Report

F74052201 陳鈺潔

**Choose a dataset：**Online Shoppers Purchasing Intention Dataset

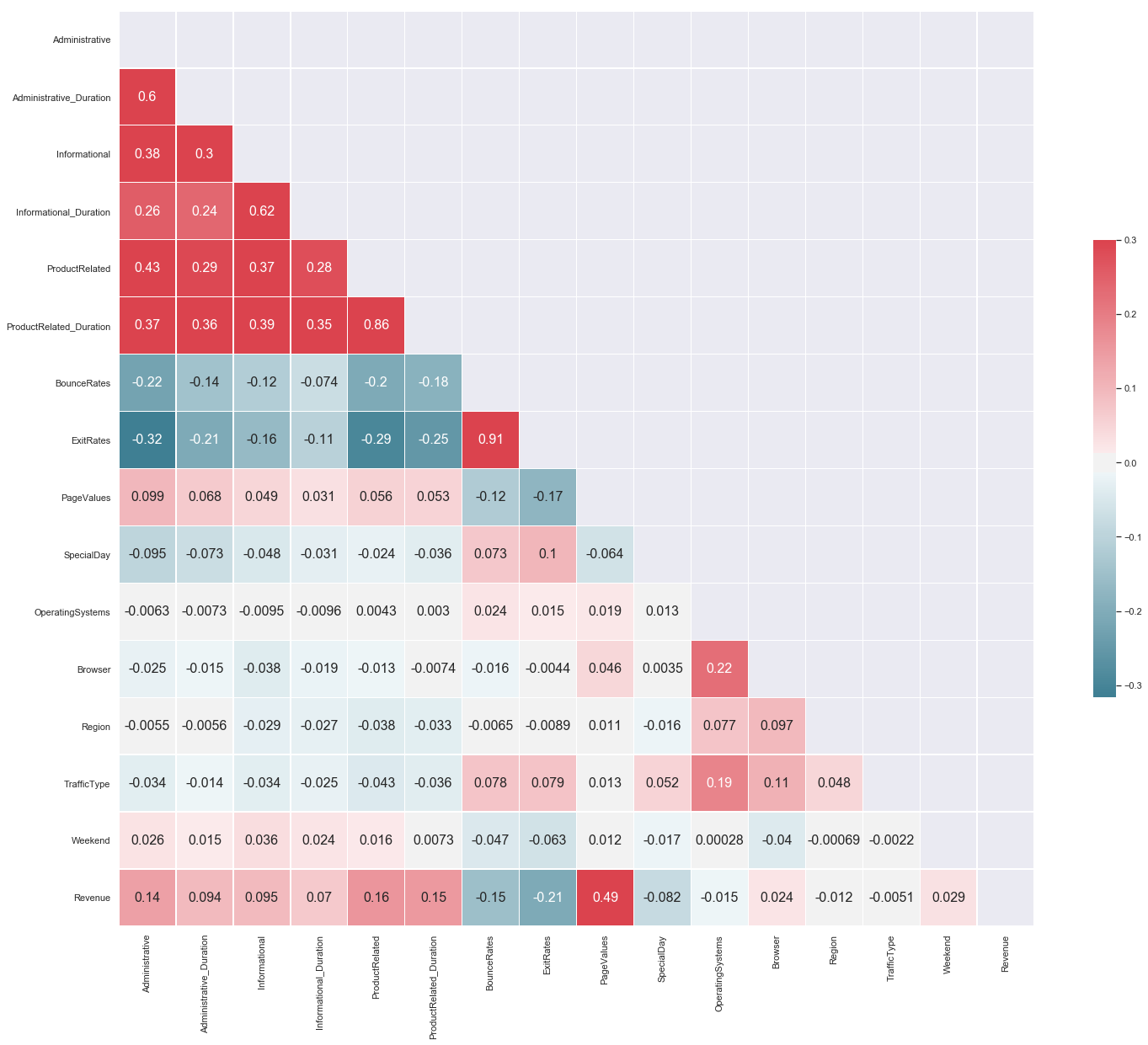
**Analyze the data：**

1. 轉換類別

* Weekend、Revenue從True/False轉成1/0。
* Month從字串轉成整數，Ex：Feb轉成2、May轉成5。
* VisitorType使用One-Hot Encoding，將New\_Visitor、Other、Returning\_Visitor三種個別建立新特徵，值為1/0。

