學號：B03902096 系級： 資工四 姓名：陳柏屹

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

答：Logistic Regression 較佳。在都有實作 feature scaling 的狀態下，我的 generative model 準確率約為 84.39%，但是 logistic regression model 的準確率可以輕鬆達到 85%以上。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：我使用了 scikit-learning package 的 ensemble.GradientBoostingClassifier()來實作 best model，其訓練方式主要是組合許多較弱的 classifier 來達到一個較強的 classifier 的效果。在經過調整參數諸如 estimator 數量、learning rate 之後準確率可以達到 87.38%。

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：

	w/ normalization	w/o normalization
generative	84.39%	84.41%
logistic	85.25%	80.62%

可以看到對於 generative model 來說是否有 feature normalization 並不太重要。而對 Logistic regression model 在沒有做 feature normalization 的情況下會非常容易使得 sigmoid 中的 np.exp() overflow，因此我在做 sigmoid 之前會先將 np.dot 的結果除以 10,000 來防止 overflow。然而我們可以看到沒有做 normalization，不但會更容易壞掉，且準確率也會變差許多。

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：

LANDA	10	1	0.1	0.01
Accuracy	83.64%	85.03%	85.32%	85.30%

可以看到 regularization 的權重 LANDA 越高，對於模型準確率反而會有負面的影響。我想這和上次一樣，都是因為模型不會過於複雜，不會造成 overfitting，因此也不需要 regularization 來優化。

5.請討論你認為哪個 **attribute** 對結果影響最大 **?**

利用 xgboost 套件的 plot_importance，我們可以發現 feature[1]、feature[0]、feature[5]都位居最重要的前三名，而他們分別是 fnlwgt, age, hours_per_week。