學號：B02303106 系級： 經濟五 姓名：廖婉琪

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

generative model		logistic regression	
Private + public 平均	validation	Private + public 平均	validation
0.843125	0.842291	0.845825	0.846613

logistic regression 的準確率較佳，但相差不大。

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：

在 best.py 中利用 XGBoost 實作收入預測，使用 XGBClassifier 訓練模型，在 training data 上得到 87.42%的準確率，在 test data 上則得到 87.48%的準確率，

在 XGBoost 中，會先使用較簡單的模型去模擬，得到較粗淺的結果，在過程中繼續增加這些簡單的模型，像是增加樹的概念，隨著樹增加讓整個 XGBoost 的模型逐漸複雜，直到接近資料本身的複雜度，使訓練達到最符合資料的狀態。

另外在 XGBoost 中，每次增加的模型複雜度都不大，也對增加的節點數做了懲罰，從而限制節點的增加，使得每個節點都是弱的。

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：
未做 feature normalization：
　　private+public 平均：0.844415
　　validation：0.835074
已做 feature normalization
　　private+public 平均：0.84386
　　validation：0.835074
在 generated model 上較無影響，因為 generative model 有模型分配的假設，所以會被模型分配影響而不致有太大誤差。

logistic regression：
未做 feature normalization：

private+public 平均：0.78654

validation：0.793039

loss：5.927884

已做 feature normalization

private+public 平均：0.845825

validation：0.846613

loss：0.294819

在 logistic model 上影響較大，因為 logistic model 是完全依照資料的情況來決定其準確率，故資料的處理對於 logistic model 會有較大的影響。

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：

λ	0.01	0.1	1	10	100	1000
準確率	0.845351	0.846275	0.845965	0.846989	0.846238	0.845453

增加 λ 的值可以幫助模型提升準確率，但 λ 值過大時則會使模型過於平滑。

但在此例子中準確率相差不大，因為此模型並未加入高次項的參數，模型本身原本就較為平滑，因此有無加上 regularization 差異不大。

5.請討論你認為哪個 attribute 對結果影響最大？

篩選 logistic regression 中的 |w| > 0.4 的 attribute，得到以下幾個重要的參數：

變數名稱	w
Age	0.47710
Capital_gain	2.60934
Hours_per_week	0.50126
Married_civ_spouse	0.92738
Never_married	-0.601143

由結果可看出 capital_gain 對於收入預測的結果影響最大，顯而易見地，收入高的人才有可能進行投資賺取資本利得，另外還有已婚且配偶為本國籍、未婚、工時高與年紀大的人，其中只有未婚對於收入的影響為負，其他皆為正。