學號：R05942107 系級： 電信所碩一 姓名：吳易霖

1.請比較你實作的 generative model、logistic regression 的準確率，何者較佳？
一開始實作的時候的想法預期應該會是 logistic regression 的方法會比較好，而最後的結果不管是 pubic 還是 privite 確實都是 logistic regression 比較好，有可能是因為我們對影響收入的這些 feature 的分布假設是有錯誤的，也可能有一部份原因是我們的 data 資料夠多，讓 logistic regression 的正確率能夠夠高。

2.請說明你實作的 best model，其訓練方式和準確率為何？
best model 是從 logistic regression 為基礎下去修改，normalization 的部分是用 Rescaling 的方式將連續型資料壓縮到[0, 1]之間，而 feature 方面拿掉了問號(未知項)，畢竟在邏輯上我們不該把所有的未知都分再同一類別裡面來用 0、1 分類，國家也拿掉了，原因是因為國家的分類太多太複雜(佔了超過一半)，在這樣的 sample 數量下對 model 可能反而會有不好的影響，然而對 feature 不管做什麼樣的刪減正確率幾乎都維持在一個一定的值，原因大概是因為當初這份 database 的蒐集者已經對該領域有相當的理解，選出了需要用到的幾乎完美的 feature 種類。

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。
在一開始實作時就有使用標準化，原因是有些連續型 feature 的分布範圍實在是太廣了，而離散型的資料範圍就只有 0、1，在尺度上面的差異事非常巨大的，因此可以很直觀的預期有做 normalization 後的準確率應該會提高不少。

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。

λ	Private+public
0	0.85161+0.85442=1.70603
10	0.84424+0.84508=1.68932
100	0.83699+0.83734=1.67403
1000	0.81464+0.81855=1.63319
10000	0.76231+0.76523=1.52754

可以看的出來隨著 λ 的增加正確率逐漸下降，有一部份的原因可能是來自於 feature 已經做過標準化，已經不會有哪一項 feature 對 output 造成過大的影響，因此正規劃的部分就變得沒有那麼必要。

5.請討論你認為哪個 attribute 對結果影響最大？
如果從方法(generative、logistic)、feature、標準化、正規化這四種面相來討論的話，我覺得影響最大的是方法和標準化，原因分別在 1、3 小題中已經詳述。而影響偏小的是 feature 和正規化，原因分別在 2、4 小題中已經詳述。