Report

F74052201 陳鈺潔

Choose a dataset：Online Shoppers Purchasing Intention Dataset

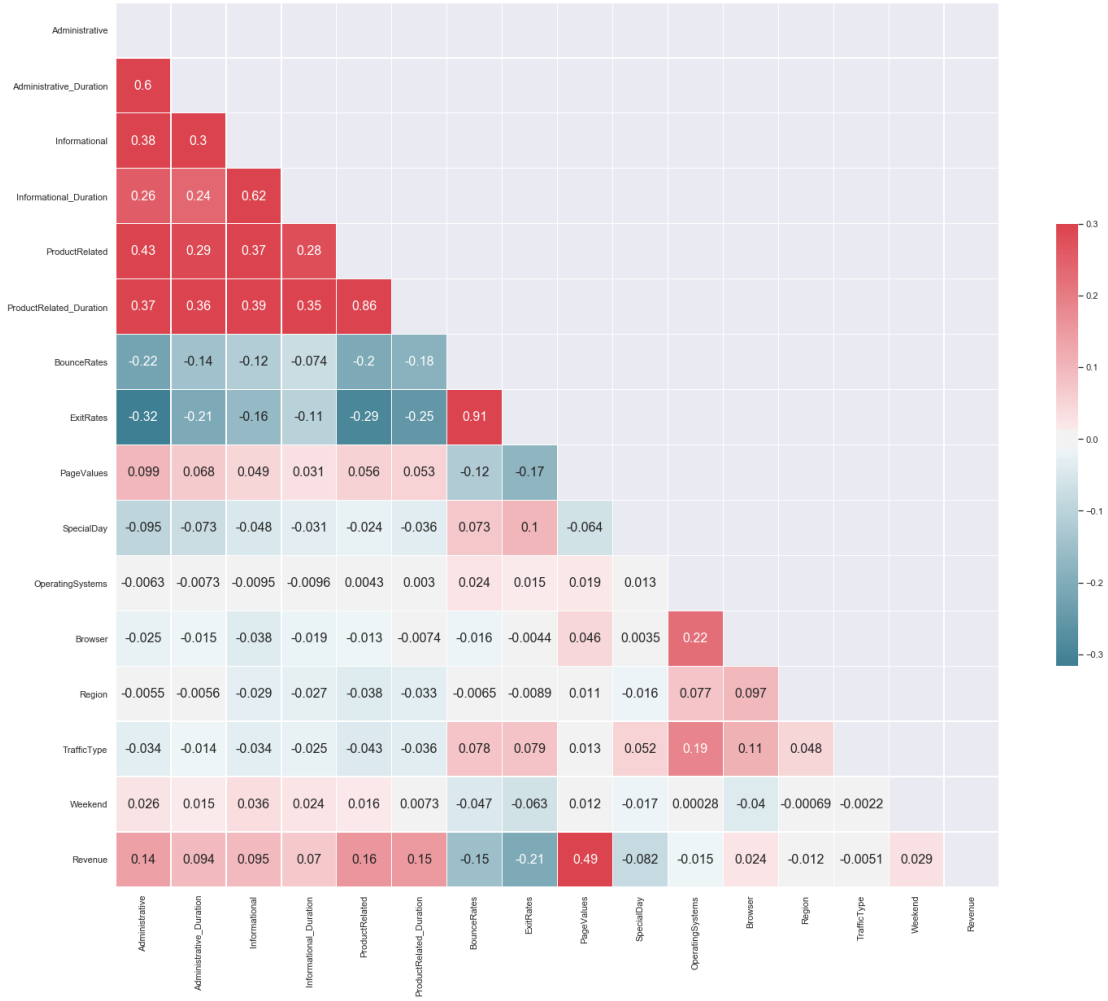
Define a reasonable problem：預測 Online shopper 是否有購買商品

