學號：R05525096 系級： 工海碩二 姓名：郭捷

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

答：在使用全部的資訊和其平方作為 feature，並對所有特徵標準化的情況下。LogisticRegression 的 lr 設為 0.1 並且用 adagrad 的優化方式進行優化。再上述條件下，我實作的 logistic regression 的準確率高於 generative model。

	Public Score	Private Score
Generative model	0.84582	0.84019
Logistic Regression	0.85540	0.85149

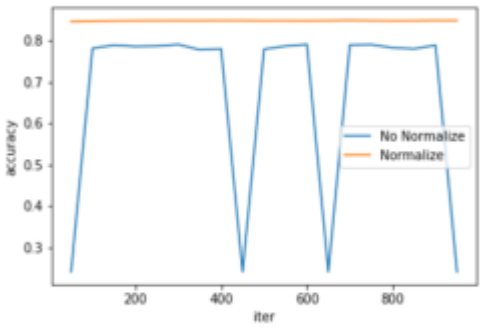
2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

我實作的 best model 是 xgboost 的 XGBClassifier，準確率 Public Score 為 0.87825，Public Score 為 0.87323。xgboost 是屬於 gbm 的一種優化實現，其訓練方式產生很多 CART 樹，在目標函數中加入正則化項，通過優化 GINI 指數、剪枝、控制樹的深度後找出最優的樹結構。

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：如果沒有做 feature normalization 的話，兩個 model 的準確率都會相對於有做 feature normalization 來得更低，並且都會出現如圖所示準確率突然爆掉的情況。



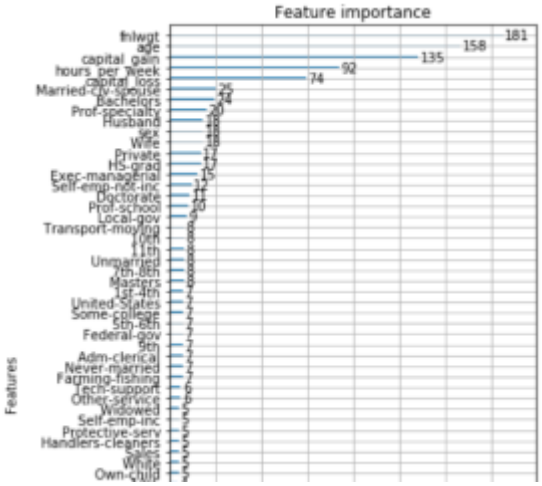
4. 請實作 logistic regression 的正規化(regularization)，並討論其對於你的模型準確率的影響。

答：圖為 r=0,r=1,r=10,r=100,r=1000 時 testing data 的準確率，可見做正規化對 model 的影響並不大

Lambda	testing
0	0.85167
1	0.85250
10	0.85272
100	0.85122
1000	0.85073

5.請討論你認為哪個 attribute 對結果影響最大？

左圖为模型最後的參數分析，右圖為各 attribute 與 label 的相關係數，從中可以發現 fnlwgt 、age、 capital_gain 、Married-civ-spouse、Husband 等幾個 attribute 都比較重要。其中年齡、是家庭男主人和是否已婚幾個都是比較可以理解的會直接影響收入的屬性，其中對結果影響最大的可能會是年齡。



capital_loss	0.150526
Prof-school	0.154907
Masters	0.174570
Bachelors	0.180485
Prof-specialty	0.185866
Exec-managerial	0.214861
sex	0.215980
capital_gain	0.223329
hours_per_week	0.229689
age	0.234037
Husband	0.401035
Married-civ-spouse	0.444696
income	1.000000