學號：R06944049 系級： 網媒碩一 姓名：黃敬庭

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

logistic model 的效果比較好,單就以 kaggle public test 的結果來看,並做完 normalization 之後, logistic model 的 0.85184 比 generative mode 的 0.84410 來得好, 考慮到 private set 的正確率之後,logistic 的 model 的 0.84854 比 generative 的 0.84110 好

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：

我的 best model 是用 XGboost 的 library 來做的,用的是好幾棵 pruned 的 gradient boosted decision tree,因為 decision tree 的特性所以資料並不需要經過 normalization, validation 的做法是將資料的後 3000 筆當作 validation set, 挑出在 validation set 上表現最佳的模型,最後結果: 1000 棵樹高為 3 的樹, 然後 learning rate 為 0.07, l2 reg 的 weight 為 1,最後在 kaggle 的 public test 上面達到準確率 0.87985,private 上面達到 0.87274

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：

以 logistic 的 model 來講，做 normalization 之前，kaggle public 的準確率是 0.76523,對「0,1,3,4,5」col 做 normalization 之後，得到的 kaggle 準確率是 0.85184, 對於 generative model 來說,做 normalization 之前，kaggle public 的準確率是 0.84410,但對「0,1,3,4,5」col 做完 normalization 之後,準確率反而下降到 0.83309

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：

Reg Weight	Validation Accuracy	Kaggle Public Accuracy
0.1	0.84700	0.85208
0.01	0.84333	0.85184

0.001	0.84433	0.85184
0.0001	0.84433	0.85245

Validation 做法一樣是將後 3000 筆資料作為 validation set ,但是傳 kaggle 的結果是用全部資料去訓練的,最後得出來的結果發現差異並不大

5.請討論你認為哪個 attribute 對結果影響最大？
答：

Capital gain, 將所有的 feature normalize 之後然後訓練出的 logistic model 去找他們的 coefficient,將之排列之後發現 capital gain 正向影響最大 ,緊接在後的還有 marital status(有些是負面影響),education-num, hours per week, age 等