學號：R06921005 系級： 電機碩一 姓名：陳昱文

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

Accuracy	public	private	total
generative	0.79729	0.79498	0.79614
Logistic	0.85393	0.85124	0.85259

Logistic 較 generative 為佳

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：

Accuracy	public	private	total
Best	0.86805	0.86610	0.86708

此 Best model 使用了 xgboost 套件的 XGBClassifier()
若實際上手刻的二階 gradient 亦可達到 0.8646 故可見在稍微簡單的 task 使用強力套件的效果較不顯著

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：

Accuracy	public	private	total
w normalize	0.85393	0.85124	0.85259
w/o normalize	NA	NA	NA

由於 feature normalization 使得值介於 0~1 (使用 max normalization)
若未經由此步驟 在計算過程中會出現 overflow 的現象
而後使用助教 clip 的方式 其值於兩端點震盪 並無 training 效果

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：

Accuracy	public	private	total
Lambda=0.01	0.85393	0.85124	0.85260
Lambda=0.1	0.85466	0.85124	0.85295
Lambda=1	0.85196	0.84645	0.84921
Lambda=10	0.84471	0.84166	0.84319

Regularization 目的是在於使其平滑化 避免 overfitting 然而過大的值則會影響到的最佳值的收斂 如表中可見當 regularization 過大時 其 performance 不論 public 或 private 都變差

5.請討論你認為哪個 attribute 對結果影響最大？

Accuracy	public	private	Total
w/o age	0.85233	0.85112	0.85173

w/o fnlwgt	0.85356	0.85014	0.85185
w/o sex	0.85442	0.85112	0.85277
w/o capital_gain	0.83894	0.83589	0.83742
w/o capital_loss	0.85073	0.84829	0.84951
w/o hours_per_week	0.85319	0.84928	0.85124
w/o workclass	0.85110	0.85063	0.85087
w/o education_num	0.85368	0.85063	0.85216
w/o education	0.85049	0.84215	0.84632
w/o marital_status	0.85393	0.85087	0.85240
w/o occupation	0.84877	0.84621	0.84749
w/o relationship	0.85565	0.84903	0.85234
w/o race	0.85331	0.85161	0.85246
w/o country	0.85442	0.85149	0.85296

由於無法檢測所有的可能性 我們已同要缺少特定一種 feature 為基準比較 看少了哪一個 attribute 會使 performance 變差最多

從表中可見 capital_gain 的移除會使得準確率下降幅度最大 此與真實生活情況是正相關的 而第二名則是 occupation 亦符合直覺