Machine Learning (2017,Fall)

Assignment 2 – Income Prediction

學號：b03901086　系級：電機四　姓名：楊正彥

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

	Public Score	Private Score	Testing Accuracy
Generative Model	0.84533	0.84215	0.84525
Logistic Regression	0.85356	0.85235	0.85811

從表格中可以看出這次作業二所實作的兩個 binary classification 模型裡，logistic regression 不論在 public set、private set 或者是自己在 training 時所額外切出來的 testing set 中均表現得比 generative model 要來得好。

2 (1%)請說明你實作的 best model，其訓練方式和準確率為何？

由於要實做 binary classification，故選用 *xgboost 以及 sklearn* 這兩個套件完成 best model 的參數選擇以及訓練過程。

首先使用 *feature_ importance()*這個函式來找出所要選用的 features，這樣對於往後的 training 速度會有相當大的幫助，最後我們選用了所有 F score 大於 10 的 features。

接下來要透過選定一個較適當的 *learning rate* 讓往後的參數調整過程更加順利，在這裡經過簡單的幾次嘗試之後，我們先選用 learning rate = 0.1。接下來用 *GridSearchCV()來選定*一組較佳的 Booster Parameters 來進行 binary classification。這邊選定最終參數的依據為 5-fold 的 cross validation 的 error 值，最後用於訓練的一些相關參數為：iterations=360、learning rate=0.1、gamma=0.6、max_depth=4、min_child_weight=1、colsample_tree=0.9。

最後訓練出的模型所 predict 出來之結果分別在 public/private 得到了 0.87776/0.87409 的準確率，與其 testing accuracy=0.875004 亦相當靠近。

3 (1%)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。

由於 *xgboost* 內建 inputs normalization，故這題僅就 logistic regression 以及 generative model 兩個模型來討論，從以下數據可以看出來 feature normalization 對於 linear regression 有一定的幫助，但 generative model 就幾乎沒有影響。

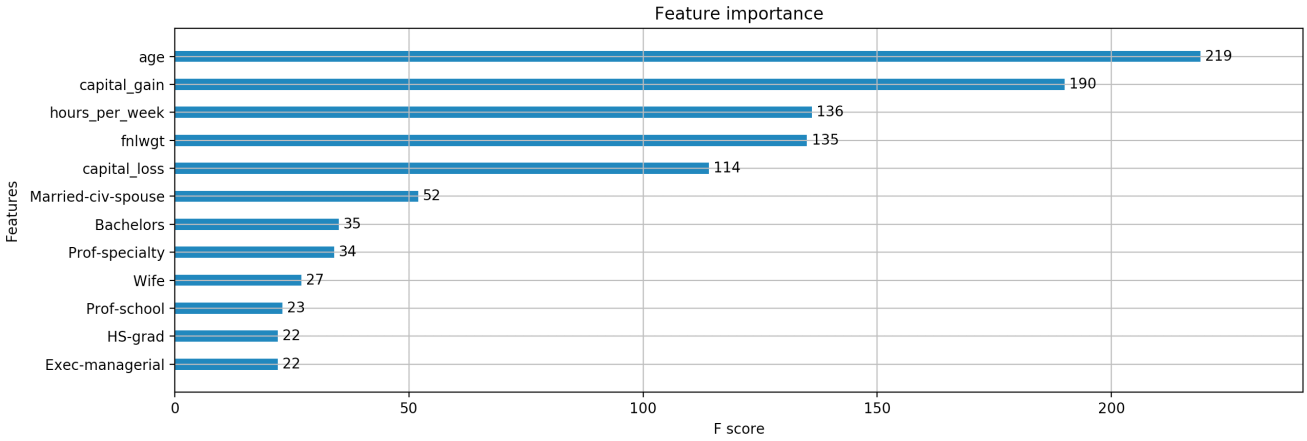
	Public Score	Private Score
Generative Model (with/without normalization)	0.84533/0.84528	0.84215/0.84240
Logistic Regression (with/without normalization)	0.85356/0.85112	0.85235/0.84977

4 (1%)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。

	Public Score	Private Score	Training Accuracy	Testing Accuracy
$\lambda = 0.5$	0.81916	0.81193	94.29%	80.27%
$\lambda = 0.01$	0.85429	0.85100	90.46%	85.59%
$\lambda = 0$	0.85393	0.84989	90.51%	85.32%

因為 feature normalization 的原因，所以正規化的λ值不能太大否則會 train 不起來（$\lambda = 0.5$），當選擇足夠小的λ時（$\lambda = 0.01$）可以開始發現 training accuracy 會開始下降，代表著開始有效控制 overfitting 的狀況，而 testing accuracy 也會比沒有做 regularization（$\lambda = 0$）的要來得好一點。

5 (1%)請討論你認為哪個 attribute 對結果影響最大？



我使用了 *xgboost* 套件中的 *feature_ importance()*來幫助我找出對結果影響最大的 attribute，可以看出來以 age、captital_gain、hours_per_week 有著較大的 F score，代表著年紀、資產增值、每週工時對於 income 是否大於 50K 有高度正相關。