學號：R06946009　系級：　資料科學碩一　姓名：林庭宇

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

	Kaggle Accurcy 平均值
Generative Model	**84.424%**
Logistic Model	**82.513%**

實作出來的結果 Generative Model 的準確率比較好

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：
利用 Sklearn 套件實作，將全部 32561 筆資料的 106 個特徵的一次項都拿來訓練，用 10-fold 交叉驗證測試各個模型，最後測出 AdaBoostClassifier 的準確率最好，其中 AdaBoostClassifier 參數為預設值

Kaggle 平均準確率：86.162%

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：
以下是以 Generative 做比較

	Kaggle Accurcy 平均值
No Feature Normalization	**79.712%**
Feature Normalization	**84.386%**

以下是以 Logistic 做比較，此次實驗的固定參數為　l_rate = 0.1　batch_size = 256　epoch_num = 350

	Kaggle Accurcy 平均值
No Feature Normalization	**78.570%**
Feature Normalization	**81.764%**

這兩個模型做完 Feature Normalization 後準確率都得到了不錯的改善，可見 Feature Normalization 對於提升模型準確率有很大的幫助

以下是以 AdaBoostClassifier 做比較，參數為套件預設值

	Kaggle Accurcy 平均值
No Feature Normalization	**86.162%**
Feature Normalization	**86.162%**

原本預期做完 Feature Normalization 後準確率會提升，但是最後的結果是準確率不變，原本以為是程式寫錯，但是 X_train 與 X_test 的值的確有 Normalization 成功，因此可能是 AdaBoostClassifier 本身比較不受 Feature Normalization 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。

答：

此次實驗的固定參數為　l_rate = 0.1　batch_size = 256　epoch_num = 350

Lamda	Kaggle Accurcy 平均值
10	**28.395%**
1	**79.172%**
0	**81.764%**
0.1	**81.524%**
0.01	**81.696%**
0.001	**81.838%**
0.0001	**81.580%**

當 lamda 過大時準確率會顯著下降，而選擇一個適當的 lamda 時（此例為 0.001），能夠使準確率稍微得到提升

5.請討論你認為哪個 attribute 對結果影響最大？

答：

一般來說

以 logistic regression 為例，從第三與第四題可以發現，做 feature normalization 跟 regularization 都可以提升準確率，但是 regularization 提升的幅度比較小，只有大約 0.1%不到的提升，相較之下，做了 feature normalization 後，準確率上升了 3%以上，因此我認為 feature normalization 對結果影響最大

如果以 attribute 來說，我認為 capital-gain 對結果影響最大，畢竟有錢才能夠投資