學號：B04902089 系級： 資工三 姓名：林政豪

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

Generative	
Training accuracy	0.8427566720923805
Testing accuracy	0.84475

Logistic	
Training accuracy	0.8522465526243052
Testing accuracy	0.850925

Logistic 的準確率較佳。
注 1:兩者 feature 相同
注 2:Logistic 使用 adagrad

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：

訓練方式: 使用 xgboost 的 XGBClassifier
Feature:使用助教整理的 feature
　　刪除與 output 相關係數較低的參數(index = 105,84,80,78,75,68,66,22)
　　將數值為連續數字的參數(age ,capital gain loss ….)加上次方或是+1 取 log 平方
max_depth = 3, n_estimators = 1700 , learning_rate = 0.04
Training accuracy = 88.45%
Testing accuracy = 87.75%

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：

Testing accuracy	Normalization = True	Normalization = False
Logistic	0.850925	0.794175
Best(xgboost)	0.8775	0.8775

Normalization 可以使 Logistic regression 的準確率上升，且可以在相同訓練次數和初始學習率之下得到更好的模型。
Xgboost 為許多 decision tree 組成，normalization 並不會影響樹的建置。

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：

Logistic	Training accuracy	Testing accuracy
Lamda = 0	0.8522465526243052	0.850925
Lamda = 0.01	0.8521851294493412	0.850869
Lamda = 0.1	0.8519701483369675	0.850316
Lamda = 0.2	0.8517551672245939	0.850439
Lamda = 0.3	0.8515708976997021	0.850377
Lamda = 0.4	0.8515094745247381	0.850070
Lamda = 0.5	0.8512330702374006	0.850009

從上表可以推論，regularization 對於 Logistic regression 的模型準確率沒有顯著效果，甚至會讓預測效果變差，可能是因為 feature 的選擇，使得模型不會 overfitting。

5.請討論你認為哪個 attribute 對結果影響最大？

Logistic	Training accuracy	Test accuracy
Origin	0.8522465526243052	0.850869
Remove age	0.848100488314241	0.848412
Remove workclass	0.8496667792758208	0.847060
Remove fnlwgt	0.8514787629372562	0.850991
Remove education num	0.8256196062774485	0.825686
Remove marital_status	0.8433401922545376	0.840673
Remove occupation	0.8496667792758208	0.850132
Remove relationship	0.8497589140382666	0.848719
Remove race	0.8512637818248825	0.850746
Remove capital gain	0.8314548078990203	0.831030
Remove capital loss	0.8487147200638802	0.845955
Remove hours_per_week	0.8483461810140966	0.849149
Remove native_country	0.850925954362581	0.850316

我認為 capital loss 對結果影響最大。

在分別拔除 attribute 中，Education num(education)的影響最大，但我認為 capital gain 的影響也排名第二，我有試過把各種 continuous 的 attribute 加上次方以及 log 轉換，而其中 capital loss 對 accuracy 的提升有顯著的效果，但 education num 卻沒辦法這樣調整，故我認為 capital gain 對結果影響最大。