學號：R05943138 系級： 電子所碩二 姓名：賴又誠

1.請比較你實作的 generative model、logistic regression 的準確率，何者較佳？
答：
我使用同樣的 feature(助教抽的 106 種 feature 全用)且都做 normalization，下表為準確率:

Model	Accuracy(public)	Accuracy(private)
Generative model	0.84606	0.84166
Logistic regression model	0.85393	0.85124

結果為 Logistic regression 較佳。

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：
我使用的 best model 為 Gradient Boosting，其訓練方式如下列公式解說:

$$Y = M(x) + error \quad (1)$$
$$error = G(x) + error2 \quad (2)$$
$$error2 = H(x) + error3 \quad (3)$$
$$Y = a*M(x) + b*G(x) + c*H(x) + error3 \quad (4)$$

原本的 loss function 為(1)，但 accuracy 假設為 80%，那我們進一步再把 error 寫成(2)、(3)並整理成(4)之後，可以看到可以當作是我們把三種 model(M、G、H)以不同的係數結合起來成新的 model 那他的 error 被進一步的壓低，使 accuracy 達到 84%甚至更高，最後我用兩個不同參數的 Gradient Boosting 做 voting，得到最後的結果。

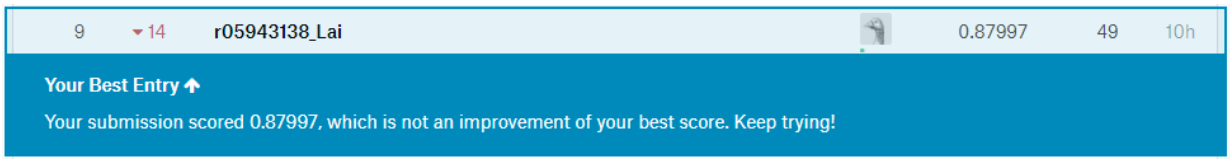
在我 learn 過不同參數後，最後使用的 Gradient Boosting 的參數如下:

Model	Learning rate	N estimator
Gradient Boosting	0.1	600
Gradient Boosting	0.1	700

最後在 Kaggle 上的精準度如下:

Model	Accuracy(public)	Accuracy(private)
Gradient Boosting	0.87997	0.87384

Kaggle public accuracy:



Kaggle private accuracy:

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：

Model	Learning rate	Cross validation	Adagrad	Feature	Epoch
Logistic regression	1	0.9	yes	106	5000

Normalization	Training (accuracy)	Valid (accuracy)
Yes	0.852580	0.853723
No	0.774638	0.332749

由上表可得知，在相同的 model 及參數下，使用 normalization 能大幅提升準確率及效能，若無使用 normalization，即使用 adagrad 也較容易在 training 中發散。