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。

2. 相關係數

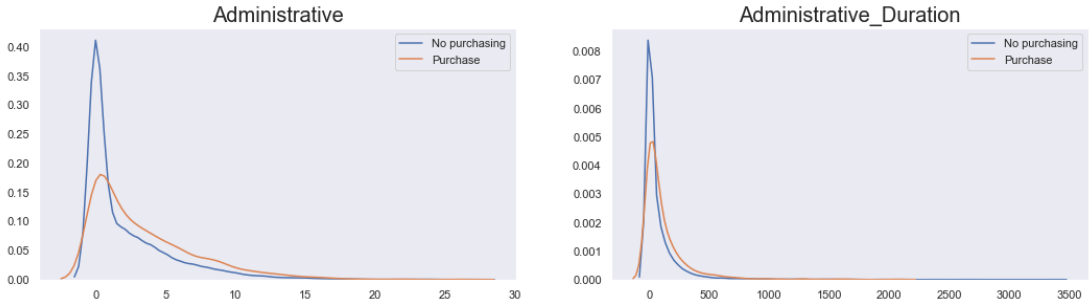
   發現 PageValues 和 Revenue 的相關性最高，適合作為特徵。



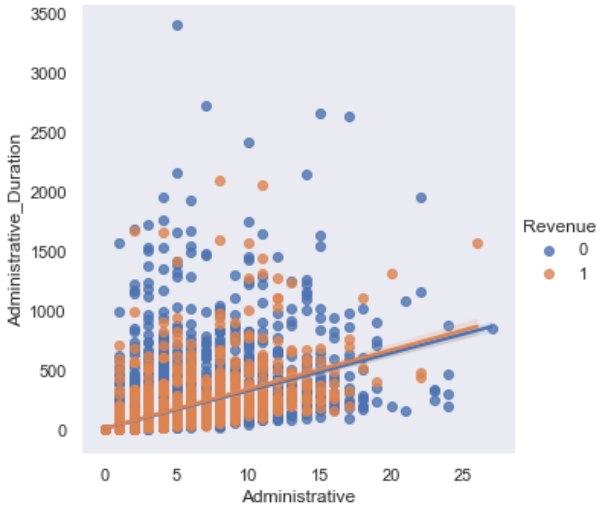
3. 資料視覺化
   - Adminisrative(觀看管理頁面的次數)：有購物的人平均觀看管理頁

面的次數，相較沒有購物的人的觀看次數略高一點。

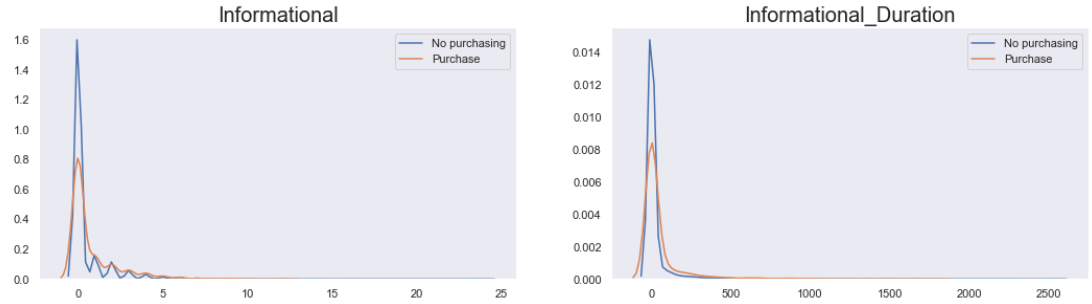
- Adminisrative_Duration(觀看管理頁面的時間)：有購物的人平均觀看管理頁面的時間，相較沒有購物的人的觀看時間稍微高一點。



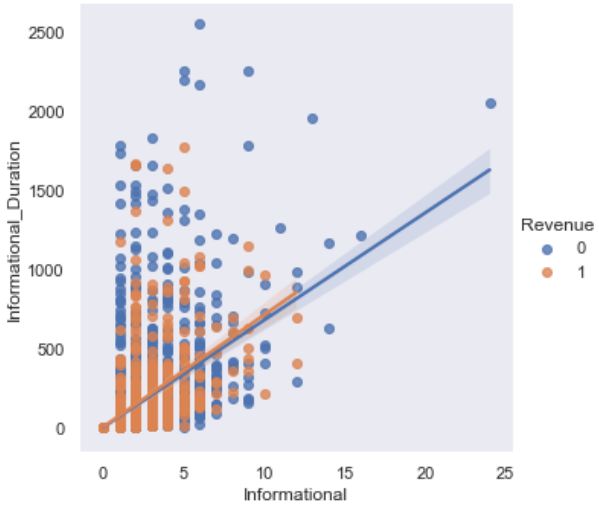
- Administrative vs Administrative_Duration：不論有沒有購物，兩者關係成正相關。有購物者的兩者值較集中左下角。



