Report

F74052201 陳鈺潔

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

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

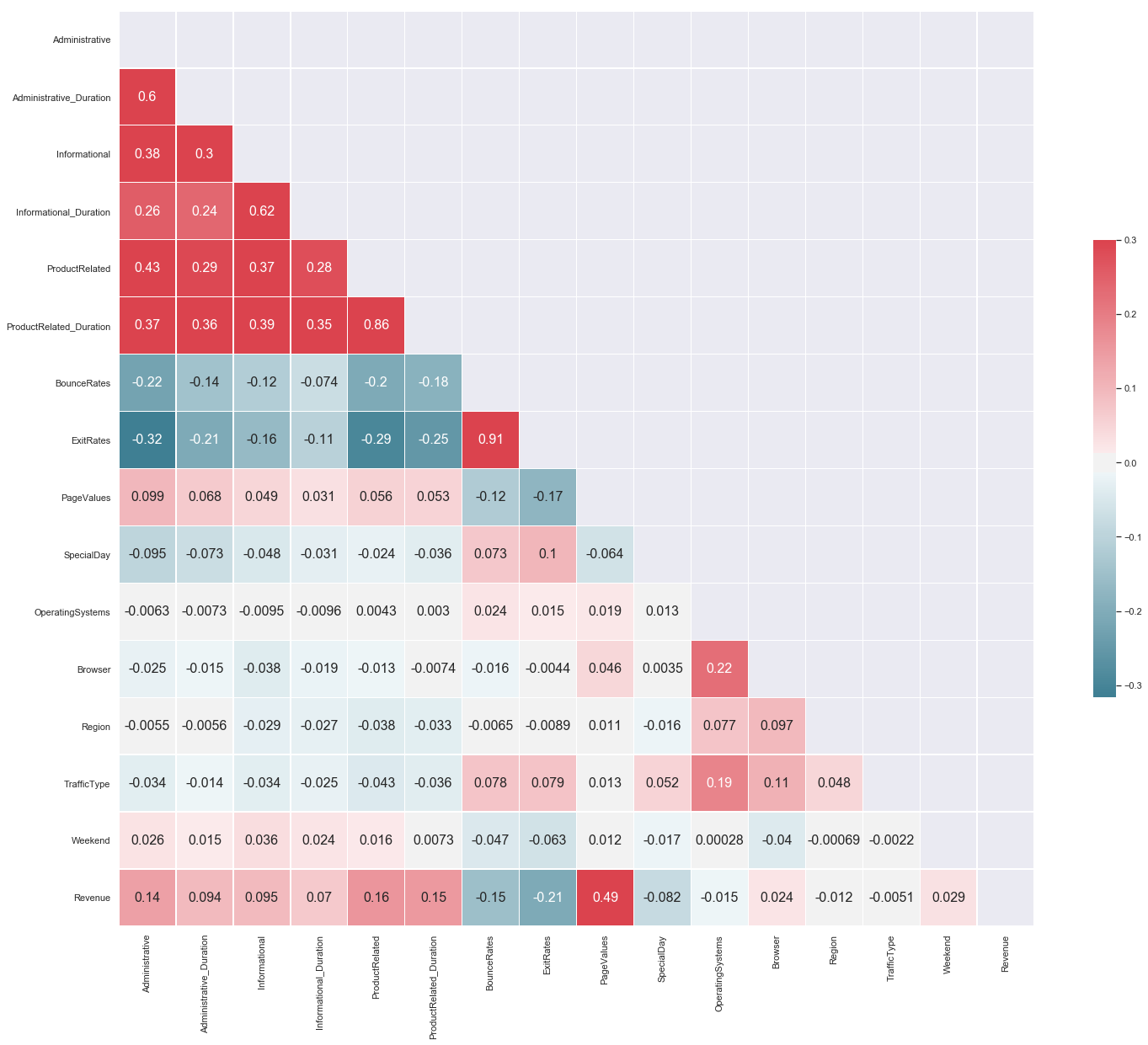
**Analyze the data：**

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。

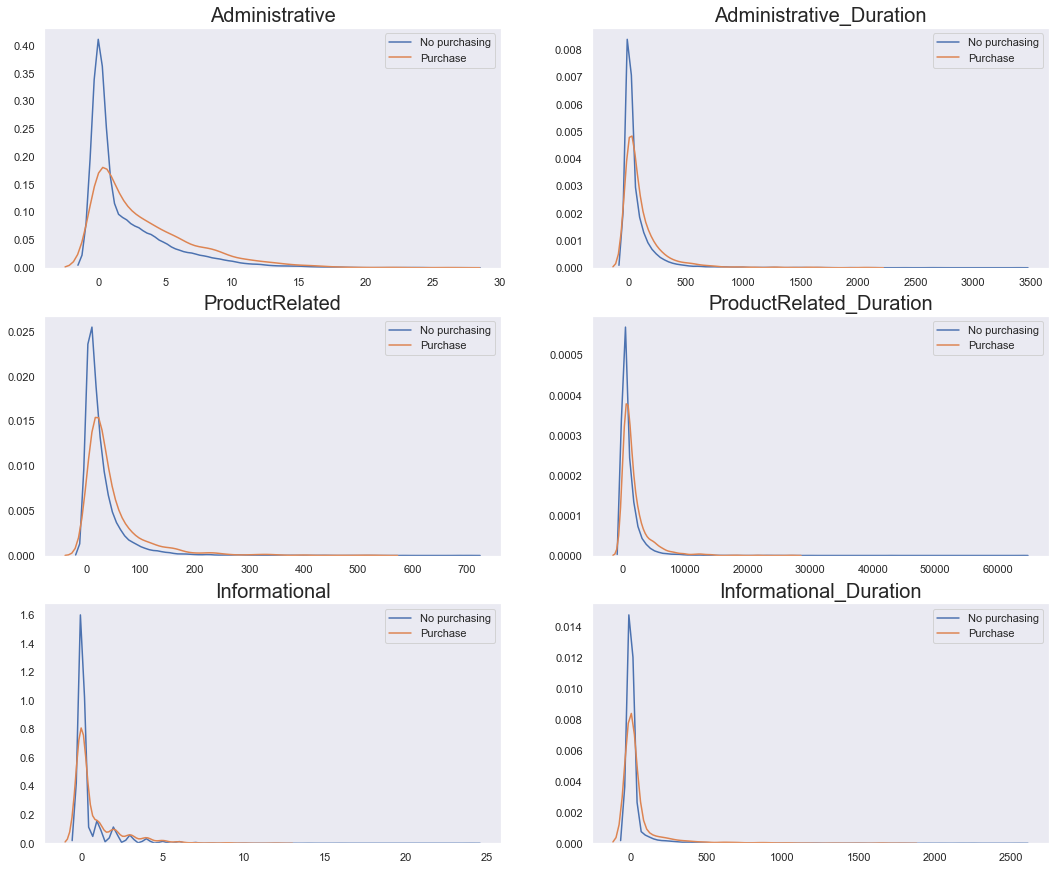
