Homework 2 Report - Income Prediction

學號：b05902042 系級：資工二 姓名：林瑋毅

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

根據 kaggle public/private score的結果：

generative model: 0.83572 / 0.83024

logistic regression: 0.85147 / 0.84289

可知 logistic regression 的準確率較佳。

2. (1%) 請說明你實作的best model，其訓練方式和準確率為何？

best model 可從兩個面向探討。首先，資料預處理的部分，對數值類的特徵做了 MinMax scale，類別式的特徵做了one-hot encoding，其中 capital_loss 雖為數值類特徵，但因為分佈及不平衡，所以有做 binarization，而從 fnlgwt 的意義可知其對分類結果不應有影響，故捨棄。其次，模型的部分，採用了典型的隨機森林，沒有做特別的處理。

kaggle public/private score 的結果如下：0.86105 / 0.85714

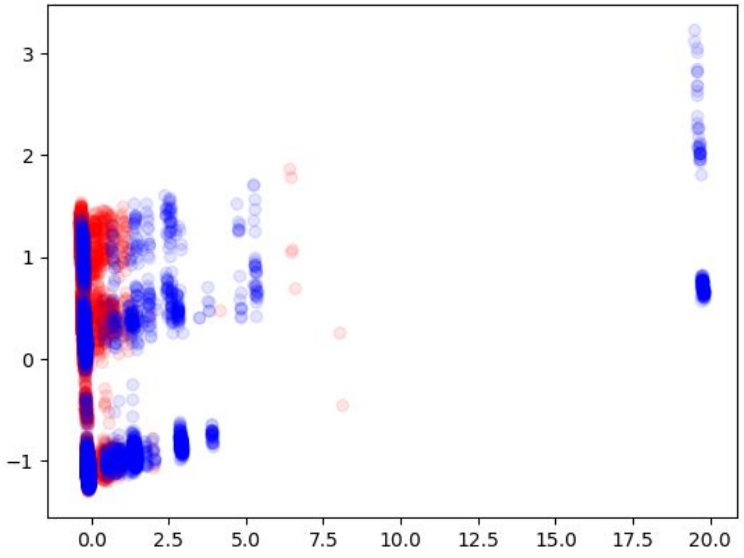
3. (1%) 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。(有關normalization請參考：https://goo.gl/XBM3aE)

以 kaggle public/private score 比較：

有標準化：0.85638 / 0.84707

無標準化：0.85614 / 0.84780

其中，兩組實驗的模型為 Logsitic Regression，其中正規化強度 lambda = 1，且有標準化組僅對數值類資料標準化，由結果可知標準化並未對準確率有太大的影響。推測原因是資料本身並沒有非常線性可分（如下圖為將資料拿去 PCA 的結果），所以不管有沒有標準化，對 Logistic Regression 來說都是難以分類的問題，所以對準確率影響不大。



4. (1%) 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。(有關regularization請參考：https://goo.gl/SSWGhf P.35)

regularization coefficient = 1000 : 0.83894 /0.82901

regularization coefficient = 100 : 0.85552 / 0.84743

regularization coefficient = 10 : 0.85675 /0.84977

regularization coefficient = 1: 0.85638 / 0.84707

regularization coefficient = 0.1: 0.85589 / 0.84670

由此可知 regularization 太大會 underfit，太小則會 overfit ，因此都會使準確率變差。

5. (1%) 請討論你認為哪個attribute對結果影響最大？

可以由經過 feature normalization 的 logistic Regression model 的係數來判斷 attribute 的影響力，其中影響最大的是 edu_num 以及 hours_per_week，且兩者與 income >= 50K 的關係回歸後呈正相關。

註：實驗時，數值類資料有經過類似 min-max scale (將5%及95%的feature 線性調整至0和1)，故數值範圍與 one hot encode 後的類別式特徵相似，所以能夠與類別式特徵比較。