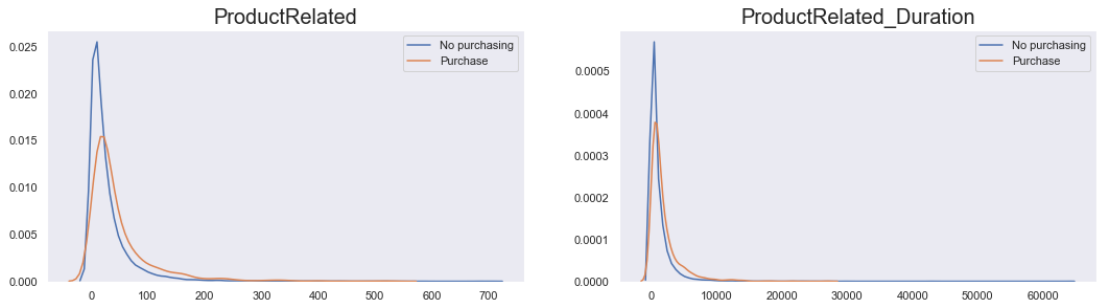
- Informational(觀看資訊頁面的次數)：有沒有購物觀看資訊頁面的平均次數幾乎相同。
- Informational_Duration(觀看資訊頁面的時間)：有沒有購物觀看資訊頁面的平均時間幾乎相同。



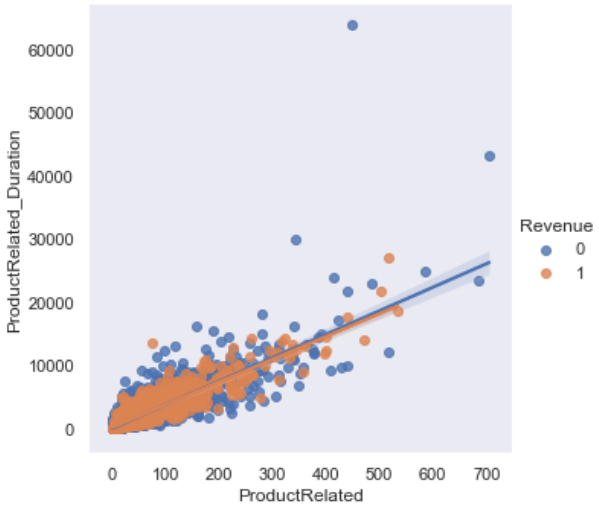
- Informational vs Informational_Duration：不論有沒有購物，兩者關係成正相關。有購物者的觀看次數值偏低、觀看時間長短不一。



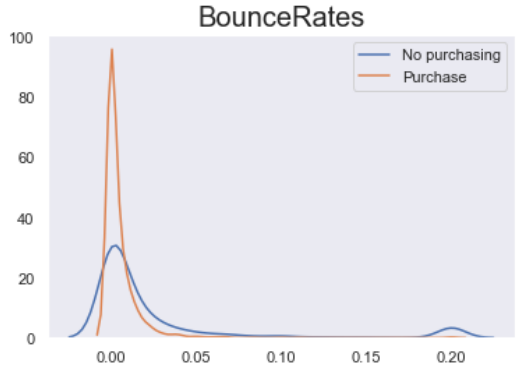
- ProductRelated(觀看商品相關頁面的次數)：有購物的人平均觀看商品頁面的次數，相較沒有購物的人的觀看次數略高一點。
- ProductRelated_Duration(觀看商品相關頁面的時間)：有購物的人平均觀看商品頁面的時間，相較沒有購物的人的觀看時間稍微高一點。



