學號：R04546031 系級： 工工碩二 姓名：洪唯凱

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

答：

	Public Score	Private Score
generative	0.84557	0.84240
logistic	0.78832	0.78675

根據上述表格，generative model 在 private 和 public 都贏過 logistic

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

我使用的是 xgboost，準確率為 0.86413(public 和 private 的平均)

由於時間的不足，其實沒有什麼訓練方式，我只是把 xgboost 整套搬進來，丟上去 kaggle 就過了!之後可能會開始著墨他的一些參數調整的變化!

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。

答：

我用 generative model 來進行說明，

	Public Score	Private Score
有做標準化	0.84557	0.84240
沒做標準化	0.76523.	0.76231

沒有做標準化的 model，精準度明顯小於有做的 model，並且在打開 output 的 file 之後，發現裡面全部都是 0，再做進一步的敘述統計。可以發現變數之間差距很大!

	age	fnlwgt	sex	capital_gain	capital_loss	hours_per_week	Federal-gov	Local-gov	3
count	32561.000000	3.256100e+04	32561.000000	32561.000000	32561.000000	32561.000000	32561.000000	32561.000000	
mean	38.581647	1.897784e+05	0.669205	1077.648844	87.303830	40.437456	0.029483	0.064279	
std	13.640433	1.055500e+05	0.470506	7385.292085	402.960219	12.347429	0.169159	0.245254	
min	17.000000	1.228500e+04	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	
25%	28.000000	1.178270e+05	0.000000	0.000000	0.000000	40.000000	0.000000	0.000000	
50%	37.000000	1.783560e+05	1.000000	0.000000	0.000000	40.000000	0.000000	0.000000	
75%	48.000000	2.370510e+05	1.000000	0.000000	0.000000	45.000000	0.000000	0.000000	
max	90.000000	1.484705e+06	1.000000	99999.000000	4356.000000	99.000000	1.000000	1.000000	

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。

答：

	Public Score	Private Score
regularization	0.80663	0.80444
No regularization	0.78832	0.78675

Lamba 值 = 0.1，雖然都還沒有 overfitting 的程度，但是我們可以發現比較 smoother 的線，預測起來的準確率，都大於沒有做正規化。

5.請討論你認為哪個 attribute 對結果影響最大？

我以 logistic regression 作為舉例，要看哪一個 attribute 影響最大，我的想法是，在最後 train 完所得到的 106 個 feature 中，我取絕對值最大的權重，當成是我 attribute 最大的因子，所以我的答案是 Hungary