學號：R07922134 系級：資工碩一 姓名：陳紘豪

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

	Private	Public
Generative model	0.84080	0.84705
Logistic regression	0.85345	0.85171

由觀察可得知，Logistic regression在Private和Public的表現皆比Generative model還要好，原因可能是Generative model用到了隨機分佈的概念，所以表現得會比較差。

2. 請說明你實作的best model，其訓練方式和準確率為何？

	Private	Public
Best	0.85763	0.85958

I. 對 continuous features [**age**, **fnlwgt**, **capital_gain**, **capital_loss**, **hours_per_week**] 做 min-max normalization
II. 再來取 [**age**, **fnlwgt**, **capital_gain**, **capital_loss**, **hours_per_week**]，這五個維度的 log 和 2次方形成總共116 維的 features
III. 拿來做 logsitic regression，其中 optimizer 是 Adagrad，learning rate是 1，iteration是10000。

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

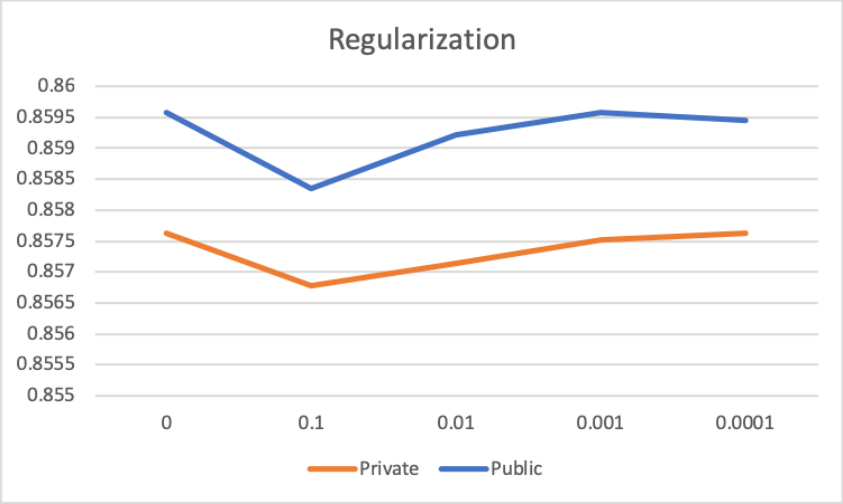
所選的 model 採用和 best model 一樣具有116維度的 features，以下表格為針對不同的 normalization 方法以及 normalization 在不同 features 上所對準確率的影響

Normalization	Private	Public
無	0.53077	0.52051
min-max	0.85763	0.85958
z-score	0.760470	0.76707

由此可知，在沒有採用任何 normalization 方法時，整體的表現是最差的，而 min-max normalizaiton 對整體 performance 的影響最好。

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。

所選的 model 採用和 best model 一樣具有116維度的 features， 以下為針對不同的 lambda 值對準確率的影響



由此可見在0.1時準確率是最低的， 而在不做 regularization 時的準確率反而是之中最高的。

5. 請討論你認為哪個attribute 對結果影響最大？

所選的 model 只採用最基本的 106 個 features， 並針對其中的 continuous features 去找出哪一項對結果的影響最大， 如以下表格

Attribute	Private	Public
106個全取	0.85345	0.85171
拿掉 age	0.85173	0.85171
拿掉 fnlwgt	0.85136	0.85233
拿掉 capital gain	0.83417	0.84066
拿掉 capital loss	0.84780	0.85122
拿掉 hours per week	0.84940	0.85307

由觀察可知， capital gain 對結果的影響最大， 可導致 performance 在 private set 上下降了 2%