學號：R03922165 系級： 資工碩二 姓名：李冠穎

1.

	Public score	Private score
Logistic Regression	**0.85405**	**0.84866**
Generative Model	0.84582	0.84240

這邊的 feature 是助教所提供的 X_train 並對所有 feature (包括 dummy variable)做 Standardization (減 mean 除以 std.)。

可發現 Logistic Regression 的準確率較高，因為 Generative Model 在 fit training data 的時候會預設一個分佈 (通常為 Gaussian)，但時常真實資料並不符合高斯分佈，此一假設就有可能造成錯誤的 fitting。

2.

最好的 model 利用 xgboost 所訓練出來的三個 model 進行 ensemble 所得到的。這些 model 都是從 20-fold cross-validation 裡面 validation accuracy 最高的 model (不同 round)

	Max Tree Depth	Boosting Iteration	Objective	Eta	Subsample	Column Subsample	Early Stopping	Private Acc.
1	3	30	Binary:logistic	1	1	1	NaN	86.67%
2	5	50	Multi:softmax	1	1	1	NaN	87.21%
3	6	500	Binary:logistic	0.1	0.8	0.65	50	87.36%

這邊的 feature 是另外抽取的並只對 numerical feature 做 Standardization。抽取方式為保留 numerical feature 的值，並將 categorical feature 做 one-hot encoding 外加一維 unknown，最後總共有 113 維。

Model 3 的 parameter 是透過 step-wise grid search 所找出的組合，方法為 fix 初始參數找另一至二個參數的 optimal，再 fix 最新的參數尋找下一組，如此遞迴 。(可見 tune_gbdt.py)

參數 grid search 依序為 boosting iteration -> max tree depth, min child weight -> gamma -> subsample, column subsample -> regularization alpha。

Ensemble 則是使用最簡單的方式 max voting, 輸出最多票數的結果。

3.

由於 logistic regression 在表現上優於 generative model，因此 3,4 題都以 logistic regression 為探討對象。

以下 model 皆為 epochs=1000, lr=0.001, adamOptimizer, lambda=0.01 的結果。

	Public score	Private score
Feature standardization	**0.85294**	**0.84657**
w/o Feat. Standardization	0.78488	0.78344

可以看到基本上沒有 feature standardization 是 train 不大起來的，model 的 training accuracy 只有 0.795，且事實上 training data 收入低於 50k 的人 (label 0) 就佔了超過 7 成 5，更能確定 feature standardization 的重要性。

4.
由 3 我們已經可以知道 feature normalization 的重要性，這裡以此標準討論。
以下 model 皆為 epochs=1000, lr=0.001, adamOptimizer 的結果。

	Public score	Private score	Training Accuracy
Lambda=0	0.85343	**0.85100**	**0.8530**
Lambda=0.1	0.84226	0.83896	0.8414
Lambda=0.01	**0.85393**	0.84731	0.8516

可以看到 training accuracy 是符合邏輯的，也就是 lambda 愈大，regularization 效果愈好，在 training data 上的表現就會相對差一些。但在 testing 上，由於一些隨機的效果 (包括來自 batch shuffling)，似乎 lambda 比較大時候(0.01 v.s. 0.001)，testing accuracy 並沒有比較好。當然每個 dataset 也不盡相同。

5.
我認為 **occupation** 應該佔很大的關係，直覺上來說。我利用 xgboost booster 所提供的 function 將各個 feature 的 f-score 印出來，發現第一二名不外乎 age 跟 fnlwgt，另外前幾名都是 numerical 的 features。不確定是否是因為 numerical feature 才有 normalize 的緣故 (但正常來說 encode categorical feature 成的 one-hot vector 不會 Normalize)

下次也許可以透過一些 correlation 或是 statistical 的方法做一些分析，或許可以挑出比較適合的 feature。