1. 相關係數

發現PageValues和Revenue的相關性最高，適合作為特徵。

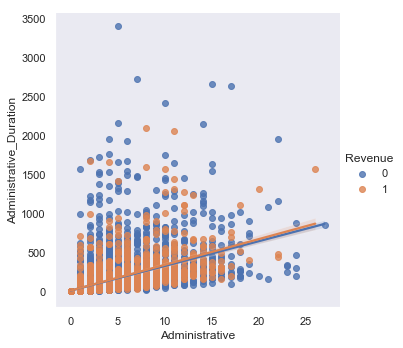


1. 資料視覺化

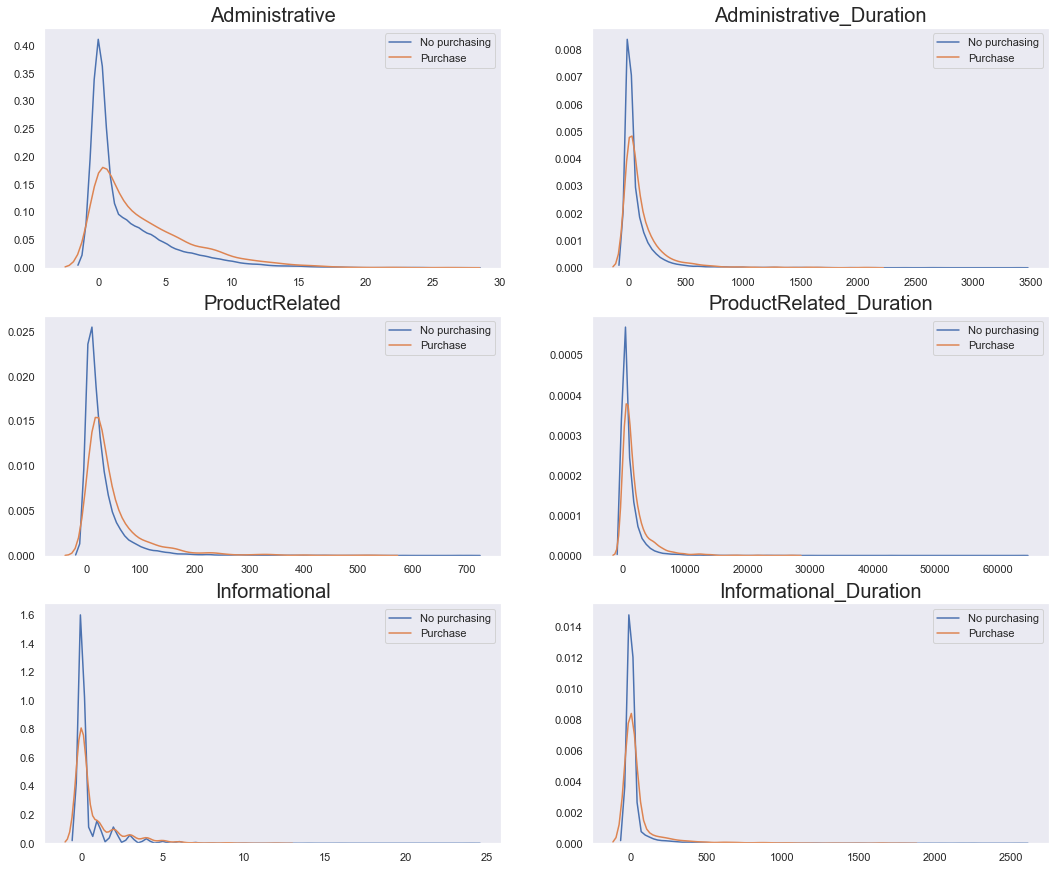
* Adminisrative(觀看管理頁面的次數)：有購物的人平均觀看管理頁面的次數，相較沒有購物的人的觀看次數略高一點。
* Adminisrative\_Duration(觀看管理頁面的時間)：有購物的人平均觀看管理頁面的時間，相較沒有購物的人的觀看時間稍微高一點。