1. 相關係數

發現PageValues和Revenue的相關性最高，適合作為特徵。



圖表 1

1. 資料視覺化

* Adminisrative(觀看管理頁面的次數)
* Adminisrative\_Duration(觀看管理頁面的時間)
* Informational(觀看資訊頁面的次數)
* Informational\_Duration(觀看資訊頁面的時間)
* ProductRelated(觀看商品相關頁面的次數)
* ProductRelated\_Duration(觀看商品相關頁面的時間)
* BounceRates(進入頁面後直接從該頁離開的機率)
* ExitRates(最後從此頁離開的機率)
* PageValues(在抵達目標網頁前的拜訪的網頁平均價值)
* SpecialDay(該筆資料多接近特殊節日)
* Month(該筆資料觀看月份)
* OperationSystems(該筆資料使用之作業系統)
* Browser(該筆資料使用之瀏覽器)
* Region(該筆資料觀看者之區域)
* TrafficType(該筆資料觀看者交通類別)
* VisitorType(該筆資料觀看者種類)
* Weekend(該筆資料是否在周末觀看)

1. 轉換數值

* ProductRelated vs ProductRelated\_Duration：假設兩者關係成正相關，且購物機率越高。下圖橘點和藍點大部分重疊，無法看出購物機率，且兩者皆高的case(靠近右上角的點)最後並沒有購買，但是觀看次數與觀看時間成正相關，且有沒有購物的相關程度相近，假設正確。
* ProductRelated：假設觀看較多次的人購買的可能性越高。平均而言，有購物的人觀看商品頁面的次數，相較沒有購物的人的觀看次數稍微高一點，且觀看次數越多，購買機率越高，假設正確。
* ProductRelated\_Duration：假設觀看較久的人購買的可能性越高。平均而言，有購物的人觀看商品頁面的時間，相較沒有購物的人的觀看時間稍微高一點，且觀看時間越長，購買機率越高，假設正確。
* ExitRates：假設離開商品畫面機率越低，購買機率越高。平均而言，有購物的人離開商品頁面的機率，相較沒有購物的人的離開商品頁面機率低一點，且離開商品畫面機率越低，購買者比沒購買者多，假設正確。
* SpecialDay：假設越接近特殊節日(SpecialDay越大)，購買機率越高。但下圖反映，購買大多集中在不是特別日子，假設錯誤。
* Weekend：假設在假日的購買機率較高。假設正確。

最後觀察Revenue和其他特徵的相關性。發現目前使用的特徵並不是最強的特徵，像是PageValue的相關度最高，但沒有使用，所以還能再提高準確度。

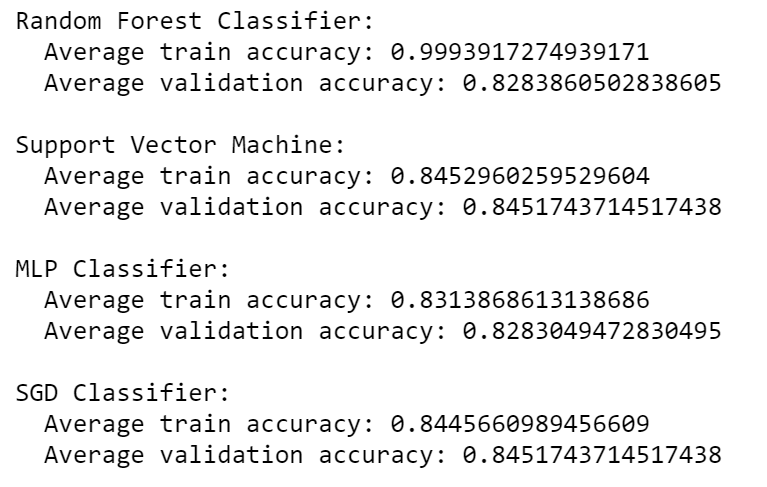
**Define a reasonable problem：**預測Online shopper是否有購買商品

**Explain how you improved your results step-by-step：**

1. Initial Result：

直觀使用五個可能會使線上逛街者購買的特徵(ProductRelated、ProductRelated\_Duration、ExitRates、SpecialDay、Weekend)，只將Weekend從True/False轉換成1/0。

使用Random Forest、SVM、MLP三種未調整超參數的模型進行訓練(5-cross validation)。Random Forest、MLP的驗證準確率82較低、SVM有驗證84準確率。



1. Reasons：使用ProductRelated、ProductRelated\_Duration、ExitRates、SpecialDay、Weekend做了五種假設。

一、觀看越多次商品頁面的人購買的可能性越高。

二、在商品頁面時間越長的人購買的可能性越高。

三、最後從離開商品畫面機率越低，購買機率越高。

四、越接近特殊節日，購買機率越高。

五、周末的購買機率較高。

以上五點的假設在分析資料後發現，僅有第四點假設不完全正確，但是沒有使用到最強的特徵PageValues，所以準確率約8成。

1. My approaches：

選擇使用()feature，使用()模型，設定超參數為()。

1. Improvement：
2. 特徵篩選：PageValue與Revenue相關性最高，Administrative、Month、BounceRates相關係數也高(不管正、負相關)，將此四項特徵加入；Special Day的相關性低，刪除此特徵。三者準確度都有上升。
3. 修改原特徵：
4. 調整模型超參數