學號：B06902052 系級：資工二 姓名：張集貴

以下準確率為 Kaggle 上的 Private Score。

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

   generative: 0.84092; logistic: 0.85345

   logistic 較佳。

2. 請說明你實作的best model，其訓練方式和準確率為何？

   使用 scikit-learn 的 GradientBoostingClassifier，訓練的參數為 (n_estimators=500, max_leaf_nodes=128)，其餘皆為 default 值。

   準確率為 0.87605。

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

	generative	logistic
without normalization	0.84117	0.78553
with normalization	0.84092	0.85345

   normalization 對 generative 的影響不大。

   而 logistic 則受到很大的影響，做 normalization 之後的準確率明顯的提高，這是因為 gradient descent 收斂的速度會受到 feature 的 scale 的影響。

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	0	100	1000	10000
準確率	0.85345	0.85210	0.84633	0.81844

   λ 越大準確率越低，因為 regularization 會懲罰 weight 絕對值較大的那些 function，使得一些較為準確的 function 不會被挑到。

5. 請討論你認為哪個attribute 對結果影響最大？

我認為 capital_gain 的影響最大。

對 logistic regression with feature normalization 訓練出來的 weight 做排序，找出最大跟最小的十項 weight，可發現 capital_gain 是當中絕對值最大的，而且是第二大的三倍多，所以是影響很顯著的 feature。

而且 capital gain 的意義是 capital asset (股票、債券或不動產等) 價值提高帶來的獲利，跟高所得成正相關是很合理的。

smallest 10

(-0.5956342912779331, ' Preschool')

(-0.5417162496640042, ' Never-married')

(-0.2961218183970013, ' Own-child')

(-0.28013316650109427, ' Priv-house-serv')

(-0.2557373437944402, ' Other-service')

(-0.23044189032019158, ' Divorced')

(-0.22363199113612894, ' 7th-8th')

(-0.19455097252709916, ' 11th')

(-0.18821167737693684, ' 10th')

(-0.17468948184479588, ' Farming-fishing')

largest 10

(2.3579534274959335, 'capital_gain')

(0.7649754808264984, ' Married-civ-spouse')

(0.4054488602621081, 'sex')

(0.3667651632298394, 'hours_per_week')

(0.3480656228461773, 'age')

(0.29953016213532324, ' Bachelors')

(0.260866678217521, 'capital_loss')

(0.25984931784198395, ' Masters')

(0.2582086515696291, ' Wife')

(0.2524484044847201, ' Exec-managerial')