學號：r07942092 系級：電信所碩一 姓名：白佳灝

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

	Testing (public)	Testing (private)
generative model	0.76707	0.76047
logistic regression	0.85319	0.85136

logistic regression 比較好。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

我使用 random forest classifier，因為在討論分類問題時，decision tree 可能會有比較好的效果，且 RF classifier 又是其 ensemble。而我的訓練方式為:

- 調整 estimator 的數目，考慮整體樹投票決定的影響
- tree 的深度(max_depth)，深度亦會影響決策
- oob_score 設為 True，考慮採用袋外樣本評估模型好壞
- 劃分時考慮的特徵樹目 max_feature 選擇

最後參數選擇 : RandomForestClassifier(n_estimators=350,max_depth=25,random_state=0,
max_features='log2',oob_score=True)

Training accuracy : 93.3693%

Testing accuracy(kaggle public score): 86.474%

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

Testing acc(public)	normalization	Non-normalization
Logistic regression	0.85319	0.79950
Generative model	0.76707	0.16511

特徵標準化會讓模型的準確率提升。其實我們觀察一下 training data，106 維的資料裡面其實佔大部分的數字是 0 和 1，但是有少數幾筆特徵是突然爆高的，像是 age、fnlwgt、hours_per_week 等…，這些值如果沒有做 normalize(換句話說將值弄在 0 到 1 附近)，對於 model 的 learning 是會有影響的。 又 Generative model 比 Logistic regression 受到更嚴重的影響。

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。

Iteration 1000 times, learning rate 0.1

lambda	Testing (public)	Testing (private)
1	0.83906	0.83380
0.1	0.83882	0.83478
0.01	0.83882	0.83478
0.001	0.84054	0.83478
0	0.84004	0.83392

從上面的數據觀察，我們發現其實有無加 lambda 對於 testing 的分數沒有到太劇烈的影響，可能是因為我們在做 logistic regression 時，x 的值都會先經過 sigmoid function，而吐出的值都是藉在 0-1 之間的。Lambda 基本上是對 x 的輸入值有劇烈的變化時才能觀察到明顯的變化，但對這邊的 x 是還好的。

5. 請討論你認為哪個 attribute 對結果影響最大？

如果將所有 feature 的 correlation 畫成 heatmap，我們可以得到一下這張圖。

根據觀察結果發現，影響 income 比較關鍵的(選擇 correlation > 0.2)因素有: age, education_num, sex, capital_gain,hours_per_week。