學號：R06945003 系級： 生醫電資碩一 姓名：林鈺盛

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

答：

我的 generative model 在 kaggle 上得到的 public 分數為 0.84533，而 logistic regression 得到的 public 分數為 0.85466，因此 logistic regression 的準確率較佳。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

我利用 keras 撰寫，model 有兩層 layer，第一層有 28 個神經元，第二層有 1 個神經元，兩層的 activation 都是 sigmoid，optimizer 為 Adam，loss function 為 binary cross-entropy，epochs 為 150 次，batch size 為 64，其中我切了 0.3 的 data 當作 validation。

Model 在 kaggle 的 public 獲得 0.86142 的準確率，在 training data 上有 0.8653 的準確率，在 validation 上有 0.85997 的準確率。

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：

logistic regression 沒有 normalization 的 public 分數為 0.77002，而 normalization 之後的分數為 0.85466

generative model 則是從 0.84520 進步到 0.84533

我覺得兩個 model 的差異是來自於 logistic 需要不斷地去更新參數，所以每次更新都會受到 feature 間的差異影響，因此沒有 normalization 的情況下，準確率就會下降許多，而由於 generative model 是直接計算出參數，所以受到的影響較小。

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：

本題的結果上作業一滿像的，當我不斷調整 lambda 值從 1 到 0.0001，在 public 上獲得的分數幾乎都沒有改變，我覺得 regularization 對於 logistic regression 的影響並沒有 normalization 來得大。

5.請討論你認為哪個 attribute 對結果影響最大？

本題中我挑選了 age、fnlwgt、sex、capital_gain、capital_loss、hours_per_week 共六個 feature 測試，總共測試六次，每次刪除一種 feature，看訓練出來的 model 的準確率，其中除了 capital_gain 之外，刪除其他五種 feature 都還能在 training data 中得到 0.85 以上的準確率，而刪除 capital_gain 之後只剩下 0.8383，因此我再嘗試只用 capital_gain 來 train model，發現在 kaggle public 上也可以得到 0.80171 的準確率，因此我認為最重要的 feature 為 capital_gain，而事實上也滿符合現實中的情況，因為資本利得往往可以帶來龐大的利潤。