Homework 2 Report - Income Prediction

學號：r06725035 系級：資管碩一 姓名：陳廷易

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
generative model	0.76474	0.76280
logistic regression	0.85257	0.84289

在同樣皆有做 feature normalization 的情況下，Logistic regression 準確率較 generative model 高出不少。在此 classification task 中，由於 logistic regression 並不考慮 jointly distribution，不關心數據是如何生成的，僅是對給定的變量進行分類(直接對 posterior distribution 進行建模)，或許在此分類問題中較不易模擬數據的生成分布，因此在這個監督式 task 下會有相對較好的表現。

2. (1%) 請說明你實作的 best model，其訓練方式和準確率為何？

因為了通過 strong baseline，若僅使用 logistic regression 搭配全部 attributes 將有很大難度。因此首先選取認為重要度較高的 72、77 項 attributes(依據不同 model 表現不同來選擇適合者)，利用 scikit-learn linear model 套件(包含 logistic regression、Lasso、Ridge、ElasticNet、OMP、ARDRegression、TheilSenRegressor、LogisticRegression 等)以及 SVM、DecisionTree、KNeighbors 等方法，再輔以 NN 模型(兩種四層 fully-connected model)。

將上述所有模型將資料切為 75% training 25% validation，利用 validation 的準確率成績作為算術平均數的權重值。而後將所有 training data 對全部 model 重新訓練一遍，最後將預測出來的值乘上對應權重值進行 ensemble 得到最終答案。最後在 public leaderboard 上 score 為 0.85444 在 private leaderboard 上 score 為 0.85958。

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。(有關 normalization 請參考：https://goo.gl/XBM3aE)

	Public Score	Private Score	Training loss change
有特徵標準化	0.85257	0.84289	0.3575-> 0.3486
無特徵標準化	0.23525	0.23719	15.5034-> 6.6655

在其他條件相同情況下，此 task 中有否進行 feature normalization 對結果影響相當大，因為此次訓練資料在某些 attributes 資料分布極為分散，在原始的資料中，各變數的範圍大不相同，若沒有進行標準化，函數很可能會無法正確計算出兩點差異，而過度受某特徵左右，因此透過標準化才能使個特徵依比例影響距離。

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。(有關 regularization 請參考：https://goo.gl/SSWGhf P.35)

	Public Score	Private Score	val_accuracy
λ=0 (無正規化)	0.85257	0.84289	0.850071
λ=1e-3	0.85356	0.84510	0.851913
λ=0.2	0.84778	0.84510	0.841076

在本 task 其他參數皆一致的情況下，logistic regression 的正規化並不會對準確率帶來太大的影響。但從表中可發現當 λ 值較大時，對 w 的 penalty 較大，雖然在 training performance 可能會稍微差一點，但從結果上可看到較能避免 overfitting。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

在挑選 attribute 的部分，因為用肉眼難以縮小判斷範圍，故借用 extra trees、random forest、XGBoost 套件的 feature_importances_來縮小篩選範圍，接著再剔除重要度最高的 feature 來觀察對結果的影響。

依據實驗結果，我認為 age 對結果的影響最大，能令 logistic regression model 的 public 準確率從原本的 0.85257 下降至 0.84338，為所有實驗 attribute 中下降幅度最大者。

※註：以上若未調整實驗用之預設參數皆如以下：

Validation ratio = 50% training data
Learning rate = 0.2
Epsilon = 1e-8
Learning rate decay = 0.995
Batch size = 32
Training epoch = 4100
Regularization lambda = 1e-6

(機器設備：Ubuntu 16.04、anaconda3、GTX1080Ti、i7-8700、64GB ram、sklearn 0.19.1、keras 2.1.4、tensorflow-gpu 1.6.0)