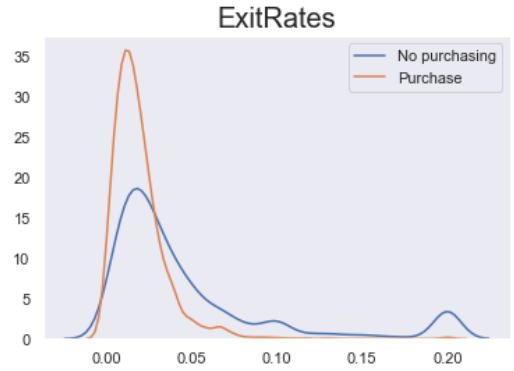
- ProductRelated vs ProductRelated_Duration：不論有沒有購物，兩者關係成正相關。有購物者的兩者值較集中左下角，且接近回歸線分布。



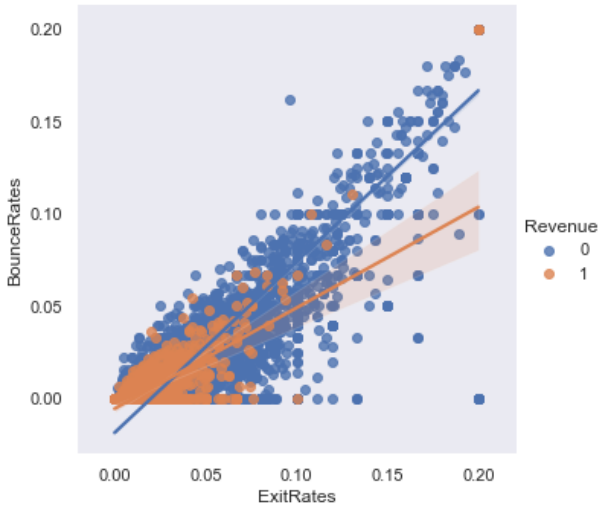
- BounceRates(進入頁面後直接從該頁離開的機率)：直接進入頁面後離開機率越低，購買機率越高。不管有沒有購物，BounceRates 平均都極低且幾乎相同。



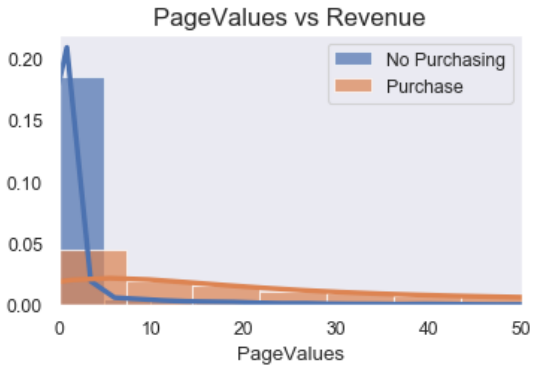
- ExitRates(最後從此頁離開的機率)：離開該商品畫面機率越低，購買機率越高。有購物者的平均 ExitRates 較沒購物者低。



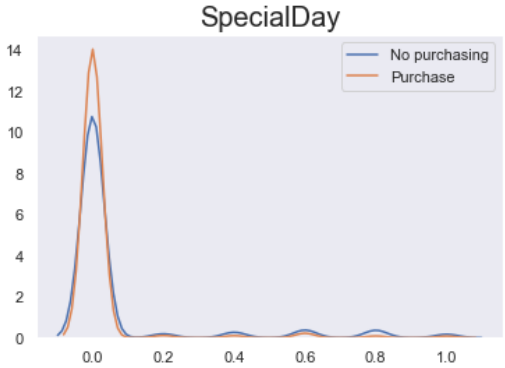
- ExitRates vs BounceRates：不論有沒有購物，兩者關係成正相關、但有購物的相關度較沒購物低。有購物者的兩者值較集中左下角。



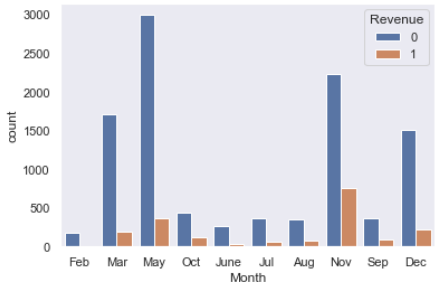
- PageValues(在抵達目標網頁前的拜訪的網頁平均價值)：平均網頁價值越高，購買機率較高。

