面試題目 許詠淇

資料預處理:這次我是在colab的環境下實作。使用的是XGBoost分類模型，為了做模型的評估，我先在train.csv內的資料先做切分及訓練，資料集的切分分為8成訓練集及2成驗證集;因為此資料集中並沒有發現明顯的離群值，所以我並沒有對資料集做正規化。

參數的調整及特徵工程:在參數的調整上我利用了GridSearch來幫助我做參數的調校，因為資料的數量較多所以在GridSearch上先調整某一變數並固定其他變數以加快GridSearch的運行，在選擇特徵值上，我使用了SHAP及xgboost內的feature importance來做輔助，在測試後我drop掉了三個特徵值，分別為:SpecialDay、Weekend及Region

一張含有 文字 的圖片

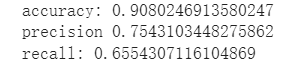
自動產生的描述

模型評估:

|  |  |  |
| --- | --- | --- |
|  | 預測為真 | 預測為假 |
| 實際為真 | 175 | 92 |
| 實際為假 | 57 | 1296 |

一張含有 文字 的圖片

自動產生的描述



加分題:

因模型A的Accuracy:0.89811

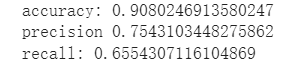
Precision:0.7144

Recall:0.5996

而模型B的Accuracy:0.9001

Precision:0.7280

Recal:0.5920

而我這次訓練模型的

B模型較接近我訓練模型的數值，所以我認為B模型較適用於本案例