學號：B03902101 系級： 資工四 姓名：楊力權

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

答：

	public score	private score
generative model	0.84533	0.84215
logistic regression	0.85442	0.85087

logistic regression的結果較好

2.請說明你實作的best model，其訓練方式和準確率為何？

答：

最好的結果是使用Gradient Boosting Regressor

使用套件sklearn的GradientBoostingClassifier，使用logistic regression的loss function，設定初始learning rate=0.1並作adaboost。

得到public score:0.87788 private score:0.87409

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。

答：

使用106維的feature，並且把feature做標準化與沒做標準化訓練model。

	public score	private score
有feature normalization	0.85442	0.85087
無feature normalization	0.78734	0.78368

有把feature標準化的結果明顯好多了，因為X_train中有一些feature的標準差很大或是均值就很高，而有些feature其實是one hot所以均值很低標準差小，因此在train的過程中，一樣的learning rate，標準差大的feature就會對weight有不良的影響，因此把feature都normalize到標準差為1的feature，才會得到較好的結果。

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。

答：

在有做normalization的情況下，且訂在learning rate=0.0002且500個iteration的時候，觀察無正規化及正規化λ=0.1、0.01、0.001、0.0001的效果

	public score	private score
無regularization	0.85466	0.85075
regularization(λ=0.1)	0.83280	0.82839
regularization(λ=0.01)	0.85061	0.84670
regularization(λ=0.001)	0.85368	0.85050
regularization(λ=0.0001)	0.85528	0.85063
regularization(λ=0.00001)	0.85506	0.85087

在正規化係數為0.0001或0.00001時有比沒有正規化的結果好一點點，而正規係數大於0.01後反而導致不好的結果，或許是因為weight的數值不小，因此若λ太大會影響到結果。

5.請討論你認為哪個attribute對結果影響最大？

在一樣的條件下，我把各個feature抽掉，觀察失去哪個feature會造成training的結果變差。

	public score	private score
全部都在	0.87788	0.87409
無age	0.86855	0.86733
無fnlwgt	0.87457	0.87188
無sex	0.87457	0.87077
無capital gain & loss	0.84656	0.84301
無hours per week	0.87592	0.86991
無workclass	0.87235	0.86905
無education	0.87076	0.86377
無marital	0.87235	0.87053
無occupation	0.86683	0.86598
無relationship	0.87579	0.87200
無race	0.87469	0.87065
無native country	0.87285	0.86893

發現拔掉capital gain與loss的結果變差最多，所以將兩個feature再分別抽掉如下

	public score	private score
無capital gain	0.85037	0.85087
無capital loss	0.86953	0.86257

得知在沒有capital gain的情況之下拿到的結果是最差的，因此猜測capitcal gain與loss是非常重要的attribute，尤其是capital gain這項attribute，對結果影響非常大。