學號：R06922118 系級： 資工碩一 姓名：吳政軒

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

generative model: 0.84533
logistic regression: 0.85296

logistic regression 較佳，符合預期，通常 generative model 是在 data 較少時可能會有比較好的表現。

2.請說明你實作的 best model，其訓練方式和準確率為何？

我使用 xgboost，是一種 gradient boosted decision tree，原理上是將很多個比較弱的 decision tree 合在一起，讓他有較好的表現，準確率平均為 0.866405

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。

準確率從 0.74817（regularized）上升到 0.85296，因為做了 feature normalization 後，可以將 scale 壓縮到一個範圍內，就好像降低了 variance，因此表現上升。

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。

我使用 L2 regularization，lambda 設為 1，accuracy 從 0.85227（normalized）上升到 0.85296，表示原本可能有一點點的 overfitting，藉由 regularization 限制 overfitting 的程度。

5.請討論你認為哪個 attribute 對結果影響最大？

我將 logistic model 訓練出來的 weight 印出來後，發現絕對值最大的是 capital gain 這一項，capital gain 是指買賣資產所產生價差的收益，所以這項數據是直接影響收入，當 capital gain 高時表示他賺很多，其他像是 hours_per_week 的數據感覺就比較間接，因為工時長如果時薪低不一定會賺很多。