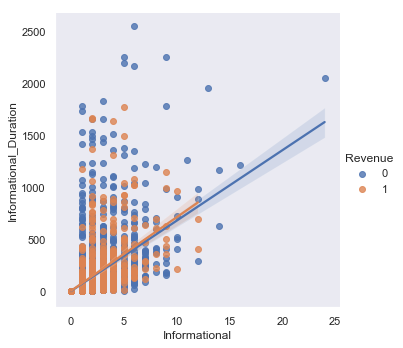
* Administrative vs Administrative\_Duration：不論有沒有購物，兩者關係成正相關。有購物者的兩者值較集中左下角。



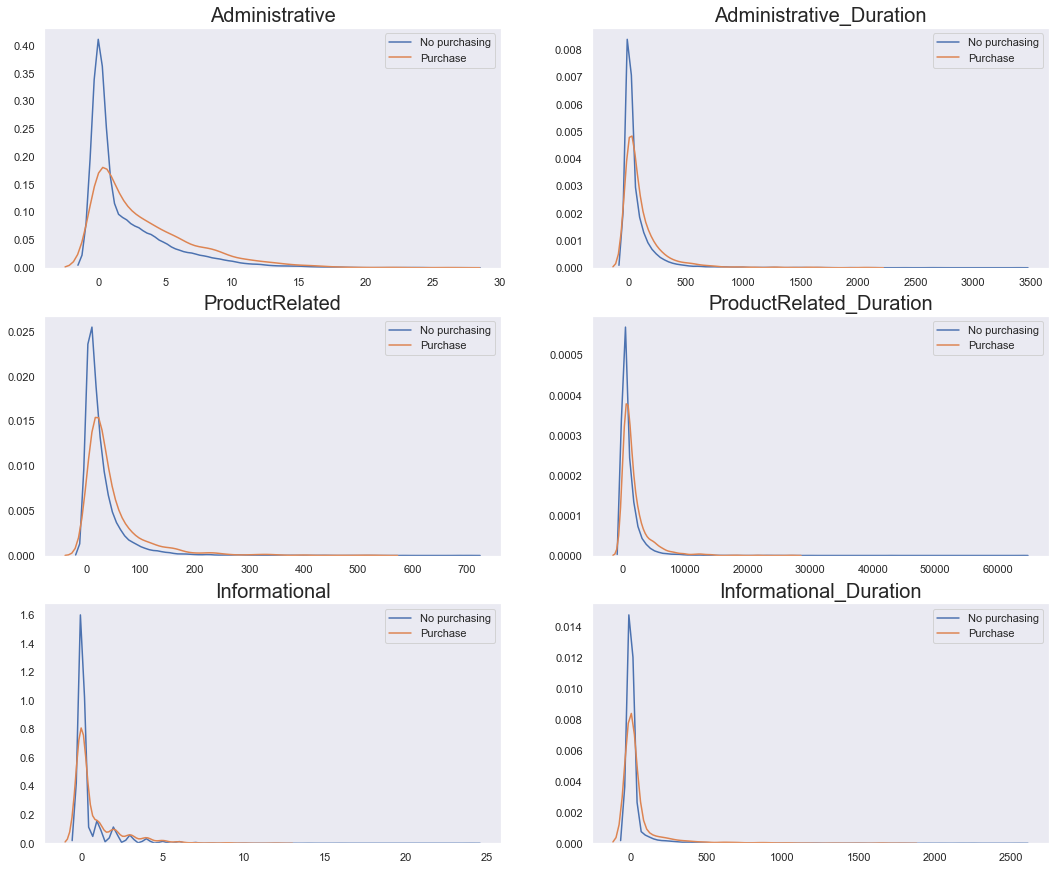
* Informational(觀看資訊頁面的次數)：有沒有購物觀看資訊頁面的平均次數幾乎相同。
* Informational\_Duration(觀看資訊頁面的時間)：有沒有購物觀看資訊頁面的平均時間幾乎相同。



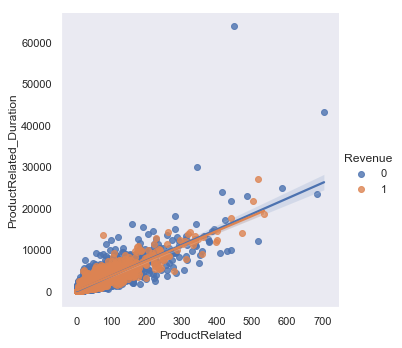
* Informational vs Informational\_Duration：不論有沒有購物，兩者關係成正相關。有購物者的觀看次數值偏低、觀看時間長短不一。



