Report

F74052201 陳鈺潔

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

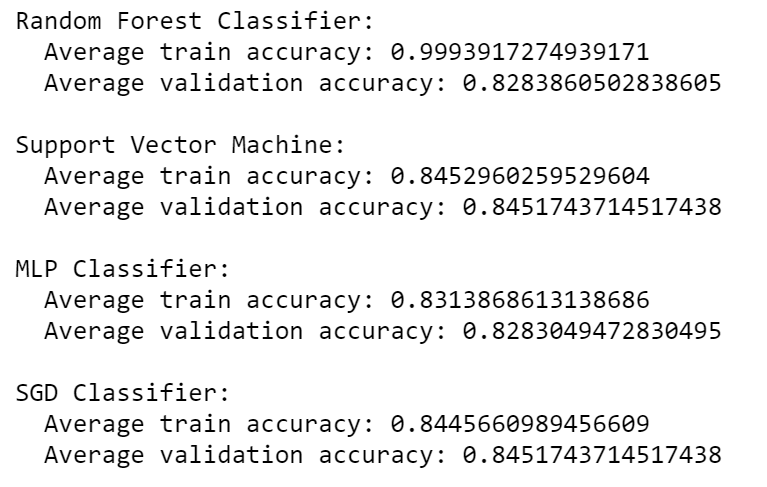
**Analyze the data：**

* ProductRelated vs ProductRelated\_Duration
* ProductRelated
* ProductRelated\_Duration
* ExitRates
* SpecialDay
* Weekend

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

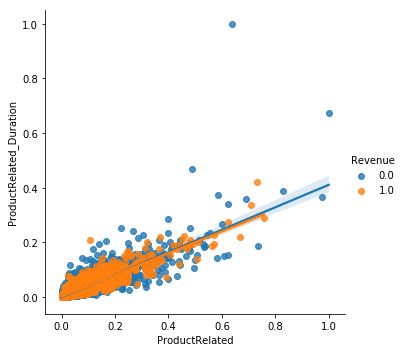
**Explain how you improved your results step-by-step：**

1. Initial Result：先直觀使用五個可能會使線上逛街者購買的特徵(觀看跟商品相關頁面次數(ProductRelated)、停留在商品相關頁面的時間(ProductRelated\_Duration)、從頁面離開的機率ExitRates、 多靠近特殊節日SpecialDay、 是否在週末Weekend)，並不做任何資料分析。使用Random Forest、SVM、MLP、Stochastic Gradient Descent四種模型(不調整超參數)進行訓練(5-cross validation)，Random Forest、MLP的準確率82%較低、SVM有84%準確率。

