學號：B06902028　系級：資工二　姓名：林柏劭

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

	Public score	Private score
generative model	0.84154	0.84643
logistic regression	0.85345	0.85171

根據老師影片的敘述，推測是因為這次的 data 夠多，因此 logistic regression 的表現才能超越 generative model。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

Public score	Private score
0.87618	0.87395

根據第 5 探討，我將 5 個 continuous 的 feature 分別加上 2~20 次方以及 sin,cos,tan 項。並使用 GradientBoostingClassifier，經歷多次的試驗後，透過 validation 找出最佳解會是使用 max_depth=3, n_estimators = 1000 這組參數。

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

	Public score	Private score
Without normalization	0.79891	0.80061
With normalization	0.85345	0.85171

可發現有沒有 normalization 對結果有很大的影響。觀察 data 後，推測是因為這次的 data 中同時存在 one hot encoding 以及 continuous 的 feature。兩者數值相差過大，使得最後 train 出來的結果並不能正確反應出各 feature 的重要程度。

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。

	Public score	Private score
λ =0.0001	0.85345	0.85171
λ =0.001	0.85345	0.85171
λ =0.01	0.85345	0.85171
λ =0	0.85345	0.85171
λ =1	0.85333	0.85171
λ =10	0.85370	0.85196
λ =100	0.85210	0.85257
λ =1000	0.84633	0.84852
λ =10000	0.81844	0.81904

只有在 λ =100 時，regularization 有些許幫助，使得 private score 些許上升。不過隨著 λ 的繼續增加，推測其 underfitting，score 因此下降。

5. 請討論你認為哪個 attribute 對結果影響最大？

我使用 logistic regression，並將 106 個 feature 分別去掉，每次使用 105 個 feature 去做。發現 w 只有在去掉 continuous 的 feature 時(分別是第 0,1,3,4,5)會對整體表現有較大的影響，影響如下:

	Public score	Private score
原本的 logistic regression	0.85345	0.85171
去掉的 0 個 feature	0.85173	0.85171
去掉的 1 個 feature	0.85136	0.85233
去掉的 3 個 feature	0.83417	0.84066
去掉的 4 個 feature	0.84780	0.85122
去掉的 5 個 feature	0.84940	0.85307

因此我認為第 3 個 feature，也就是 ”capital gain”，對結果影響最大。