學號：B06902066 系級：資工二 姓名：蔡秉辰

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

	Private	Public	Total
Logistic regression dim10	0.85935	0.87420	0.86678
Logistic regression	0.84436	0.84631	0.84536
Generative model	0.79842	0.80626	0.80234

註：

(1) Logistic regression dim10：將['capital_gain', 'capital_loss', 'hours_per_week', 'age']等項不僅使用線性，將其用 10 次方來做 fitting，除此之外亦有使用 normalization 和 regularization。繳交之 logistic 即用此方式。

(2) Logistic regression：單純使用線性做 logistic regression，且無使用 normalization 和 regularization。

結論：

使用 Logistic regression，無論有沒有將特定項丟到高次方、無論有沒有做 normalization 和 regularization(前提是 eta 沒有太大 or 太小)，都較 Generative model 的表現來得更好。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

(1) 準確率：

Private-0.87605; Public-0.8742 Total-0.87513

(2) 訓練方式：

使用 sklearn 中的套件：GradientBoostingClassifier

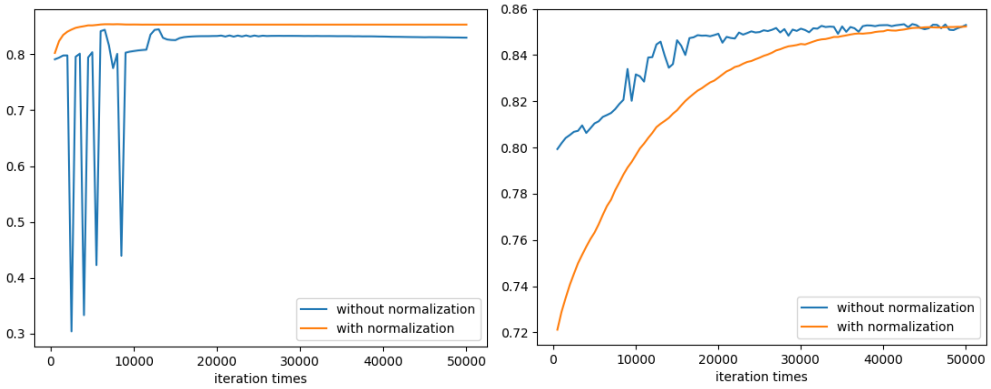
即為採用 Gradient boosting 之方式。Boosting 即為 blending 之方式，結合許多組 weak model，來使其變成 strong model，且較不容易 overfit。

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

	Cor_train	Cor_test
$\eta = 10^{-3}, \text{normalization}$	0.85335	0.84687
$\eta = 10^{-3}$	0.82994	0.82047
$\eta = 10^{-4}, \text{normalization}$	0.85231	0.84755

$\eta = 10^{-4}$	0.85295	0.84860

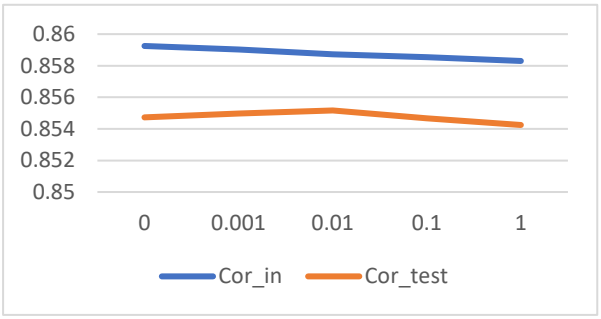
註：Cor-…表示正確率



(註：縱軸為 training data 正確率，橫軸為 iteration)
左圖為使用$\eta = 10^{-3}$，右圖則是使用$\eta = 10^{-4}$。Iteration 均為 50000。
在 η 較小時，沒有 normalization 的會較快達到好的 performance，但也較為不穩定；在 iteration 到 50000 次時，兩者在 training data 正確率差不多，在 testing data 亦差不多；在 η 較大時，有 normalization 的準確率明顯較沒有做 normalization 的好(traing/testing data 均是)，且也穩定得多，到達穩定值所需之 iteration 也沒有到太多。

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。

因為原本的 feature 太少，regularization 的效果會不顯著，故用第一題所提之 Logistic regression dim10，並調整 λ ，結果如下圖：



可以看出在有適當之 λ 時，對準確率的確有些許提升，但並不顯著。

5. 請討論你認為哪個 attribute 對結果影響最大？

用單純之 logistic regrssion 做出之 w 來看，可以看出擁有最大和最小的兩項分別為：Doctorate(此項>0), Jamaica(此項<0)。其代表：擁有 Doctorate 對判斷結果產生年收大於 50W 的影響最大；國籍為 Jamaica 對判斷結果產生年收小於 50W 的影響最大。