FDA_HW3_2 Online Shoppers Purchasing Intention

分類問題-藉由獲得之資料預測網路 Shoppers 是否會購買商品

交管 109 劉冠廷 學號：H54051261

- Analyze the data:.
  - Data information:

```
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
```

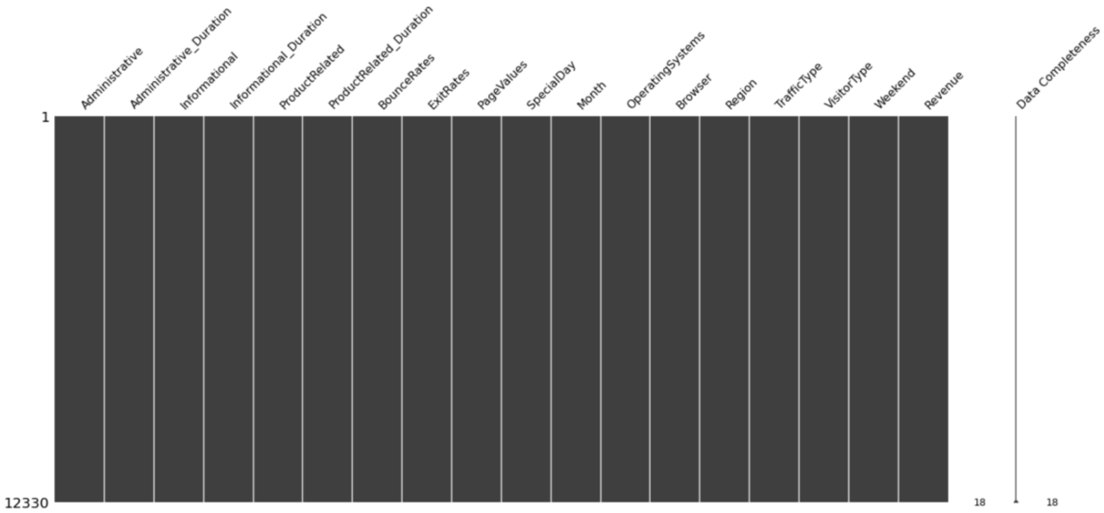
  - Correlation:

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration
Administrative	1.000000	0.601583	0.376850	0.255848	0.431119	0.373939
Administrative_Duration	0.601583	1.000000	0.302710	0.238031	0.289087	0.355422
Informational	0.376850	0.302710	1.000000	0.618955	0.374164	0.387505
Informational_Duration	0.255848	0.238031	0.618955	1.000000	0.280046	0.347364
ProductRelated	0.431119	0.289087	0.374164	0.280046	1.000000	0.860927
ProductRelated_Duration	0.373939	0.355422	0.387505	0.347364	0.860927	1.000000
BounceRates	-0.223563	-0.144170	-0.116114	-0.074067	-0.204578	-0.184541
ExitRates	-0.316483	-0.205798	-0.163666	-0.105276	-0.292526	-0.251984
PageValues	0.098990	0.067608	0.048632	0.030861	0.056282	0.052823
SpecialDay	-0.094778	-0.073304	-0.048219	-0.030577	-0.023958	-0.036380
Month	0.048560	0.029061	0.019743	0.005987	0.070299	0.061186
OperatingSystems	-0.006347	-0.007343	-0.009527	-0.009579	0.004290	0.002976
Browser	-0.025035	-0.015392	-0.038235	-0.019285	-0.013146	-0.007380
Region	-0.005487	-0.005561	-0.029169	-0.027144	-0.038122	-0.033091
TrafficType	-0.033561	-0.014376	-0.034491	-0.024675	-0.043064	-0.036377
VisitorType	-0.025820	-0.023940	0.055828	0.044677	0.126656	0.119329
Weekend	0.026417	0.014990	0.035785	0.024078	0.016092	0.007311
Revenue	0.138917	0.093587	0.095200	0.070345	0.158538	0.152373

BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
-0.223563	-0.316483	0.098990	-0.094778	0.048560	-0.006347	-0.025035	-0.005487	-0.033561	-0.025820	0.026417	0.138917
-0.144170	-0.205798	0.067608	-0.073304	0.029061	-0.007343	-0.015392	-0.005561	-0.014376	-0.023940	0.014990	0.093587
-0.116114	-0.163666	0.048632	-0.048219	0.019743	-0.009527	-0.038235	-0.029169	-0.034491	0.055828	0.035785	0.095200
-0.074067	-0.105276	0.030861	-0.030577	0.005987	-0.009579	-0.019285	-0.027144	-0.024675	0.044677	0.024078	0.070345
-0.204578	-0.292526	0.056282	-0.023958	0.070299	0.004290	-0.013146	-0.038122	-0.043064	0.126656	0.016092	0.158538
-0.184541	-0.251984	0.052823	-0.036380	0.061186	0.002976	-0.007380	-0.033091	-0.036377	0.119329	0.007311	0.152373
1.000000	0.913004	-0.119386	0.072702	-0.023763	0.023823	-0.015772	-0.006485	0.078286	0.135536	-0.046514	-0.150673
0.913004	1.000000	-0.174498	0.102242	-0.039049	0.014567	-0.004442	-0.008907	0.078616	0.179144	-0.062587	-0.207071
-0.119386	-0.174498	1.000000	-0.063541	0.021780	0.018508	0.045592	0.011315	0.012532	-0.111228	0.012002	0.492569
0.072702	0.102242	-0.063541	1.000000	0.079341	0.012652	0.003499	-0.016098	0.052301	0.085557	-0.016767	-0.082305
-0.023763	-0.039049	0.021780	0.079341	1.000000	-0.029580	-0.045913	-0.032530	0.041839	0.026481	0.029132	0.080150
0.023823	0.014567	0.018508	0.012652	-0.029580	1.000000	0.223013	0.076775	0.189154	0.001504	0.000284	-0.014668
-0.015772	-0.004442	0.045592	0.003499	-0.045913	0.223013	1.000000	0.097393	0.111938	-0.021867	-0.040261	0.023984
-0.006485	-0.008907	0.011315	-0.016098	-0.032530	0.076775	0.097393	1.000000	0.047520	-0.036191	-0.000691	-0.011595
0.078286	0.078616	0.012532	0.052301	0.041839	0.189154	0.111938	0.047520	1.000000	-0.002839	-0.002221	-0.005113
0.135536	0.179144	-0.111228	0.085557	0.026481	0.001504	-0.021867	-0.036191	-0.002839	1.000000	-0.043679	-0.104726
-0.046514	-0.062587	0.012002	-0.016767	0.029132	0.000284	-0.040261	-0.000691	-0.002221	-0.043679	1.000000	0.029295
-0.150673	-0.207071	0.492569	-0.082305	0.080150	-0.014668	0.023984	-0.011595	-0.005113	-0.104726	0.029295	1.000000

- How did you preprocess this dataset？

1. 首利用 info()和 msno.matrix()函式得知 train data 裡的資料筆數及型態，每個欄位皆有 12330 筆的資料，並無缺失，因此不須填補資料。



2. 再來，由於 Month,VisitorType,Weeken,Revenue 的資料型態前兩者為 object 後兩者為 boolean，因此利用 LabelEncoder()給予類別數值資料。

```
le = LabelEncoder()

for i in ('Month', 'VisitorType', 'Weekend', 'Revenue'):
  df[i] = le.fit_transform(df[i])
```

3. 利用 corr()函式得出各資料欄位間的相關係數，可觀察出 ExitRates 和 BounceRates 之相關係數高達 0.91，因此排除 ExitRates 資料欄位。

```
df.corr()
```

