學號：R07942091 系級：電信碩一 姓名：許博閔

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

	Public score	Private score
generative model	0.84152	0.83920
logistic regression	0.85270	0.85161

無論是 kaggle 的 public 或 private，都是 logistic regression 的準確率較高，都大約高 0.01%

2. 請說明你實作的 best model，其訓練方式和準確率為何？

我用了 sklearn 的 GradientBoostingClassifier，參數的設置如下:

loss = 'deviance', n_estimators = 150 , max_depth = 5, random_state = 0

我不確定 windows 和 linux 的 random_state = 0 會不會造成不同結果，但 best.py 直接讀我存好的 model.pickle 來作預測，應該沒問題。

準確率 : kaggle public : 0.87776，kaggle private : 0.87483

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

	With normalization	Without normalization
generative model	0.84152 / 0.83920	0.81867 / 0.81034
logistic regression	0.85270 / 0.85161	0.77383 / 0.77607

可以看到 normalization 對模型的準確率有很大的提升，尤其是 logistic regression，多做了 normalization 就可以過 kaggle 的 simple baseline

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。

Regularization constant	public / private accuracy
0.0	0.85270 / 0.85161
0.001	0.85233 / 0.84780
0.0001	0.85294 / 0.85112

根據結果，加上正規化對結果的影響不大，我認為比較可能的原因是 model 並不算

複雜，因此沒有發生 overfitting，所以正規化的影響不大。

5. 請討論你認為哪個 attribute 對結果影響最大？

Drop attribute	public / private accuracy
None	0.85270 / 0.85161
age	0.85147 / 0.85161
fnlwgt	0.85442 / 0.85186
sex	0.85221/ 0.85186
capital_gain	**0.84078/ 0.83404**
capital_loss	0.85245 / 0.84817
hours_per_week	0.85331 / 0.84903
workclass	0.85208 / 0.84842
education	**0.84557 / 0.84043**
marital_status	0.85221 / 0.85173
occupation	0.84643 / 0.84412
relationship	0.85319 / 0.85100
race	0.85331 / 0.85124
Native country	0.85270 / 0.85149

以上圖表是 logistic regression 分別 drop 掉特定 attribute 的準確率，有比較明顯下降的分別是 capital_gain 和 education，而 capital_gain 下降最多，少了大約 0.01%，因此我認為 capital_gain 對結果的影響最大。