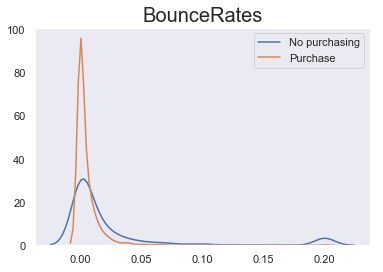
* ProductRelated(觀看商品相關頁面的次數)：有購物的人平均觀看商品頁面的次數，相較沒有購物的人的觀看次數略高一點。
* ProductRelated\_Duration(觀看商品相關頁面的時間)：有購物的人平均觀看商品頁面的時間，相較沒有購物的人的觀看時間稍微高一點。



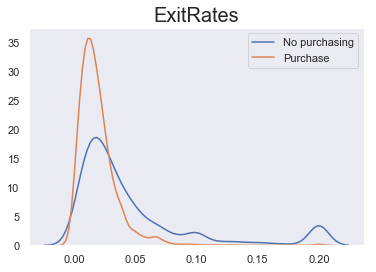
* ProductRelated vs ProductRelated\_Duration：不論有沒有購物，兩者關係成正相關。有購物者的兩者值較集中左下角，且接近回歸線分布。



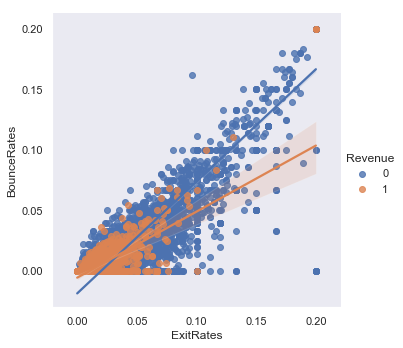
* BounceRates(進入頁面後直接從該頁離開的機率)：直接進入頁面後離開機率越低，購買機率越高。不管有沒有購物，BounceRates平均都極低且幾乎相同。



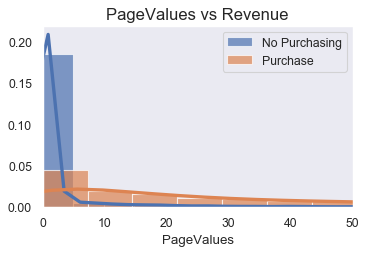
* ExitRates(最後從此頁離開的機率)：離開該商品畫面機率越低，購買機率越高。有購物者的平均ExitRates較沒購物者低。



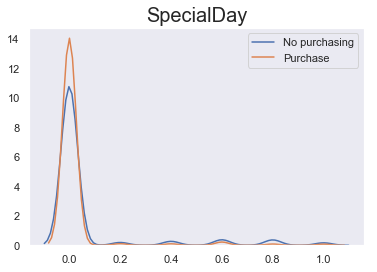
* ExitRates vs BounceRates：不論有沒有購物，兩者關係成正相關、但有購物的相關度較沒購物低。有購物者的兩者值較集中左下角。



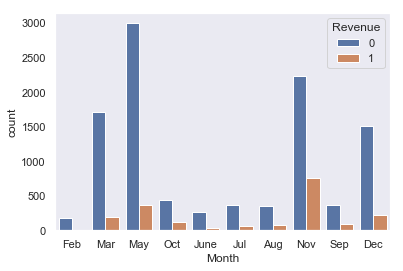
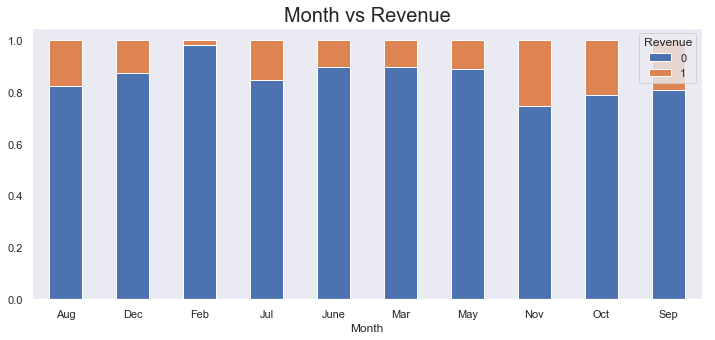
* PageValues(在抵達目標網頁前的拜訪的網頁平均價值)：平均網頁價值越高，購買機率較高。