4. 將 Revenue 欄位的值取出，並設為答案資料。

```
data_y = df['Revenue']
```

5. 由於 TraffcType 在顧客尚未訂購的狀況下，無法得知其運送方式，因此不應列為幫助預測之變數。而 Browser 則是主觀認為現代人所使用的瀏覽器幾乎大同小異並不會造成太大影響，在第一次實驗中先將 Browser 資料刪除，因此將上述兩項一同從預測變數中排除，加上 Revenue 答案欄位共計排除四個資料欄位。在第二次實驗才將 Browser 資料重新加入預測變數，觀察其變化。

```
data_x = df.drop(['Revenue', 'ExitRates', 'TrafficType', 'Browser'], axis=1)
```

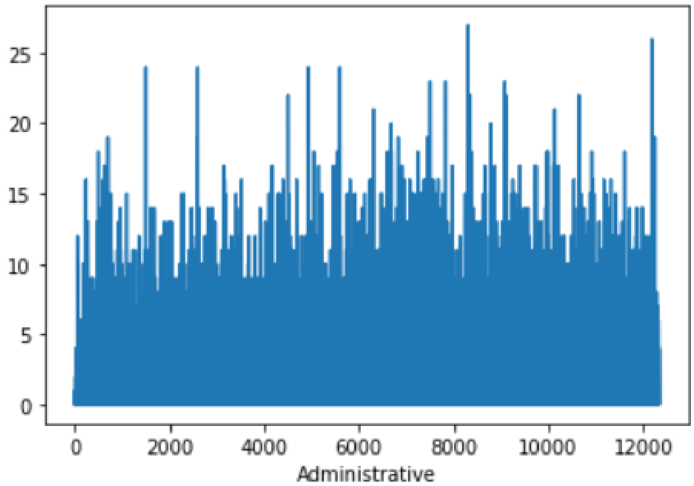
6. 將資料分為 80%的 train set 和 20%的 test set

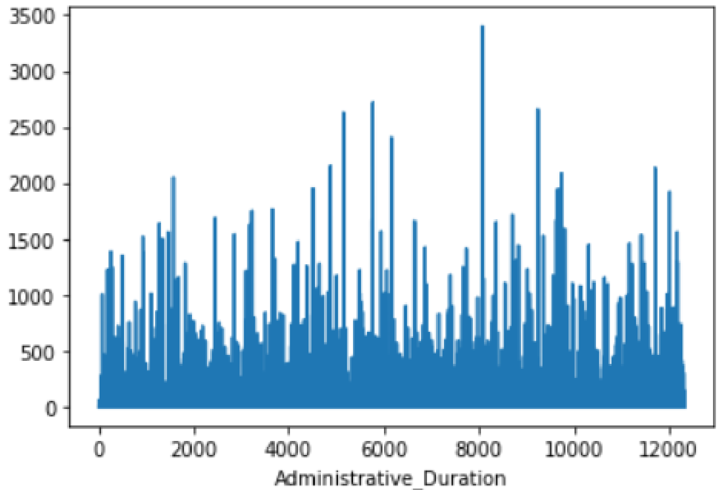
```
       Administrative  Administrative_Duration  ...  VisitorType  Weekend
8063                0                 0.000000  ...            0        0
3334                2                98.000000  ...            2        0
1769                1                14.000000  ...            2        0
10020               1                 0.000000  ...            2        0
9031                9               189.109848  ...            2        1
...               ...                      ...  ...          ...      ...
4426                0                 0.000000  ...            2        0
8593               12               651.875000  ...            2        0
3525                0                 0.000000  ...            2        0
10214               0                 0.000000  ...            2        0
7510                3                40.200000  ...            2        1

[9864 rows x 14 columns]
```

```
       Administrative  Administrative_Duration  ...  VisitorType  Weekend
7888                2               108.800000  ...            2        0
658                 0                 0.000000  ...            2        0
3237                1               260.000000  ...            0        0
11924               0                 0.000000  ...            0        0
6505                0                 0.000000  ...            0        0
...               ...                      ...  ...          ...      ...
5351                0                 0.000000  ...            2        1
695                 5               100.916667  ...            2        1
10774               8               146.000000  ...            2        0
5434                9               194.416667  ...            2        0
5560                0                 0.000000  ...            2        0

[2466 rows x 14 columns]
```

```
8063     1                              7888     0
3334     0                              658      0
1769     0                              3237     0
10020    0                              11924    1
9031     1                              6505     1
        ..                                      ..
4426     0                              5351     0
8593     0                              695      1
3525     1                              10774    0
10214    0                              5434     0
7510     0                              5560     1
Name: Revenue, Length: 9864, dtype: int64 Name: Revenue, Length: 2466, dtype: int64
```

