ML2017FALL HW2 Report

學號：B04902090 系級： 資工三 姓名：施長元

1.請比較你實作的generative model、logistic regression的準確率，何者較佳？
答：
兩者都做了 Normalization，Feature 使用 X_train 中的全部 106-dim 的 feature
沒做 regularization

	Private	Public
generative model	0.84166	0.84520
logistic regression	0.85075	0.85417

結果而言，Logistic regression 表現較佳，在 Private 以及 Public 都有壓倒性的勝利

2.請說明你實作的best model，其訓練方式和準確率為何？
答：
Feature : 取用全部 106-dim feature 的一次方，加入 age、fnlwgt、capital_gain、capital_loss、hours_per_week 共五項 Feature 的二次、三次、四次方、以及取 Log，並對這些特徵都做 Normalization
訓練方式使用 XGBoost，Learning Rate 為0.06，n_estimators 為 1100
準確率 : Private : 0.86561
Public : 0.87162

3.請實作輸入特徵標準化(feature normalization)，並討論對於你的模型準確率的影響。
答：
以 Logistic Regression 為例，只取用 106-dim 的一次方做比較

	Private	Public
未做特徵標準化 :	0.79179	0.79545
特徵標準化 :	0.85075	0.85417

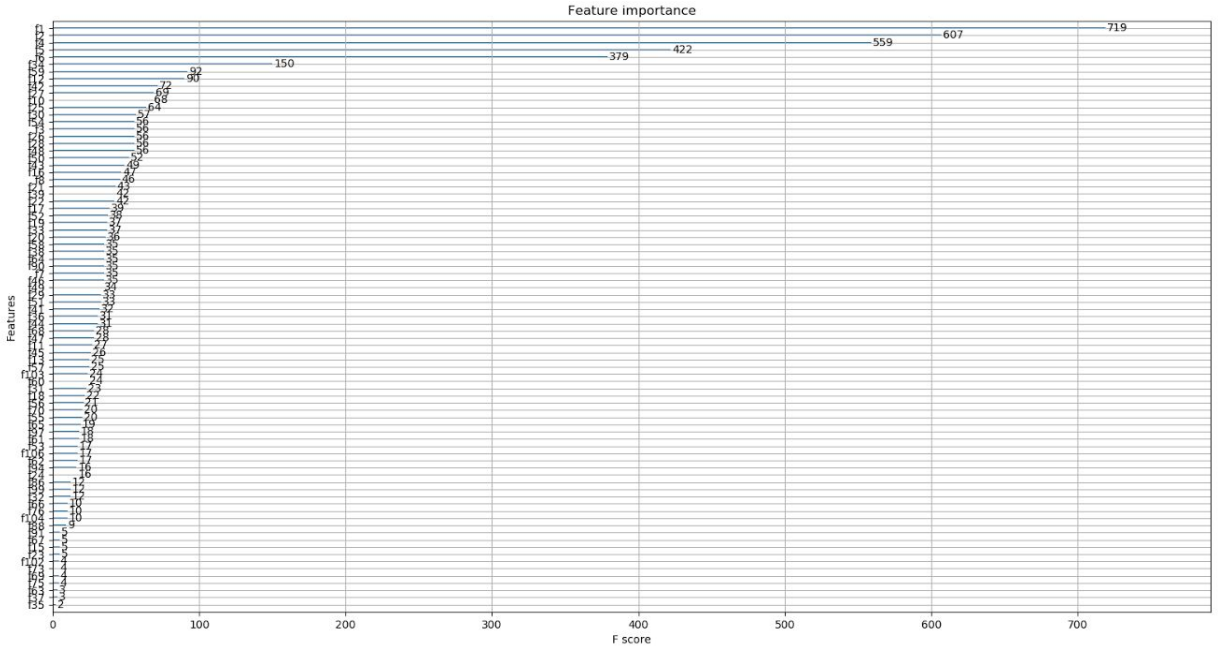
準確率有顯著的增加

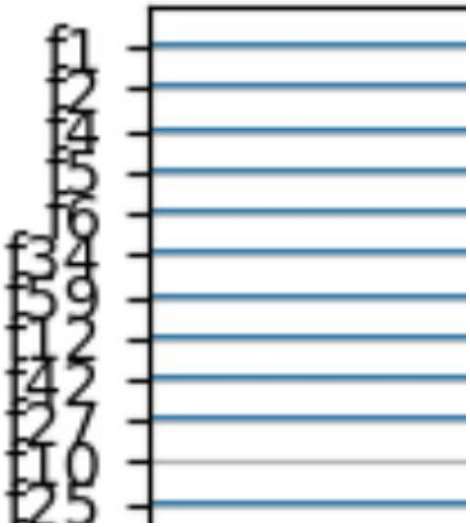
4. 請實作logistic regression的正規化(regularization)，並討論對於你模型準確率的影響。
答：

	Private	Public
0.0 (無regularization)	0.84989	0.85442
0.01	0.84989	0.85442
0.1	0.85014	0.85442
1	0.85014	0.85454

因為題目沒有給予數值，所以個人依照上次作業經驗測試
可以見得，λ=1時，準確率最高
再往上一路加到10時，結果一樣，就不放上了

5.請討論你認為哪個attribute對結果影響最大？




(貼心放大圖)

本圖使用 XGBoost 中的Plot_importance，可以看到 1、2、4、5、6 相對來說比較重要，f1 是 age，F score 高達 719，可以得知，現在戶頭沒什麼錢沒關係，隨著年齡增長，>50K 的薪水將會漸漸的不再是夢想 !!