學號：R06922006 系級：資工碩一 姓名：劉宏國

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

答：

(1) generative model　public score：0.84582　private score：0.84227

(2) logistic regression　public score：0.85368　private score：0.85026

➔ 根據 public score 和 private score 的成績來看，logistic regression 的準確率較佳。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

(1) Xgboost　public score：0.87800　private score：0.87397

➔ 使用的是 eXtreme Gradient Boosting 的分類器學習模型，他會將許多分類準確率較低的樹模型組合起來，成為一個準確率較高的模型，而在產生每一個樹時會使用梯度下降的方式，造成生成一定數量的樹，較能達到較高的準確率。且他還可以利用 CPU 多工的運算方式達到較高的效率。

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：

一、Normalization：

(1) generative model　public score：0.84582　private score：0.84227

(2) logistic regression　public score：0.85368　private score：0.85026

(3) Xgboost　public score：0.87800　private score：0.87397

二、Without Normalization：

(1) generative model　public score： 0.84570 private score：0.84191

(2) logistic regression　public score：0.78771　private score：0.78565

(3) Xgboost　public score：0.87800　private score：0.87397

➔ Logistic 及 generative 的 model 透過標準化後，準確率都有明顯提升，但是 Xgboost 則是無明顯的提升。

➔ 未使用標準化的話，那些數量級小的變數要變化很大才會對 y 有影響，因此準確率不會有較高的成果，而 Xgboost 是用 tree 的演算法來計算，因此有無使用標準化效果比較不顯著。

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：

一、Regularization(0.0001)：

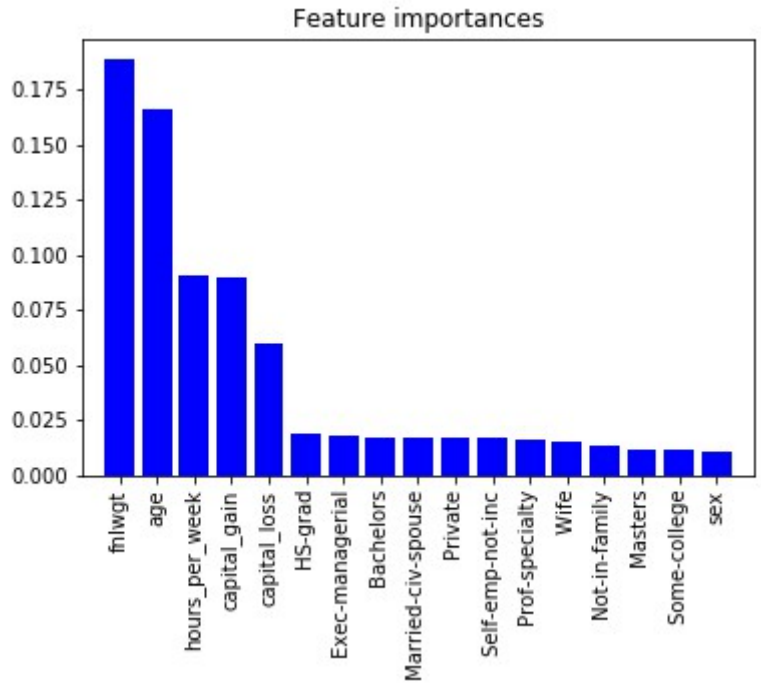
logistic regression　public score：0.85368　private score：0.85026

二、Without Regularization：

logistic regression　public score：0.85343　private score：0.85063

➔ 此次結果對於有沒有做 regularization 來說，並無明顯差異。

5.請討論你認為哪個 attribute 對結果影響最大？



上圖為透過 sklearn 的套件找出影響 y 最大的 feature 排名。

➔ fnlwgt (final weight)

最後分析權重是美國人口普查局用來獨立估計美國平民非機構人口統計變異的控制權重，美國人口普查局使用 3 套控制包括：

(1) 每一州每一劃分區域的 16+人口估計

(2) 西班牙裔根據年齡和性別來源控制

(3) 種族、年齡和性別控制。

其估計是依據人口社會經濟特徵來進行加權計分，具有相似人口統計學特徵的人會有類似的權重。因此 fnlwgt 反映出相似社會經濟特徵背景的人的權重值，而具有相似社會經濟特徵背景的人，其收入也很有可能類似。所以 fnlwgt 在收入高低的分類判斷上，相對於其它屬性較為重要。

依據 XGboost (Gradient boosting Decision tree) 梯度提升決策樹機器學習方法，挑出建構分類決策樹重要的特徵屬性共 17 個特徵屬性，其中 fnlwgt 是屬性權重最高的。