7. 畫出各資料欄位的長條圖





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

1. Original result

   利用羅吉斯回歸、隨機森林、簡單貝氏和神經網路進行模型的訓練，預測準確度皆高於 0.85，其中又以隨機森林之預測準確率最高。

```
logistic regression average train accuracy: 0.8816402532914406
                        min train accuracy: 0.8802433151691801
                        max train accuracy: 0.8838063862138875
logistic regression average valid accuracy: 0.8820953048713414
                        min valid accuracy: 0.8701825557809331
                        max valid accuracy: 0.8920425747592499
logistic regression test accuracy: 0.884022708840227

random forest average train accuracy: 0.999974654669877
                  min train accuracy: 0.9998732733493854
                  max train accuracy: 1.0
random forest average valid accuracy: 0.898620370951044
                  min valid accuracy: 0.8899594320486816
                  max valid accuracy: 0.9072478459199189
random forest test accuracy: 0.9087591240875912

naive bayes average train accuracy: 0.8533049192870704
                min train accuracy: 0.8486883791661386
                max train accuracy: 0.8567988848054746
naive bayes average valid accuracy: 0.8533030084641648
                min valid accuracy: 0.8341784989858012
                max valid accuracy: 0.8677141409021795
naive bayes test accuracy: 0.8536090835360909

     neural network test accuracy: 0.894566098945661
```

2. Reasons

   一開始因為主觀認為 Browser 這項資料不太重要因此進行排除，發現所預測之結果準確率還算不錯，因此想要知道若將 Browser 資料也加入預測變數會有什麼效果，準確率是否會有一定的提升。

3. Your approaches

   所利用的方法為將一開始刪除的資料集 Browser 資料欄位重新加入預測變數，並再次透過羅吉斯回歸、隨機森林、簡單貝氏和神經網路進行模型的訓練，並以訓練好的模型預測結果，以此觀察 Browser 欄位對於整體預測結果之影響程度。

4. Improvement

   經過實驗結果觀察後，發現除了隨機森林以及神經網路所訓練出的模型外，羅吉斯回歸和簡單貝氏所產生之結果並無變化。隨機森林之預測測試準確度從 0.90875 上升至 0.91281，微幅上升。然而，神經網路之預測測試準確度卻從 0.89457 下降至 0.88888。因此推測 Browser 資料欄位對於預測效果雖有部分提升但幫助實在有限，不是個良好的預測變數。

```
logistic regression average train accuracy: 0.8816149143843612
                        min train accuracy: 0.880623495121024
                        max train accuracy: 0.8835529650278763
logistic regression average valid accuracy: 0.8818925165186405
                        min valid accuracy: 0.8696754563894523
                        max valid accuracy: 0.8915357323872276
logistic regression test accuracy: 0.8844282238442822

random forest average train accuracy: 0.999974654669877
                  min train accuracy: 0.9998732733493854
                  max train accuracy: 1.0
random forest average valid accuracy: 0.9017627936575823
                  min valid accuracy: 0.8899594320486816
                  max valid accuracy: 0.9102889001520527
random forest test accuracy: 0.9128142741281428

naive bayes average train accuracy: 0.8515561621620685
                min train accuracy: 0.8455202129007731
                max train accuracy: 0.8548979850462552
naive bayes average valid accuracy: 0.8505658540396777
                min valid accuracy: 0.8321501014198783
                max valid accuracy: 0.8626457171819564
naive bayes test accuracy: 0.8536090835360909

     neural network test accuracy: 0.8888888888888888
```