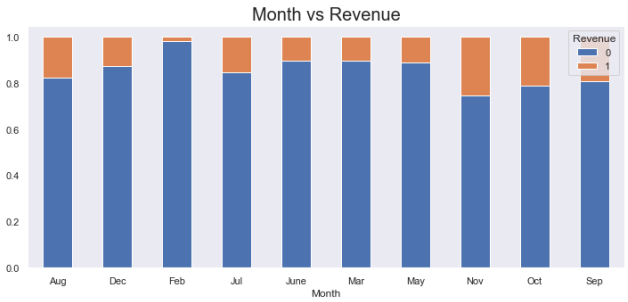


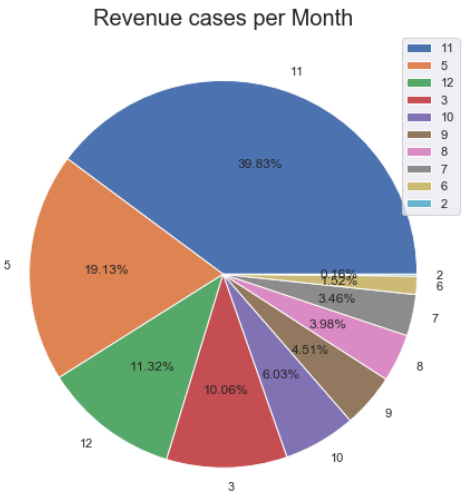
- SpecialDay(該筆資料多接近特殊節日):購物機率在一般日子最高。



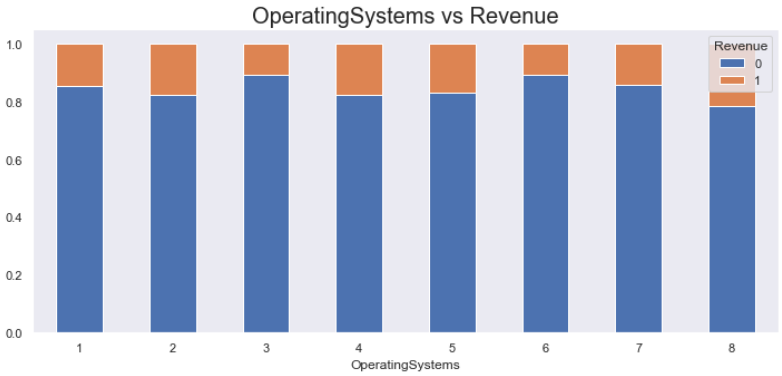
- Month(該筆資料觀看月份):11 月的購物機率最高,2 月購物機率最低,5 月的觀看數量最多。有購物的部分,11、5、12 月最多筆購物,2 月最少。



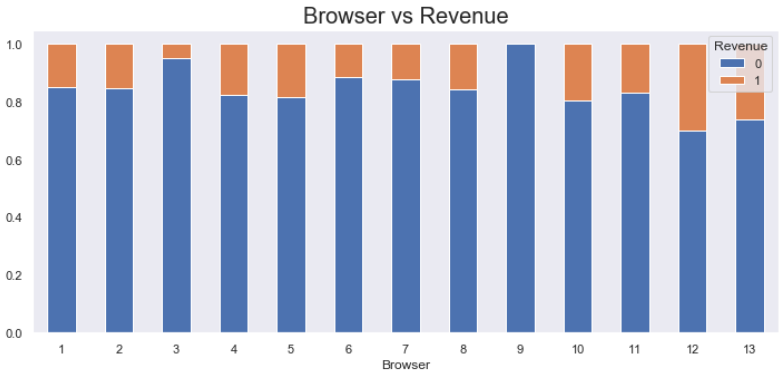




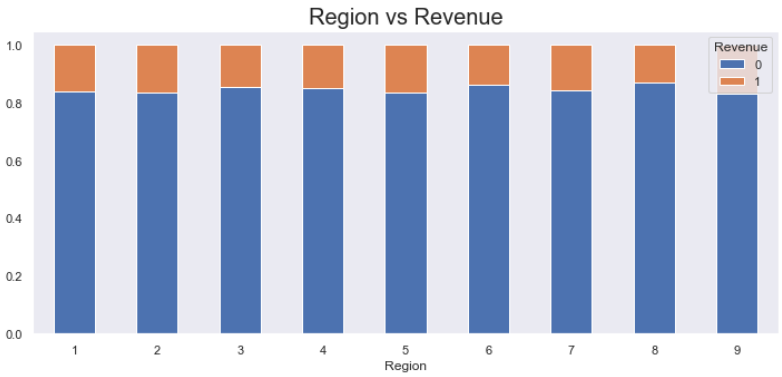
- OperationSystems(該筆資料使用之作業系統)：作業系統與購物機率影響不大。



