學號：R05546014 系級： 工工所碩二 姓名：洪紹綺

準確率計算方式:Kaggle Public、Private 平均準確率

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

答：

參數設定:Epoch = 5000, Learning_Rate = 0.1，無 Normalization；

無平行運算(batch_size = 資料量)

Model	準確率
Generative Model	(0.76231+0.76523)/2=0.76377
Logistic Regression	(0.78909+0.79385)/2=0.791469

無特徵標準化的情況下，Logistic Regression 準確率較 Generative Model 佳。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

Best model 使用 xgBoost 方法，經過 cross validation 後選出一組準確率最高的參數設定:max_depth = 6, nthread = 4, learning_rate = 0.1, n_estimators = 1000, gamma = 1, objective= 'binary:logistic' ；其 Public、Private 平均準確率(0.87163+0.87567)/2=0.87365

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：

參數設定:Epoch = 5000, Learning_Rate = 0.1，有 Normalization；

無平行運算(batch_size = 資料量)

Model	準確率
Generative Model	(0.84166+0.84484)/2=0.84325
Logistic Regression	(0.84584+0.84938)/2=0.84761

加上特徵標準化後，兩模型的準確率大幅提升；Logistic Regression 仍有較高準確率。

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：

參數設定:Epoch = 5000, Learning_Rate = 0.1，有 Normalization；

無平行運算(batch_size = 資料量)

Logistic Regression	準確率
Lambda=0	(0.84584+0.84938)/2=0.84761
Lambda=0.001	(0.84645+0.85024)/2=0.848345
Lambda=0.01	(0.84301+0.84545)/2=0.84423
Lambda=0.1	(0.82274+0.82764)/2=0.82519

從 Lambda = 0.001 時可以看出模型準確率略為提升，但模型準確率隨 Lambda 增加而降低。

5.請討論你認為哪個 attribute 對結果影響最大？

在 Best Model 內使用 xgBoost 的 Feature_Importance 功能，列出特徵重要性，取前 10 名如附表。重要性最高的為 fnlwgt，高出第二名的特徵 age 年紀許多，fnlwgt 為 Final Analysis Weights，具有相同背景的人的 fnlwgt 應類似，合理推測背景及年紀對年收入影響大，現實中也是如此；就模型結果: fnlwgt 此項特徵對結果影響最大，其次影響較大的特徵為年紀。

Feature	Importance
fnlwgt	0.228284
age	0.158770
hours_per_week	0.089974
capital_gain	0.070591
capital_loss	0.050849
Married-civ-spouse	0.016391
Prof-specialty	0.015554
Private	0.015075
Bachelors	0.015075
Exec-managerial	0.014836