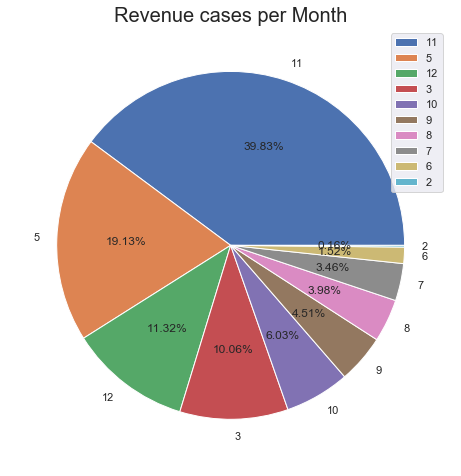


* SpecialDay(該筆資料多接近特殊節日)：購物機率在一般日子最高。

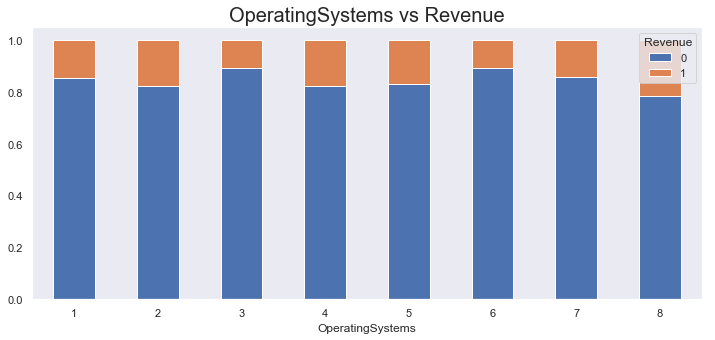


* Month(該筆資料觀看月份)：11月的購物機率最高，2月購物機率最低，5月的觀看數量最多。有購物的部分，11、5、12月最多筆購物，2月最少。

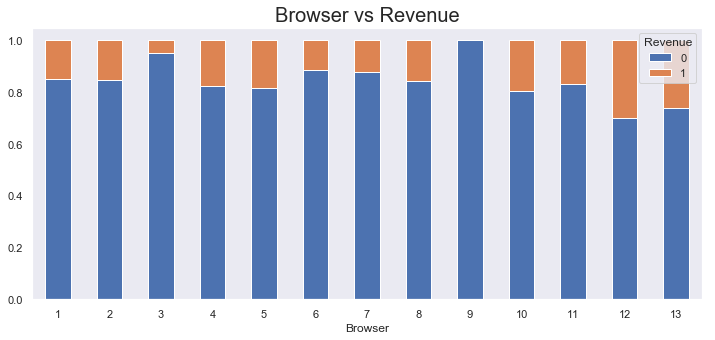
 



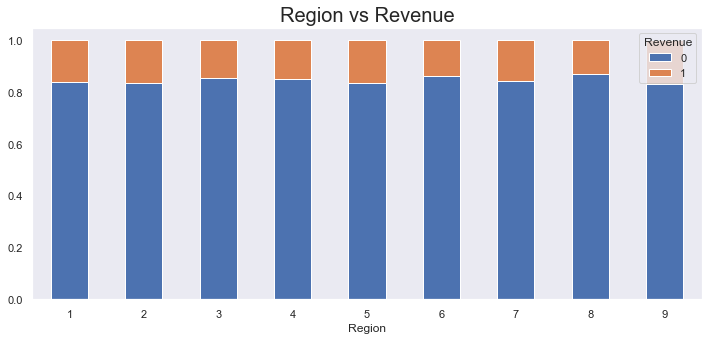
* OperationSystems(該筆資料使用之作業系統)：作業系統與購物機率影響不大。



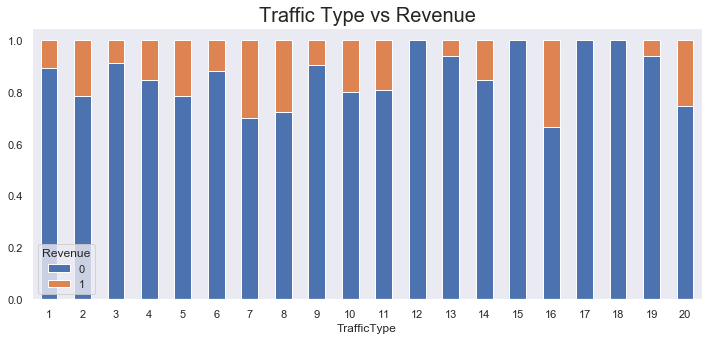
* Browser(該筆資料使用之瀏覽器)：瀏覽器在第9種瀏覽器的購物機率最低，但是只有一筆資料，無法評估。



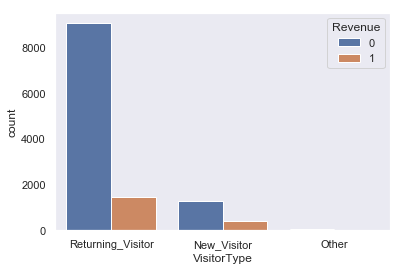
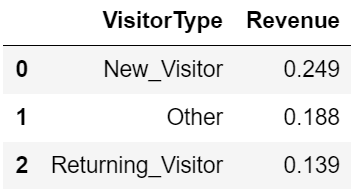
* Region(該筆資料觀看者之區域)：每個區域的購物機率差不多。



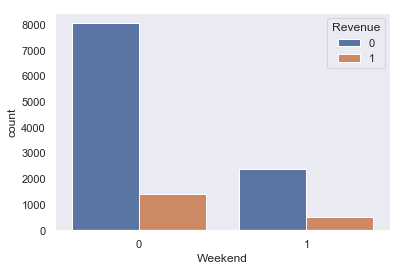
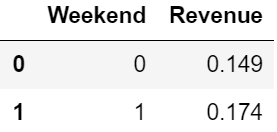
* TrafficType(該筆資料觀看者交通類別)：在第12、15、17、18類的購物機率最低。