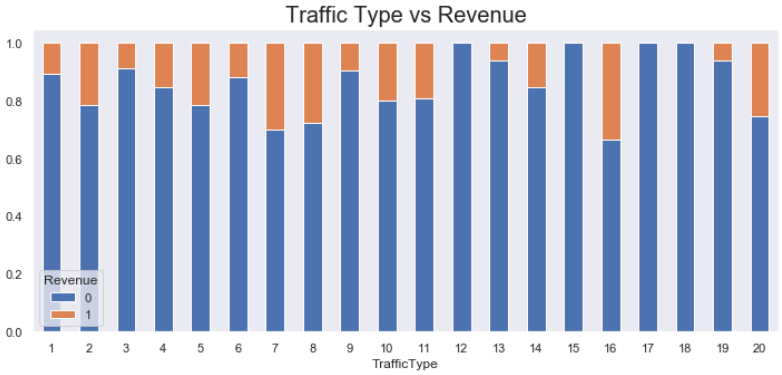
- Browser(該筆資料使用之瀏覽器):瀏覽器在第 9 種瀏覽器的購物機率最低，但是只有一筆資料，無法評估。



- Region(該筆資料觀看者之區域)：每個區域的購物機率差不多。

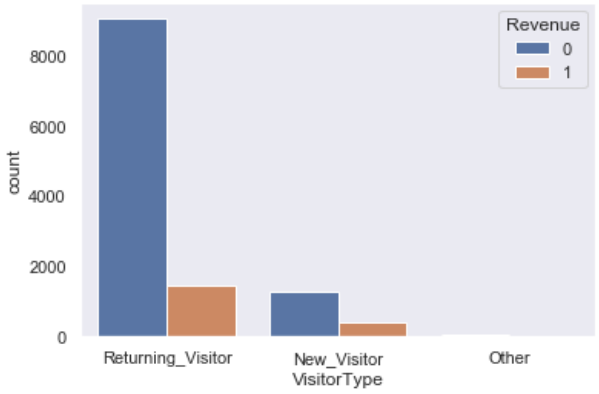


- TrafficType(該筆資料觀看者交通類別)：在第 12、15、17、18 類的購物機率最低。



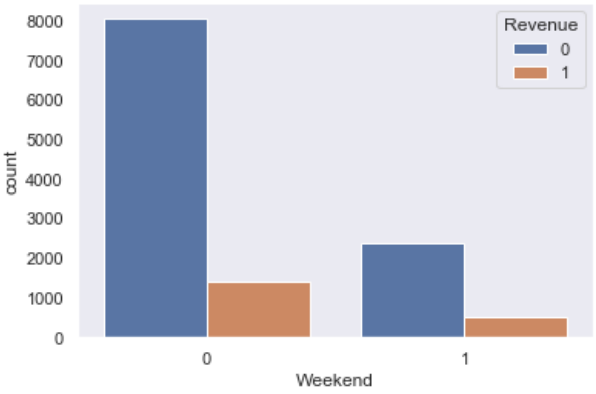
- VisitorType(該筆資料觀看者種類)：新顧客的購物機率最高，但是是因為回訪的顧客觀看次數較多的關係，回訪顧客的購物數最高。

