學號：R05944041 系級： 網媒所碩二 姓名：戴長昕

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

Logistic regression在 training data 和 testing data 中均有較佳的表現。

| **Model** | Generative | Linear Regression |
|---|---|---|
| **Kaggle Public Score** | 0.81534 | 0.85896 |
| **Kaggle Private Score** | 0.81157 | 0.85333 |

2.請說明實作的best model，其訓練方式和準確率為何？

Best model 使用 logistic regression，並在特徵中加入原有的特徵中加上 (age, fnlwgt, capital_gain, hours_per_week) 等純數值類特徵的二次項，並實作 feature normalization 和 regularization。

| **Model** | Linear Regression |
|---|---|
| **Kaggle Public Score** | 0.85896 |
| **Kaggle Private Score** | 0.85333 |

3.請實作輸入特徵標準化(feature normalization)，並討論其對於模型準確率的影響。

Feature normalization對於訓練的幫助很大，有多項特徵的標準差和數值都很大，導致這些特徵主導參數的訓練，也容易造成輸出overflow，這裡使用標準歸一化的標準化畫方法。

| **Model** | Feature Normalization | No Feature Normalization |
|---|---|---|
| **Kaggle Public Score** | 0.85896 | 0.83442 |
| **Kaggle Private Score** | 0.85333 | 0.83112 |

4. 請實作logistic regression的正規化(regularization)，並討論其對於模型準確率的影響。

分別實作 L2 正規化，即將 bias 以外的 $\theta^2$ 相加再乘上一個純量 regularization lambda $\lambda$，和 L1 正規化，即將 bias 以外的 $\theta$ 相加再乘上一個純量 regularization lambda $\lambda$，後者訓練出的參數具有較佳的稀疏性，具有特徵選擇的功用。兩者都能提升模型在 Kaggle 上的表現。

| **Model** | L1 Regularization | L2 Regularization |
|---|---|---|
| **Kaggle Public Score** | 0.85896 | 0.85123 |
| **Kaggle Private Score** | 0.85333 | 0.84992 |
| **Lambda** | 1 | 1 |

5.請討論哪個attribute對結果影響最大？

因為這種各個特徵之間數值差距很大的特徵在訓練上會出現很多問題，我認為 feature normalization 對於訓練的幫助很大，讓收斂的速度更快，也能夠提升很多準確率，也能夠避免 overflow 的問題。

而參數正規化相比特徵的正規化效果較不明顯，尤其是在餐數量較少時，也可以用 cross validation 取代來避免 overfitting 的問題。