學號：R05921120 系級： 電機碩二 姓名：黃浩恩

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

答：

以下為實作之整理

Model	Public	Private	Mean
generative mode	**0.84533**	**0.84178**	**0.84355**
logistic regression	**0.79754**	**0.79314**	**0.79534**

由上可知，此次實作當中 generative mode 有較好的表現。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

Best Model 使用 Scikit-learn 0.19.0 套件中的 AdaBoost classifier 實作，透過 SAMME.R 的演算法來進行演算與訓練，learning rate 為 1.8，n_estimators=250 。

平均分數為：(0.87088+0.86807)/2=0.86947

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：

於此使用 Min-max 標準化之方法，將所有的輸入進行標準化的處理，其公式為：

$$\text{NewData} = \frac{OriginData - minimum}{Maximum - minimum}$$

透過以上之方法，再將正規化後的輸入帶進 generative mode 與 logistic regression 中，皆可得到較優化的結果，為下表。

Model	Public	Private
generative mode	**0.84557**	**0.84080**
logistic regression	**0.80798**	**0.80751**

---

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	Kaggle Public
0.001	**0.79017**
0.01	**0.78341**
0.1	**0.78992**
1	**0.78931**
10	**0.23476**

此次之結果可以得知，加了 landa 的結果皆比沒加 landa 還來的不佳。

5.請討論你認為哪個 attribute 對結果影響最大？

經過分析，其中：

→性別(sex)

在[>50k]中男性佔 84.96%，[<50k]中男性占 61.2%，差距較大，故認為性別有其影響。

→學位(Bachelors, Doctorate, HS-grad, Masters, Preschool, Prof-school, Some-college)

在資料分析中，[>50k]與[<50k]差距也相當明顯，[>50k]中 Bachelors、Doctorate、Masters 與 Prof-school 較高等的學歷佔比皆比[<50k]高。

→婚姻(Married-civ-spouse, Married-spouse-absent, Never-married)

其中[>50k]於 Married-civ-spouse 狀態下佔比為 85.34%，比[<50k]僅有的 33.64%高出 51.7%的差別，也可以透過 Never-married 狀態下知道[<50k]之佔比有 41.37%，比[>50k]之 6.26%高出許多

以上的分析認為性別、學位與婚姻為判斷結果很大的有效參考因素，

其中認為婚姻狀態影響最大。