	VisitorType	Revenue
0	New_Visitor	0.249
1	Other	0.188
2	Returning_Visitor	0.139

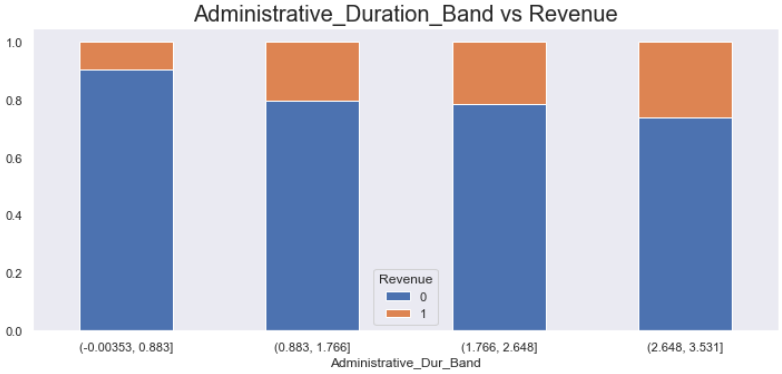


- Weekend(該筆資料是否在周末觀看)：是不是在假日的購物機率差不多。但是非假日的觀看次數較多。

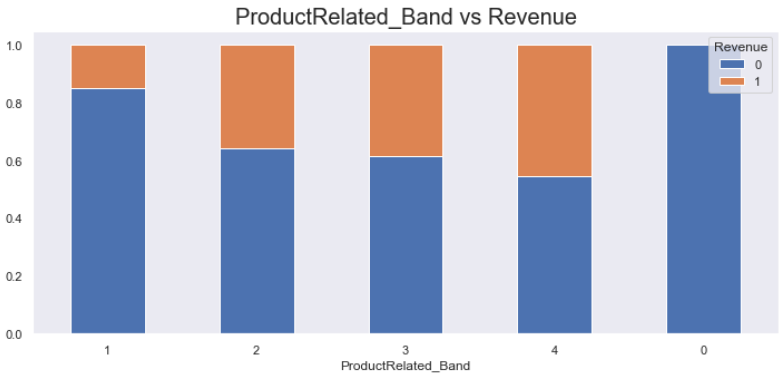
	Weekend	Revenue
0	0	0.149
1	1	0.174



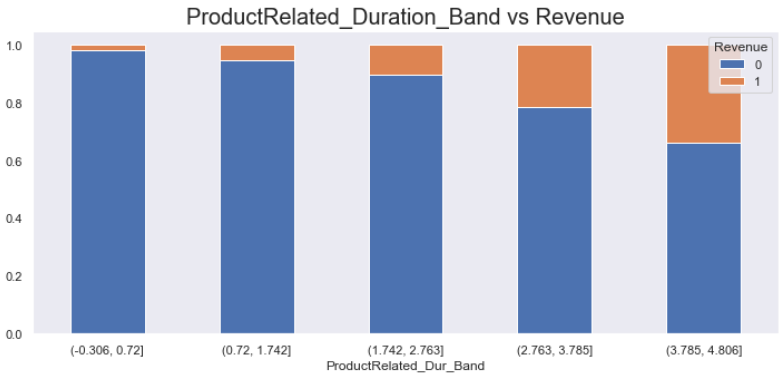
4. 轉換數值
   - Administrative_Dur_Band：由於 Administrative_Duration 範圍大，取 log 後切成 4 個數值區間，發現值越大購買機率越高。再將區間轉換為 0, 1, 2, 3。



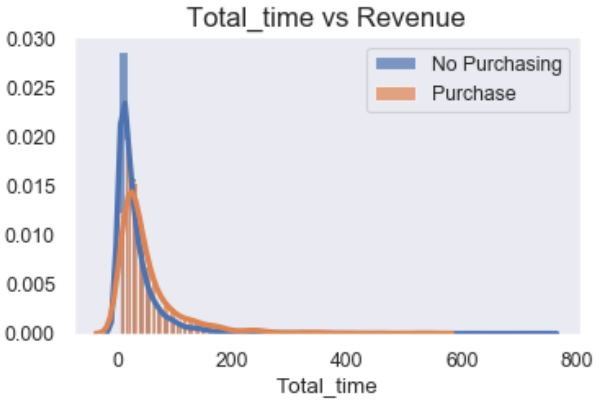
   - ProductRelated_Band：將 ProductRelated 切成 5 個數值區間，發現值越大購買機率越高，但是值在最大的區間時，購買率又為 0。按照購買機率大小將區間轉成 0~4。