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：

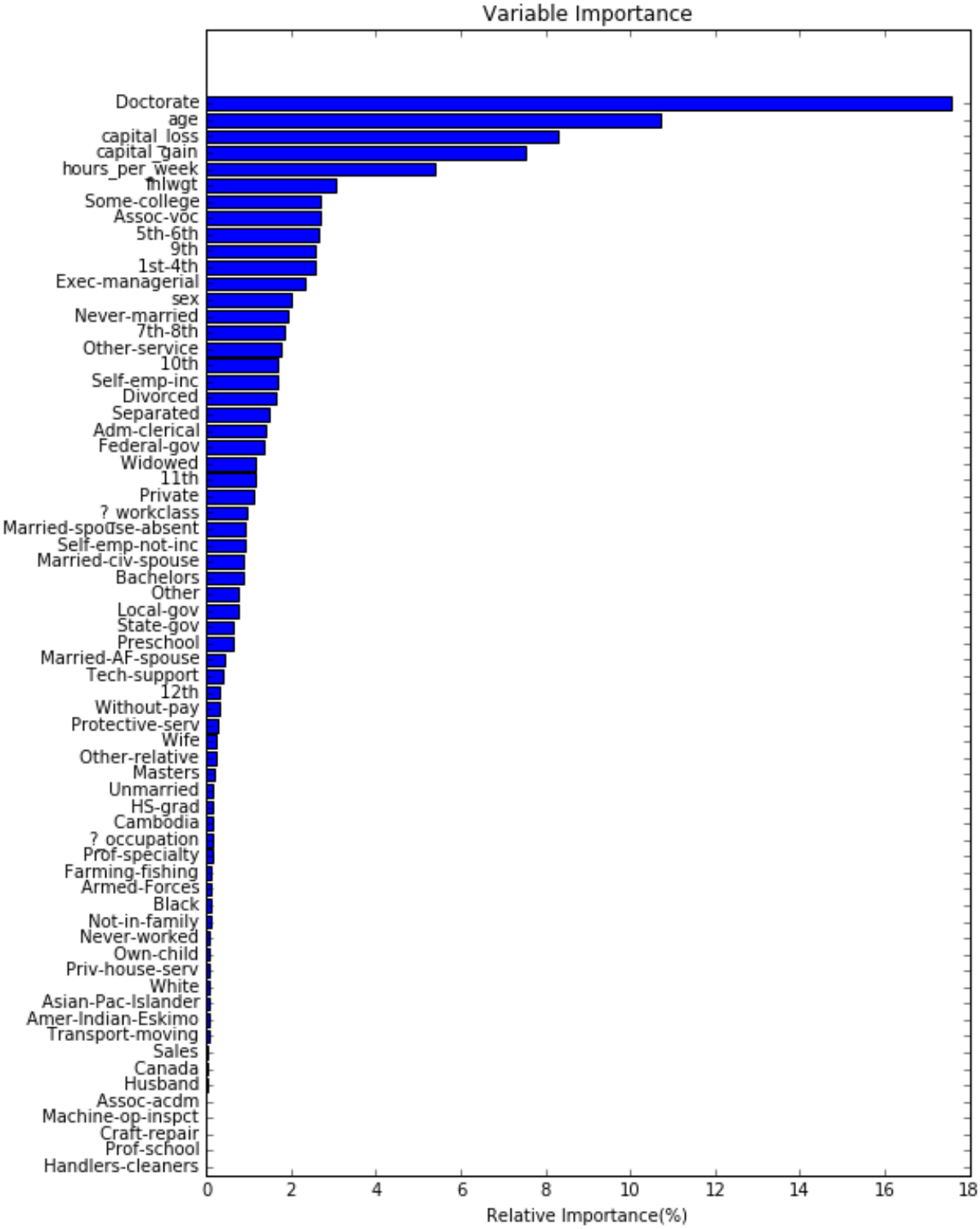
Model	Learning rate	Cross validation	Adagrad	Feature	Epoch
Logistic regression	1	0.9	yes	106	5000

Lambda	Training (accuracy)	Valid (accuracy)
0	0.852580	0.853723
0.1	0.852955	0.854113
0.01	0.852546	0.853723
0.001	0.852546	0.853723

可以看到加了 regularization 後，對模型準確率的影響並不大，我猜是 feature 的選擇，使得模型並沒有太過複雜，因此沒有產生 overfitting。

5.請討論你認為哪個 attribute 對結果影響最大？
答：

下表為使用 sklearn tool 內建好的 function，可以直接跑出各 feature 的重要性，全部相加為 100%(已經把 0%的 feature 拿掉)，藍色軸長度代表在 100%內所佔的%數，可知 education 這個 attribute 內的 Doctorate 佔的%數最大，對結果影響最大。


Variable Importance
Doctorate
age
capital_loss
capital_gain
hours_per_week
fnlwgt
Some-college
Assoc-voc
5th-6th
9th
1st-4th
Exec-managerial
sex
Never-married
7th-8th
Other-service
10th
Self-emp-inc
Divorced
Separated
Adm-clerical
Federal-gov
Widowed
11th
Private
? workclass
Married-spouse-absent
Self-emp-not-inc
Married-civ-spouse
Bachelors
Other
Local-gov
State-gov
Preschool
Married-AF-spouse
Tech-support
12th
Without-pay
Protective-serv
Wife
Other-relative
Masters
Unmarried
HS-grad
Cambodia
? occupation
Prof-specialty
Farming-fishing
Armed-Forces
Black
Not-in-family
Never-worked
Own-child
Priv-house-serv
White
Asian-Pac-Islander
Amer-Indian-Eskimo
Transport-moving
Sales
Canada
Husband
Assoc-acdm
Machine-op-inspct
Craft-repair
Prof-school
Handlers-cleaners
0
2
4
6
8
10
12
14
16
18
Relative Importance(%)