* VisitorType(該筆資料觀看者種類)：新顧客的購物機率最高，但是是因為回訪的顧客觀看次數較多的關係，回訪顧客的購物數最高。

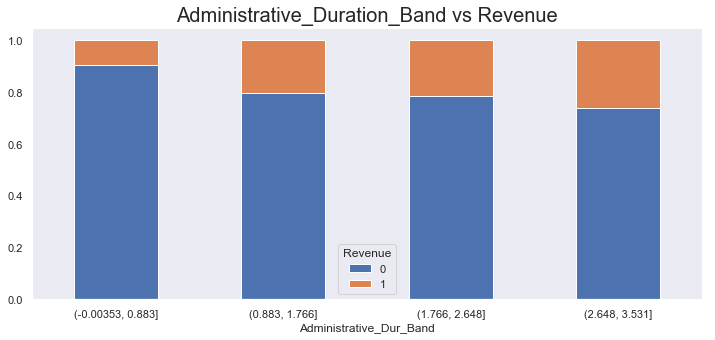


* Weekend(該筆資料是否在周末觀看)：是不是在假日的購物機率差不多。但是非假日的觀看次數較多。

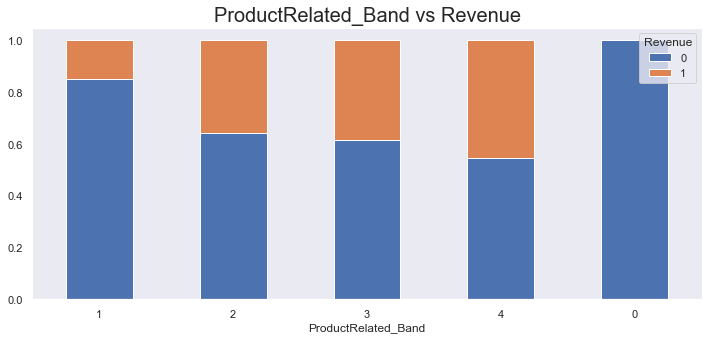


1. 轉換數值

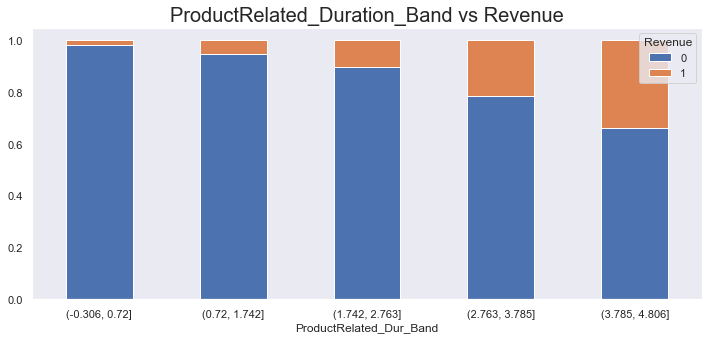
* Administrative\_Dur\_Band：由於Administrative\_Duration範圍大，取log後切成4個數值區間，發現值越大購買機率越高。再將區間轉換為0, 1, 2, 3。



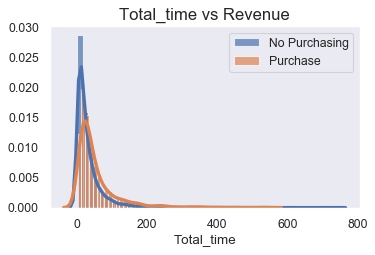
* ProductRelated\_Band：將ProductRelated切成5個數值區間，發現值越大購買機率越高，但是值在最大的區間時，購買率又為0。按照購買機率大小將區間轉成0~4。



* ProductRelated\_Dur\_Band：由於ProductRelated\_Duration範圍大，取log後切成5個數值區間，發現值越大購買機率越高。再將區間用label encoding轉換為0, 1, 2, 3。



* Total\_time：將Administrative、Informational、ProductRelated相加。有購物的人平均觀看總時間，相較沒有購物的人的觀看時間稍微高一點。



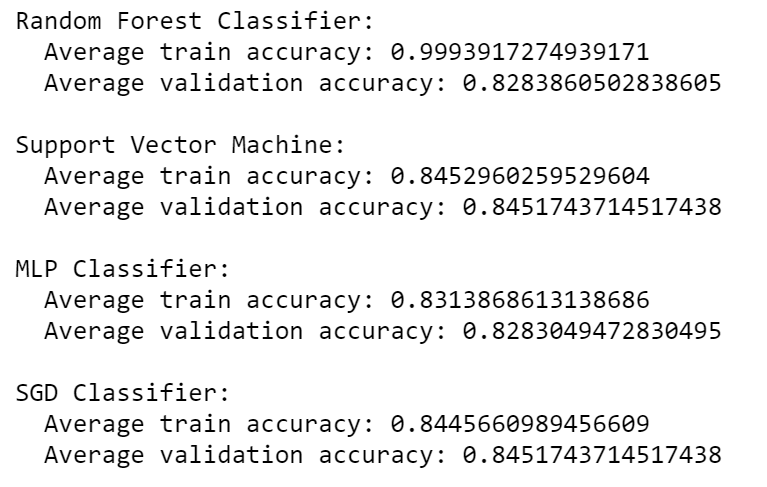
* ExitRates\_Thres、BounceRates\_Thres：分別以平均ExitRates、BounceRates值作為Threshold，高於此值為1、反之為0。
* PageValues\_Thres：Threshold為10，PageValue高於此值為1、反之為0。
* Month\_2~Month12：使用One-Hot Encoding，將Month中所有月份個別建立新特徵，值為1/0。
* Traffic\_noRevenue：把購物機率特低的12、15、17、18類值設為1，其他設為0。