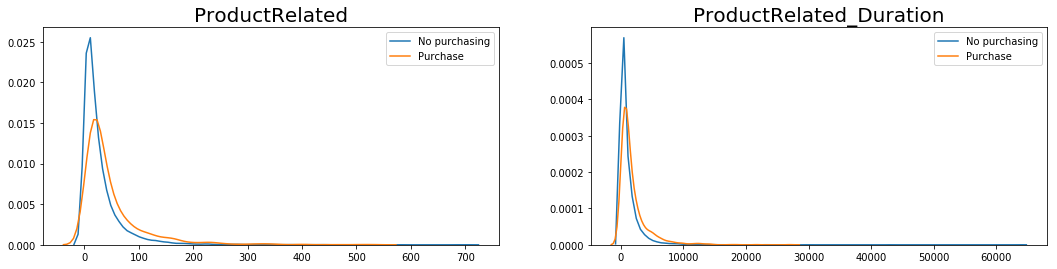


1. Reasons：在只使用五個特徵且沒有調整超參數的狀況下就能得到8成的準確度，可見推測有一定準確性，觀察這些特徵是否和有購買商品確實有關。

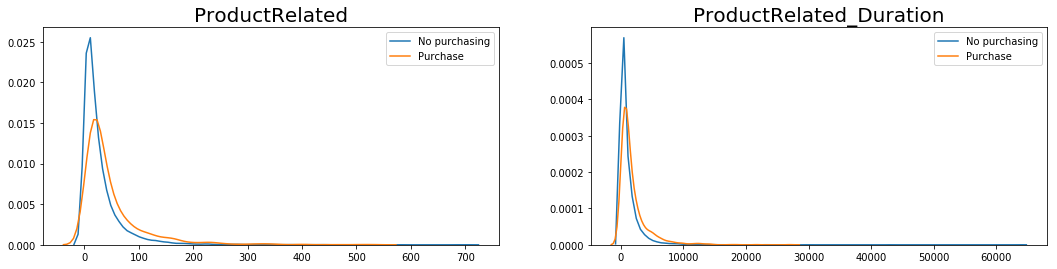
* ProductRelated vs ProductRelated\_Duration：假設兩者關係成正相關，且購物機率越高。下圖橘點和藍點大部分重疊，無法看出購物機率，且兩者皆高的case(靠近右上角的點)最後並沒有購買，但是觀看次數與觀看時間成正相關，且有沒有購物的相關程度相近，假設正確。



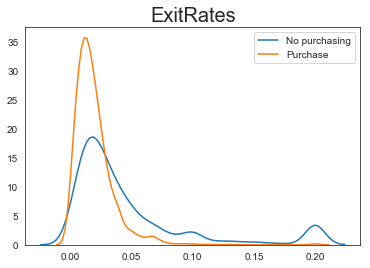
* ProductRelated：假設觀看較多次的人購買的可能性越高。平均而言，有購物的人觀看商品頁面的次數，相較沒有購物的人的觀看次數稍微高一點，且觀看次數越多，購買機率越高，假設正確。



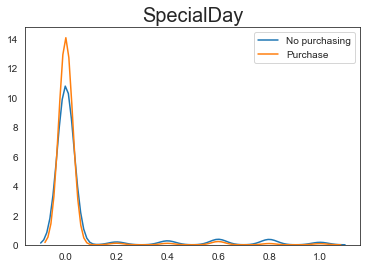
* ProductRelated\_Duration：假設觀看較久的人購買的可能性越高。平均而言，有購物的人觀看商品頁面的時間，相較沒有購物的人的觀看時間稍微高一點，且觀看時間越長，購買機率越高，假設正確。



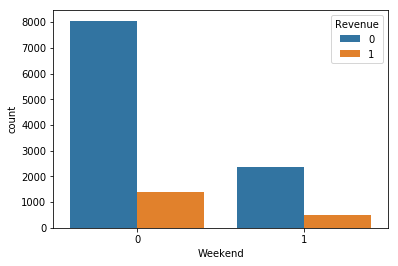
* ExitRates：假設離開商品畫面機率越低，購買機率越高。平均而言，有購物的人離開商品頁面的機率，相較沒有購物的人的離開商品頁面機率低一點，且離開商品畫面機率越低，購買者比沒購買者多，假設正確。

