學號：B04901102 系級： 電機三 姓名：簡仲明

1.請比較你實作的generative model、logistic regression的準確率，何者較佳？
答：
generative model：public accuracy = 0.85393
private accuracy = 0.85407
average = 0.85400
logistic regression：public accuracy = 0.85884
private accuracy = 0.85591
average = 0.85738
logistic regression稍微較佳。

2.請說明你實作的best model，其訓練方式和準確率為何？
答：在嘗試過了keras及scipy等等數個套件後，我使用xgboost做為基礎來實做程式。經過許多次實驗後，我將所有的training data切成10-fold的validation來達到最佳效果，其中每一個fold都是把所有的training data中的90%做為training以及10%做為validation使用。接著我調整了xgboost的參數做了10個model(其中每一個model的參數都盡量有很大的不同，希望能達到消除bias的效果)，因此我總共得到了10*10=100組prediction。

接著我利用validation的機率p，帶入公式(1/(1-p))^3來做為對每一個xgboost model的評價(公式是經過多次實驗後的結果)，然後根據此評價做加權平均，最後再對所有fold做總平均(不加權)，得到一個機率，若此機率大於0.5則預測結果為1，否則為0。
public accuracy = 0.87825
private accuracy = 0.87372
average = 0.87599
準確度遠大於前兩個model。

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：

average accuracy	normolized	unnomorlized
generative model	0.85325	0.76377
logistic regression	0.85738	0.79172
xgboost model	0.87606	0.87599

normalization對於logistic regression以及generative model的表現皆有顯著的提升。且若使用unnormalized data做為logistic regression的training data，可以很明顯的看出在training的過程中，loss會產生嚴重的上下跳動，跳動範圍甚至可達loss本身的兩倍之多！而使用normalized data則可以消除這樣的情況，loss會呈現穩定的下修。而unnormalized的generative model，甚至會出現全部都是0的預測！推測是因為矩陣中有過多0值，造成bias出現嚴重偏差所致。至於normalization對xgboost model的影響並不顯

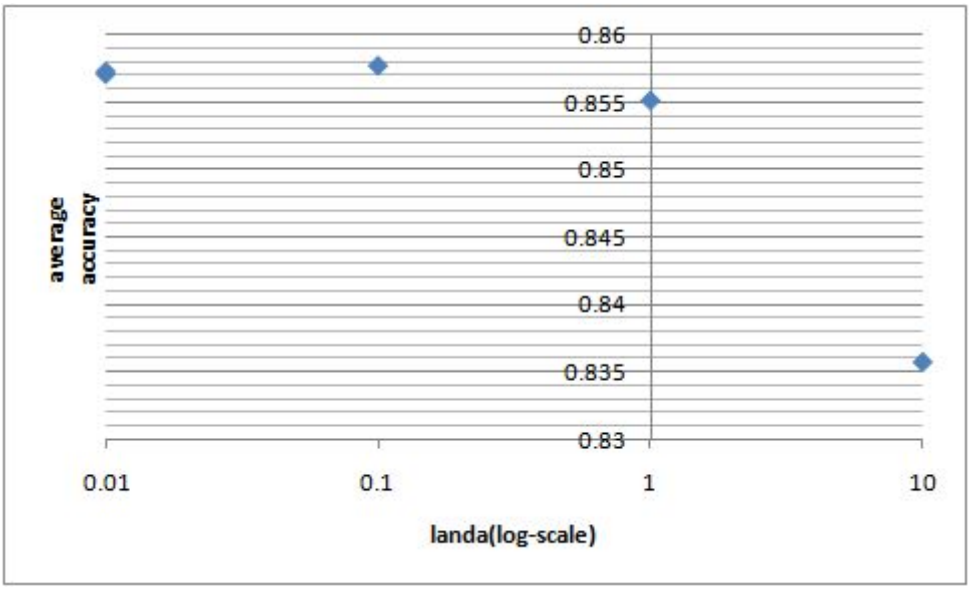
著，猜測是因為xgboost本身即包含normalization相關的程序，因此data的前處理幾乎沒有影響。

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：

landa = 0(unregularized)	0.85726
landa = 0.01	0.86707
landa = 0.1	0.85768
landa = 1	0.85511
landa = 10	0.83576

regularization coefficient(landa)=0.1左右時，logistic regression會有最佳表現，若往上即往下調整則表現皆會下滑。會有這麼小的landa值，是因為weight值本身也都非常小(如5所示，大部分都在10^-2的數量級)。作圖如下所示：



5.請討論你認為哪個attribute對結果影響最大？

觀察normalized logistic regression的係數，可以發現只有age以及capital-gain兩個attribute的係數絕對值超過0.5，其中capital-gain的係數更達到2.5左右。而觀察keras係數也有相似的情形，因此我認為這兩個是影響最大的attribute。