學號：B06502149 系級：資工二 姓名：張琦琛

1. 請比較你實作的 generative model、logistic regression 的準確率，何者較佳？

   表格的數值為 Before normalization / After normalization

	Public	Private	Average
Generative	0.84656/0.84656	0.84092/0.84092	0.84656/0.84092
Logistic	0.78869/0.85245	0.79695/0.84964	0.79282/0.85105

   由表格可以看出在標準化之後的 logistic model 大幅提升了準確率和效能，得到了較好的分數，但是 generative model 在標準化後的分數卻沒有什麼改變，因此我認為雖然 generative model 的參數容易因為資料偏移的影響而改變，但是在同一筆資料下 generative model 受到標準化的影響卻微乎其微，準確率並沒有任何改變。整體來說，如果要達到較好的準確率，logistic model 會比 generative model 更靈活，較能透過調整訓練參數、標準化等方法逼近最佳解。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

   在這次的 best model 中我用了 sklearn 中的 DecisionTreeClassifier 當作 weak learner，實作 Adaboost，在 public 和 private 的準確率分別為 0.86363 / 0.85947。Adaboost 是 Adaptive boosting 的縮寫，而他實際上是提高被前幾個分類器線性組合的分類錯誤樣本的權重，這樣做可以讓每次訓練新的分類器的時後都聚焦在容易分類錯誤的訓練樣本上。每個弱分類器使用加權投票機制取代平均投票機制，而準確率較大的弱分類器有較大的權重，反之，準確率低的弱分類器權重較低 ，這種作法也較不容易 Overfitting。

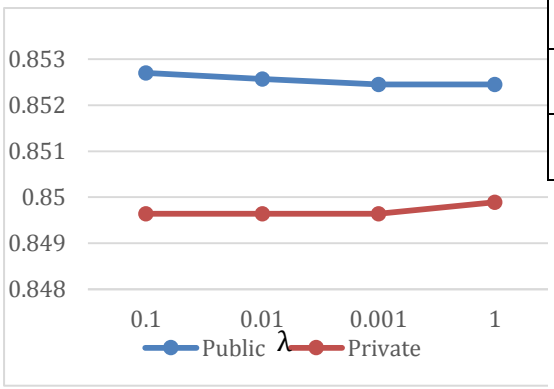
3. 請實作輸入特徵標準化(feature normalization)並討論其對於你的模型準確率的影響

   表格的數值為 Before normalization / After normalization

	Public	Private	Average
Generative	0.84656/0.84656	0.84092/0.84092	0.84374/0.84374
Logistic	0.78869/0.85245	0.79695/0.84964	0.79282/0.85105
AdaBoost	0.86363/0.86363	0.85947/0.85947	0.86115/0.86115

標準化只對 Logistic model 有較大的影響，不僅提升準確率，在 Training 的時候也較不容易 Overflow，Generative model 則沒有太大差異，而 AdaBoost 不受 Normalization 影響是因為 我是利用 DecisionTree 實作，而 DecisionTree 並不容易受 Normalization 影響。

4. 請實作 **logistic regression** 的正規化**(regularization),** 並討論其對於你的模型準確率的影響。



λ	0.1	0.01	0.001	1
Public	0.85270	0.85257	0.85245	0.85245
Private	0.84964	0.84964	0.84964	0.84989

在這邊我們可以發現隨著 λ 提高，Private 的準確率會有些許提升，Public 的會有些許下降，但是幅度並不高，我認為是因為 Logistic model 並不複雜，沒有嚴重的 Overfitting 情況，因此 regularization 在此 Dataset 並沒有得到很好的進步。

5. 請討論你認為哪個 **attribute** 對結果影響最大**?**

在將各個 **attribute** 拔掉後發現，**Captial gain+Capital loss** 的影響最大

Removed	Public	Private
None	0.85245	0.84964
age	0.85135	0.85173
fnlwgt	0.85221	0.85136
sex	0.85196	0.85284
Capital	0.83710	0.83294
country	0.85344	0.85143
Work	0.85106	0.85182
Martial	0.85329	0.85105