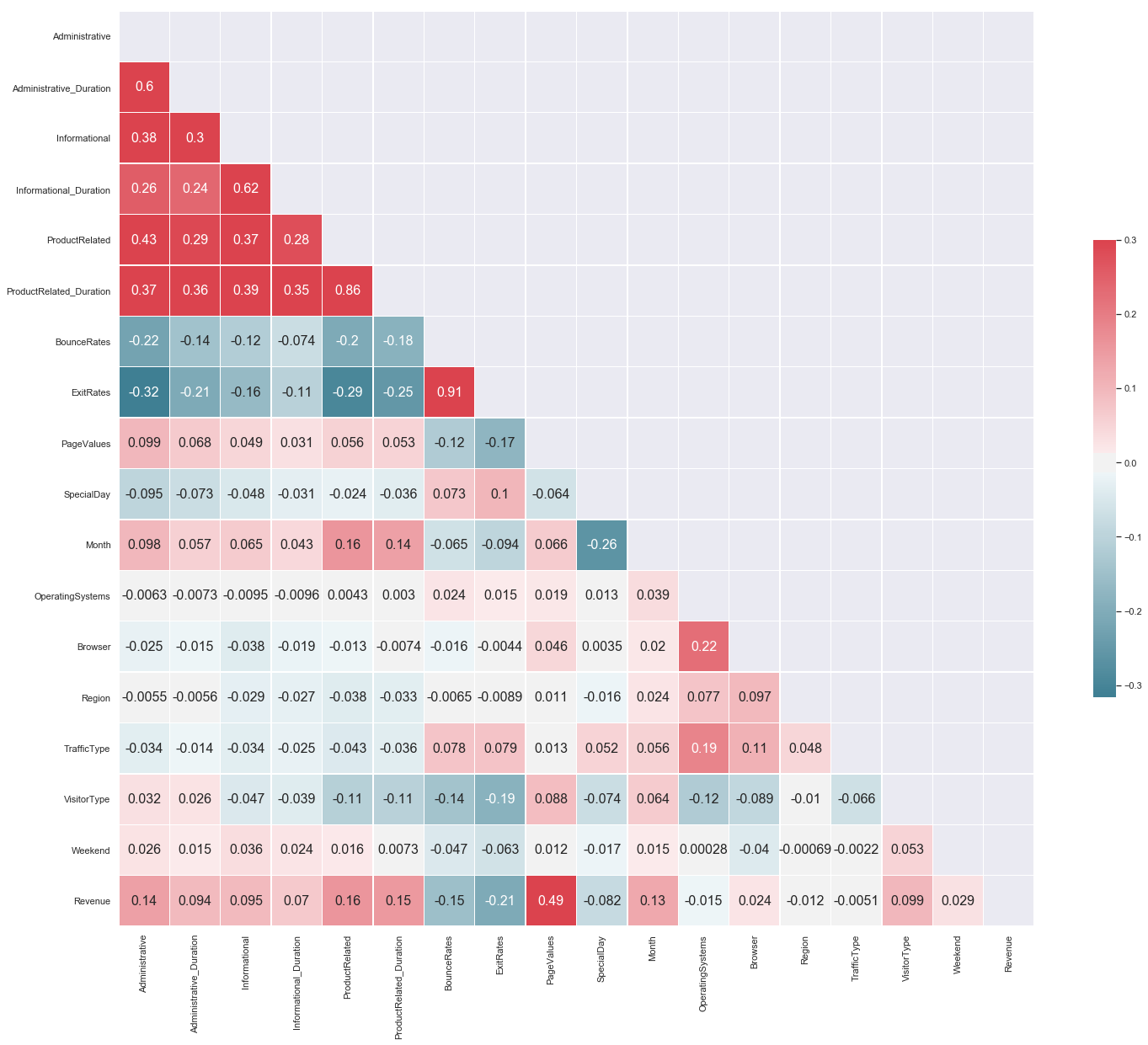


* SpecialDay：假設越接近特殊節日(SpecialDay越大)，購買機率越高。但下圖反映，購買大多集中在不是特別日子，假設錯誤。

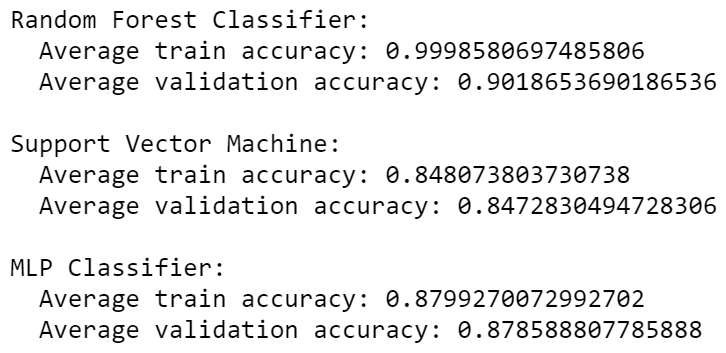


* Weekend：假設在假日的購買機率較高。假設正確。



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。

1. My approaches：
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。



1. 修改原特徵：

* Explorer
* Traffic Type

1. 修改模型
2. 調整模型超參數
3. Improvement：