**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。

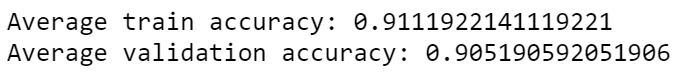
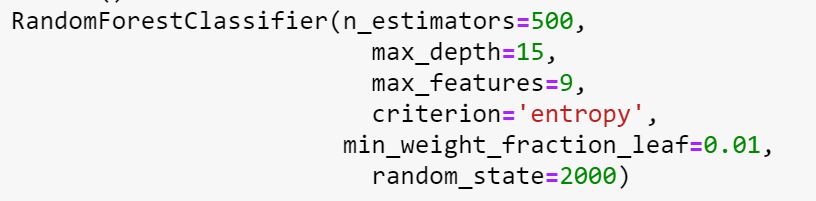
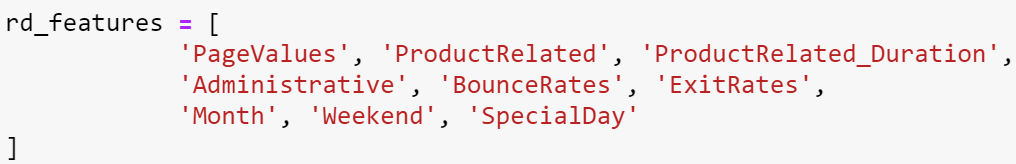
五、周末的購買機率較高。

