主題為online shopper purchasing intention

問題定義:來預測該瀏覽者最終是否會購買此商品，選擇此問題是因為可以透過每個人的瀏覽習性，讓高機率會購買商品的人，提供更多的廣告或是優惠，使他更容易產生購買的行動以此增加營收。

Original result

先將數值型態的變數取出作為線性回歸的建模與預測，由結果可知若是資料並無預處理且用簡單線性回歸，則準確率只有大概三成。

Reasons

1.有其他類別型態的變數並無考慮進去，可能會有underfitting的情形

2.像是OperatingSystems,Browser,Region,TrafficType這些變數雖然是數值型態，不過實際上是類別，若是按造原先的預設數值去進行回歸分析則會有解釋上的問題EX:假設windows代表1 ios代表2 無法說明為什麼ios的值意義上是windows的兩倍，其他的類別尺度也是有這樣的錯誤。

3.共線性問題，若兩個變數擁有高度相關，那麼在做線性回歸時就會造成係數差異會很大，導致結果不準確。

4.線性回歸主要是用來針對連續值做處理，而我們的預測資料是離散的因此在迴歸分析上準確率會很差。

My approaches

0.確認是否有NA，填補遺漏項(不過該資料是齊全的)

1.將類別屬性的資料一併考慮進去(解決underfitting)

2.將類別尺度的變數轉成dummy variable才不會有解釋上的問題

3.將高度相關的變數進行剃除 EX:ExitRates與BounceRates只保留一個

4.改使用對於分類具有較高準確度的模型EX:logstic or SVM

5.通常在預處理的時候還會有outlier的剃除，不過這裡我並無使用，原因是有outlier的地方可能是在value那邊，但極高value反而是代表使用者的資訊而不是錯誤的輸入之類的，因此更應該保留這些數值，使其變數有顯著得影響。

Improvement 27% -> 89%

Another approach may improve more

觀察資料可以發現，有些人在瀏覽網站時停留時間極短甚至是0，猜想可能是那些人無意點開這類別網站但實質上不具有興趣，因此很快地就離開或是關掉，因此這類的資料可以某種程度上算成是比較無資訊量的資料，因此把這些務處的資料進行刪除，有可能可以使得最終的預測更加準確。