學號：R06942048 系級： 電信碩一 姓名：林彥伯

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

Generative model: 84.12% Logistic regression: 80.23%

可以發現在沒有特別調整的情況下 generative model 效果較好

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：

使用 sklearn 裡面的 GradientBoostingClassifier 來做，正確率可達 87%多

此方法是對於 decision tree 的優化方法，在每次建立模型前是透過上次建立模型前所決定的梯度方向。演算法如下:

Algorithm 1: Gradient_Boost

1 $F_0(\mathbf{x}) = \arg\min_\rho \sum_{i=1}^{N} \Psi(y_i, \rho)$

2 For $m = 1$ to M do:

3 $\tilde{y}_i = -\left[\frac{\partial \Psi(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, \; i = 1, N$

4 $\mathbf{a}_m = \arg\min_{\mathbf{a},\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(\mathbf{x}_i; \mathbf{a})]^2$

5 $\rho_m = \arg\min_\rho \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i; \mathbf{a}_m))$

6 $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}; \mathbf{a}_m)$

7 endFor

end Algorithm

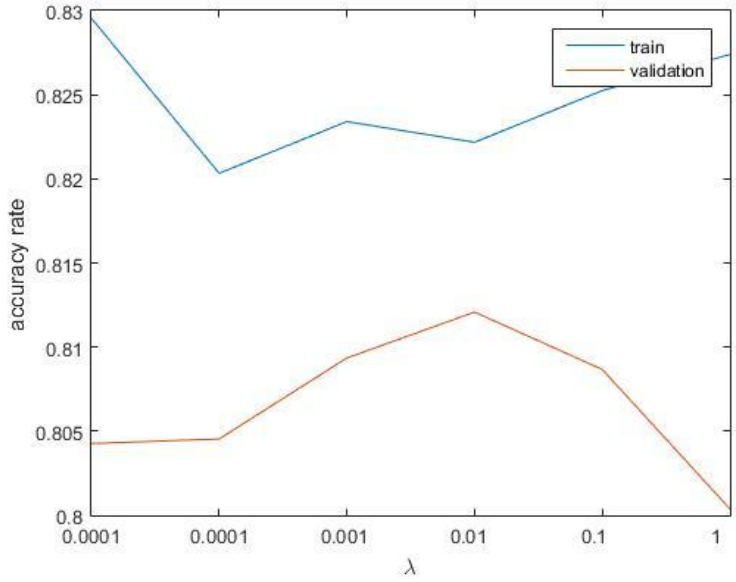
3.請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。
答：

Generative model: 84.41% (with normalization) 23.76% (without normalization)

Logistic model: 80.80% (with normalization) 78.70% (without normalization)

因為部分的 feature 是使用 one-hot encoding，若沒進行 normalization 會使資料彼此的差異性太大，但 logistic regression 並不會受到太大的影響，可能是因為 generative model 需要估算 joint probability distribution 但沒有處理好的資料會導致機率分布不好計算 (某些 feature 直接影響了輸出結果)。

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：



可以發現到在 λ 為 0.01 時準確度會上升，在比較不 smooth 的狀況下，準確度會稍稍提升(?

5.請討論你認為哪個 attribute 對結果影響最大？
本次使用了 sklearn 套件可以直接取得 feature 的影響結果，結果如下:
可以發現 capital loss 的影響最大，而 sex 跟 race 的影響最小