- 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。

```
Random Forest Classifier:
  Average train accuracy: 0.9993917274939171
  Average validation accuracy: 0.8283860502838605

Support Vector Machine:
  Average train accuracy: 0.8452960259529604
  Average validation accuracy: 0.8451743714517438

MLP Classifier:
  Average train accuracy: 0.8313868613138686
  Average validation accuracy: 0.8283049472830495
```

2. **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。

3. **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。

```
rd_features = [
            'PageValues', 'ProductRelated', 'ProductRelated_Duration',
            'Administrative', 'BounceRates', 'ExitRates',
            'Month', 'Weekend', 'SpecialDay'
]
```

```
RandomForestClassifier(n_estimators=500,
                       max_depth=15,
                       max_features=9,
                       criterion='entropy',
                     min_weight_fraction_leaf=0.01,
                       random_state=2000)
```

```
Average train accuracy: 0.9111922141119221
Average validation accuracy: 0.905190592051906
```

- SVM：使用下圖特徵與超參數，得到 89.8%驗證準確度。

```
svm_features = [
            'PageValues_Thres', 'ProductRelated_Band', 'ProductRelated_Dur_Band',
            'Administrative', 'Administrative_Dur_Band',
            'Bounce*Exit', 'BounceRates_Thres', 'ExitRates_Thres',
            'Month_2', 'Month_11',
            'Traffic_noRevenue',
            'Returning_Visitor', 'New_Visitor',
            'Weekend', 'SpecialDay'
]
```

```
SVC(C=1, kernel='rbf', gamma=0.4, random_state=2000)
Average train accuracy: 0.9046634225466341
Average validation accuracy: 0.8982968369829685
```

- MLP：使用下圖特徵與超參數，得到 89.8%驗證準確度。

```
mlp_features = [
            'PageValues_Thres', 'ProductRelated_Band', 'ProductRelated_Dur_Band',
            'Administrative', 'Administrative_Dur_Band',
            'Bounce*Exit', 'BounceRates_Thres', 'ExitRates_Thres',
            'Month_2', 'Month_11',
            'Traffic_noRevenue',
            'Returning_Visitor', 'New_Visitor',
            'Weekend', 'SpecialDay'
]
```

```
MLPClassifier(
 hidden_layer_sizes=(110,),
 activation='relu',
 solver='adam',
 learning_rate='adaptive',
 learning_rate_init=0.001,
 max_iter=500, random_state=2000)
Average train accuracy: 0.9021289537712895
Average validation accuracy: 0.8977291159772911
```

4. Improvement：

一、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。

```
Random Forest Classifier:
  Average train accuracy: 0.9998580697485806
  Average validation accuracy: 0.9022708840227087

Support Vector Machine:
  Average train accuracy: 0.8480332522303324
  Average validation accuracy: 0.8472019464720196

MLP Classifier:
  Average train accuracy: 0.8798864557988646
  Average validation accuracy: 0.8783454987834549
```

二、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。

```
Random Forest Classifier:
  Average train accuracy: 0.9220802919708028
  Average validation accuracy: 0.8823195458231956

Support Vector Machine:
  Average train accuracy: 0.8928629359286294
  Average validation accuracy: 0.8927818329278183

MLP Classifier:
  Average train accuracy: 0.896897810218978
  Average validation accuracy: 0.8932684509326846
```

三、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。

```
Random Forest Classifier:
  Average train accuracy: 0.9714517437145174
  Average validation accuracy: 0.8828872668288726

Support Vector Machine:
  Average train accuracy: 0.8936334144363342
  Average validation accuracy: 0.8927818329278183

MLP Classifier:
  Average train accuracy: 0.9011354420113543
  Average validation accuracy: 0.897566909975669
```

四、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。

```
Random Forest Classifier:
  Average train accuracy: 0.9111922141119221
  Average validation accuracy: 0.905190592051906
```

```
Support Vector Machine:
  Average train accuracy: 0.9046634225466341
  Average validation accuracy: 0.8982968369829685
```

```
MLP Classifier:
  Average train accuracy: 0.9021289537712895
  Average validation accuracy: 0.8977291159772911
```