

學號：B06902030 系級：資工二 姓名：邱譯

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

generative model : Public = 0.84029 Private = 0.83712

logistic regression: Public = 0.84828 Private = 0.84301

兩者準確率差不多，logistic regression稍微好一點。

2. 請說明你實作的best model，其訓練方式和準確率為何？

我的best model除了原本所有的feature外，還加入了所有連續項的2~25次方以及sin, cos, tan, arctan的2~25次方。訓練方式為GradientBoostingClassifier (n_estimator = 200, max_depth = 4)。

Public = 0.87641 Private = 0.87507

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

generative model : Public = 0.84729 Private = 0.84056

logistic regression: Public = 0.79840 Private = 0.79461

可看出有無標準化對generative model影響不大，對logistic regression影響很大，我認為是因為logistic用到了exp函數，如果沒有對feature做scaling，計算上的差距就會非常大。

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。

lambda = 1: Public = 0.84828 Private = 0.84301

lambda = 10: Public = 0.84815 Private = 0.84264

可見影響不大，我認為是因為此模型簡單，因此有無正規化並無太大影響。

5. 請討論你認為哪個attribute 對結果影響最大？

我認為capital gain影響最大，因為有資本利得的人，表示有閒錢進行投資，並且有賺錢，則年收入較高的可能行非常高。