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，所以準確率還有機會再提高。

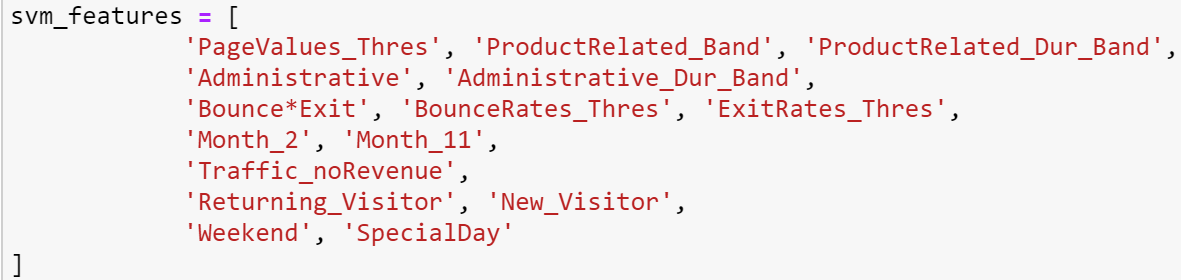
1. My approaches：

使用Random Forest、SVM、MLP三種模型，並用5-cross validation逐步調整特徵，以得到最高的驗證準確度的特徵再調整超參數。

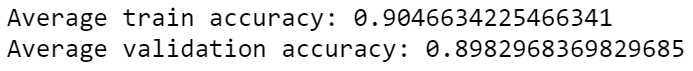
* Random Forest：使用下圖特徵與超參數，得到90.5驗證準確度。



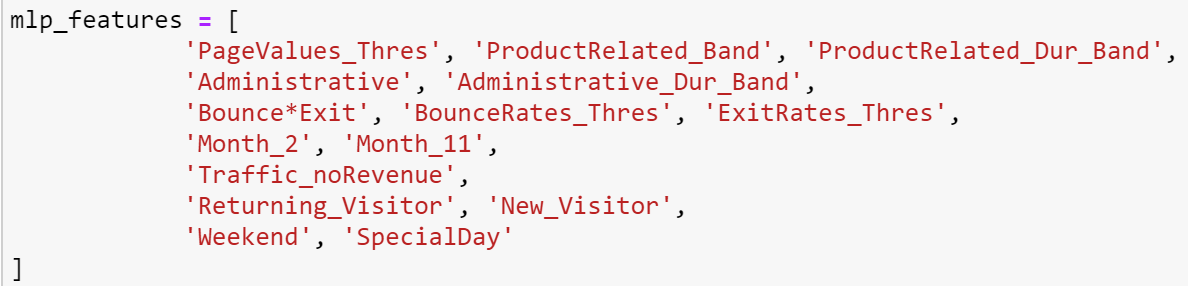
* SVM：使用下圖特徵與超參數，得到89.8驗證準確度。

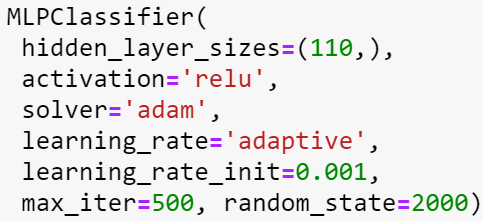


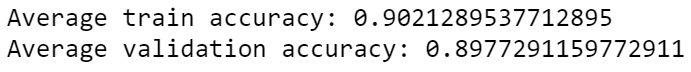




* MLP：使用下圖特徵與超參數，得到89.8驗證準確度。



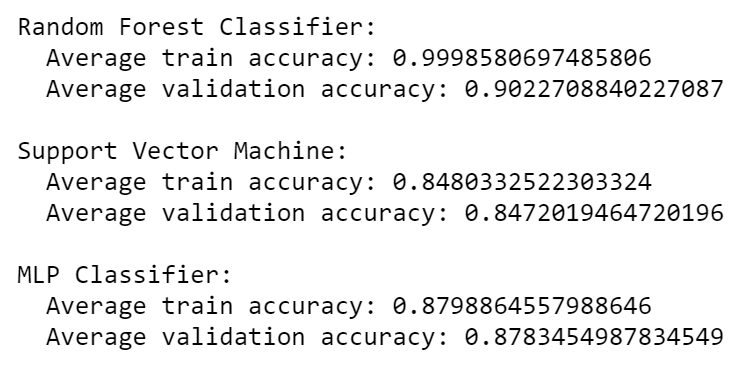




1. Improvement：
2. 特徵篩選

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。Random Forest已達9成驗證準確度，SVM進步幅度最小。

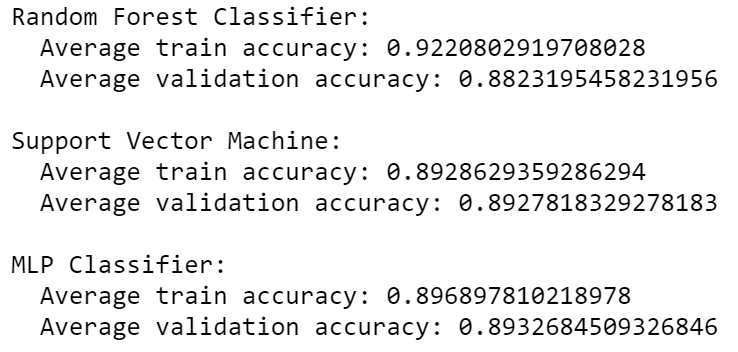


1. 修改原特徵：

將上一步取用的特徵做修改。

1. PageValues改成PageValues\_Thres
2. ProductRelated改成ProductRelated\_Band
3. ProductRelated\_Duration改成ProductRelated\_Dur\_Band
4. BounceRates改成BounceRates\_Thres
5. ExitRates改成ExitRates\_Thres

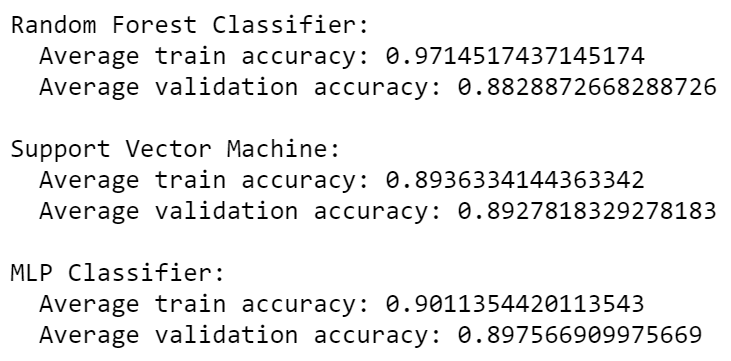
修改原特徵後，Random Forest的驗證準確度下降2，SVM和MLP上升到89。SVM進步幅度最大。



1. 加入創造的新特徵：

* Administrative\_Dur\_Band：Administrative\_Duration取log後切成4個數值區間。區間值越大、購物機率越高。
* Bounce\*Exit：將BounceRates和ExitRates相乘，再用minmax scaler設定範圍在[0,1]。
* Month\_2、Month\_11：Month的one-hot encoding結果，取出2月跟11月，因為2月購物機率最低、11月購物機率最高。
* Traffic\_noRevenue：把購物機率特低的類別和其他區分。
* Returning\_Visitor、New\_Visitor：VisitorType的one-hot encoding結果。

相較上一步的驗證準確度，SVM不變、Random Forest幾乎不變，MLP微幅上升。



1. 確定特徵與調整模型超參數

* Random Forest使用第一步的特徵，調整超參數：500棵最大深度15的決策樹、使用全部特徵、使用entropy值做為決策順序、最小的節點權重0.01。
* SVM使用第三步的特徵，調整超參數：正規化參數為1、kernel funciton使用係數0.4的Radial Basis Function。
* MLP使用第三步的特徵，調整超參數：110層隱藏層、使用relu作為激活函數，用adam做最佳化，初始的learning rate為0.001，學習率會自適應調整。

Random Forest從0.902上升到0.905。

SVM從0.892上升到0.898。

MLP從0.8975上升到0.8977。

