學號：b04104040 系級：工海三 姓名：解正安

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

	kaggle Public	kaggle Private
Generative model	0.84631	0.84092
Logistic regression	0.85724	0.85456

Logistic regression 的結果較好，原因在於 data 做過 one hot 和許多 data 屬於 binary，attribute 並不連續。此外，Gaussian 分布只能算是一個假設，可能有更好的分布去訓練。

2.請說明你實作的 best model，其訓練方式和準確率為何？

Best model 使用 sklearn 套件中的 RandomForestClassifier() 隨機森林去訓練。並 import GridSearchCV 尋找最佳參數。此外，資料處理部分刪除 fnlwgt','native_country'，評估認為他們對 income 影響較小。

參數如下:

```
(RandomForestClassifier(n_estimators=90,max_depth=13,min_samples_split=50,oob_score = True, random_state = 42,max_features='auto')
```

	kaggle Public	kaggle Private
RandomForestClassifier	0.86363	0.86095

準確率比 logistic 和 generative 皆佳，但其實 logistic 效果應該可以更好，feature 部分應該能在進一步確認。

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

	Private with normalization	Private without normalization
Logistic regression	0.85456	0.42636

由於本次作業有不少 feature 的數值差異非常大，像是 capital_gain、capital_loss、hours_per_week、education_num 等等，差距有到 10 倍以上。因此作 normalization，會使最後的模型準確率較佳。

	Private with normalization	Private without normalization
RandomForestClassifier	0.71047	0.86095

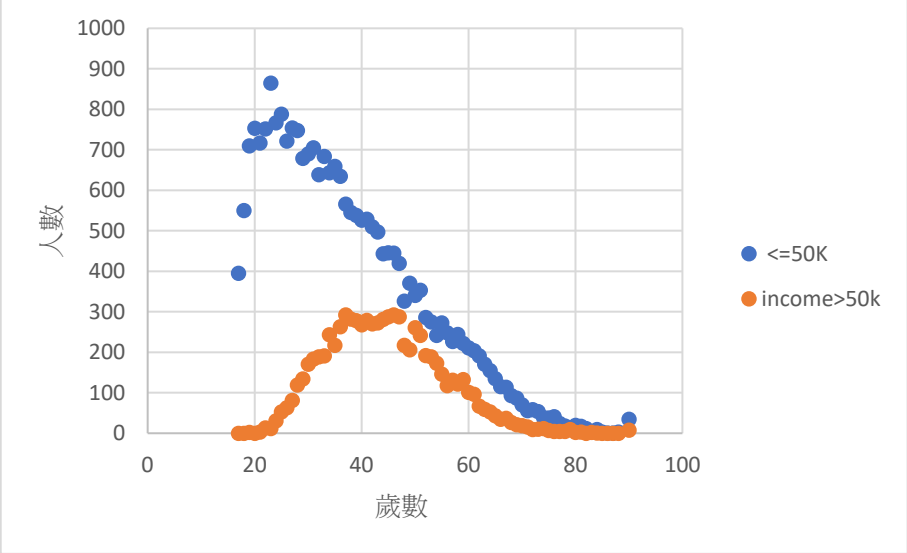
然而在 randomForst 上，如果作 normalize 對結果沒有影響或是更差，因為 randomForst 只是在每個節點找尋最佳的分裂方式。

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。

	Private with normalization
Lamda=10000	0.80309
Lamda=100	0.85210
Lamda=0.1	0.85456

Lamda 在 0.1 時和 100 時，差別不大，且準確率都算相當高，且沒有遇到 overfit 的問題。而在 lamda=10000 時，在 traning set 準確率不高，為 0.807162，private 也不高，應該是遇到 underfit。

5.請討論你認為哪個 attribute 對結果影響最大？



歲數影響最大。由圖形可以知道年輕人普遍薪水較低，但隨著年紀增長，薪水增加，因此>50k 人口增加。然而當兩種收入都下降時，代表是多數人退休年齡，分析大約是在 50~60 歲之間。