學號：R07943107 系級：電子所碩一 姓名：徐晨皓

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

	Private score	Public score
Generative model	0.84080	0.84643
Logistic regression	0.85038	0.85196

在實作中，此兩種方法都有先對每個 feature 做 normalization。以 public score 和 private score 而言，logistic regression 表現較佳。

2. **請說明你實作的 best model，其訓練方式和準確率為何？**

Best model 使用的是 sklearn 的 RandomForestClassifier(n_estimators=500, max_depth=20)，並且對於所有的 feature 都有做 normalization。Public score 為 **0.86486**；private score 為 **0.86353**。

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。**

	Private score	Public score
w/ normalization	0.85038	0.85196
w/o normalization	0.77877	0.78108

實驗設定為 logistic regression, iterations=1000, learning rate=0.5, lambda=0.001。若沒有使用 feature normalization，模型的準確率會大幅下降。原因可能是 one-hot 的 feature 為 0 和 1，但其他實數的 features (如 age, fnlwgt 等)是遠大於 1 的。這樣會使得某些 features dominate 整個模型。

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	Private score	Public score
0	0.85063	0.85110
0.00001	0.85075	0.85110
0.0001	0.85050	0.85098
0.001	0.85038	0.85196
0.01	0.85014	0.85085
0.1	0.84043	0.84004
1	0.76771	0.77272
10	0.76047	0.76707
100	0.76047	0.76707

從實驗中，我們可以發現當 λ 小於等於 0.1 的時候，對準確率沒有太大的影響。但當 λ 大於等於 1 的時候，準確率就大幅下降。

5. **請討論你認為哪個 attribute 對結果影響最大？**

在訓練後的 logistic regression 模型的 weights 中，capital_gain 的係數(1.958)最大，而 Never-married 的係數(-0.4198)最小。因此，我認為 capital_gain 與 Never-married 對結果影響最大。這也還算合理，資本收入(capital_gain)本來就應該與個人的財產收入有很大的關係，而有無婚姻(Never-married)可能某種程度上表示個人在社會上的成熟度。