學號：B04902021 系級： 資工三 姓名：陳弘梵

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

在同樣經過標準化，並取所有 feature 的情況下，準確率如下:

	generative model	logistic regression
Whole testing data Accuracy	0.8290645537743382	0.8468767274737424

我的 generative 連 public 的 simple baseline 都沒有過，logistic 卻是稍微調一下 learning rate、iteration 等就能過。但是過 strong 的 public 是在取了「連續資料」部分的「ln」之後才通過。因此以下準確率比較皆以 logistic regression 實作。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

我的 best model 是使用 xgboost 套件，當中的 XGBClassifier 模型之後達到，feature 與 logistic 相同，一次方項去除了 race 以及 native country，ln 取 column:0,1,3,4,5，訓練時 learning rate 為 0.07，iteration 為 1300。

Testing accuracy 為 0.8767888950310178。

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。

由 logistic regression model 實作時發現沒有標準化時，再計算 sigmoid 時會跳出: overflow encountered in exp 的問題，iteration 很快就結束。加上 np.clip 後再試，以下採用前 59 項 feature，再加入 1.5 次方開始每 0.5 次方，直到九次方項(col:0,3,4,5)。

	Normalize	No Normalize
Learning rate=0.1 Accuracy	0.8543701246852159	0.7825072170014127
Learning rate=0.01 Accuracy	0.8518518518518519	0.7825072170014127

應該可以說明是其中一些欄位的數字太大所造成，例如 column1，使得在 gradient 的過程中移動太遠而失去逼近 minimum 的效果。

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。

(1) 採用與第二題相同的 feature，而 learning rate=0.1，iteration=4000
(2) 比起第(1)點，加入 1.5 次方開始每 0.5 次方，直到九次方項(col:0,3,4,5)

	Lamda = 0	Lamda = 0.1	Lamda = 0.01
Accuracy (1)	0.8526503286038941	0.8526503286038941	0.8526503286038941
Accuracy (2)	0.8543701246852159	0.8543701246852159	0.8543701246852159

兩個 model 的 accuracy，都呈現出，對分類的結果沒有任何的影響。

5. 請討論你認為哪個 attribute 對結果影響最大？

以下由 Logistic regression 測試，去除特定參數並且只取一次方。

去除的 Feature	Accuracy
Race	0.8467538848965052
Native country	0.8465696210306493
Age	0.8459554081444629
Fnlwgt	0.8468767274737424
Hours per week	0.8453411952582766
Degree	0.8373564277378539
Marriage	0.8469995700509797
Capital gain	0.8328726736686936
Capital loss	0.8433757140224802
Career	0.8431300288680057

可以觀察出，再有 Normalize 的情況下，去除 capital gain 是影響 accuracy 最嚴重的項目，再來就是各項有關學歷的數據。