作業二 Report

學號：B06901020 系級：電機二 姓名：張恆瑞

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

	Generative Model	Logistic Regression
Average accuracy	84.108%	85.302%

很顯然 logistic regression 完勝 generative model。

雖說 logistic regression 是從 generative model 推導出來的，但因為 logistic regression 有更多可以調整的空間，像是 regularization 之類的，所以效果比較好應該是正常的。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

利用 gradient boosting (sklearn.ensemble 的 GradientBoostingClassifier)

```
# Start training
print('Start training...')
gbm0 = GradientBoostingClassifier(n_estimators = 160, learning_rate = 0.110, \
    min_samples_split = 200, min_samples_leaf = 50, max_depth = 8, \
    max_features='sqrt', subsample = 0.8, \
    random_state = 10, verbose = 1)
gbm0.fit(x, y)
```

無腦亂調參數並且硬 train 一發就得到這樣的準確率了，難怪此次作業限制使用 decision tree 相關的套件，但聽同學說這個套件漏掉沒有被限制，所以就用了，效果真是好啊～

Average accuracy: 87.114%

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

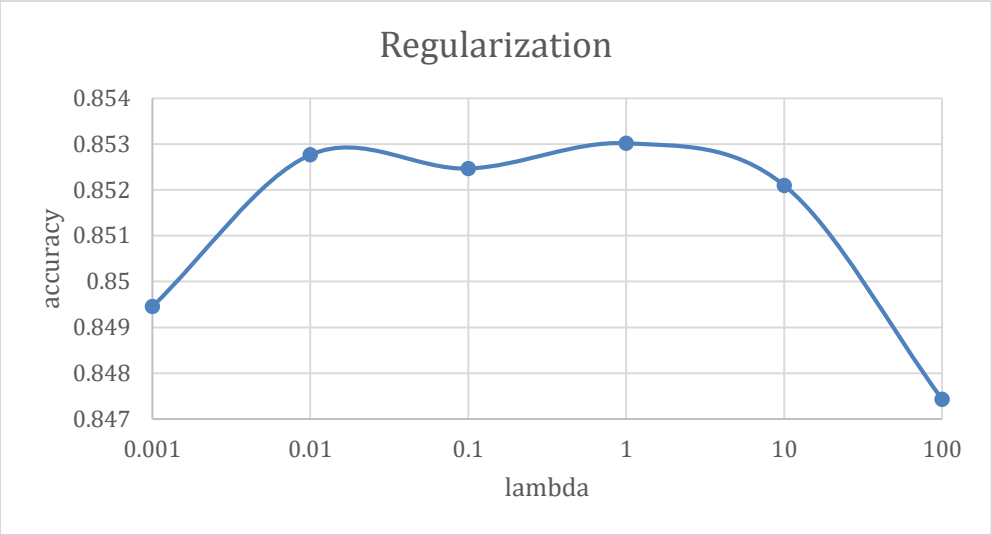
	Without Normalization	With Normalization
Logistic regression	83.558%	85.302%
Generative model	78.914%	84.108%

看得出來，經過 normalization 後，accuracy 都提高不少。

而且 logistic regression 也要因為 normalization 調整 learning rate，才能得到比較好的 accuracy。

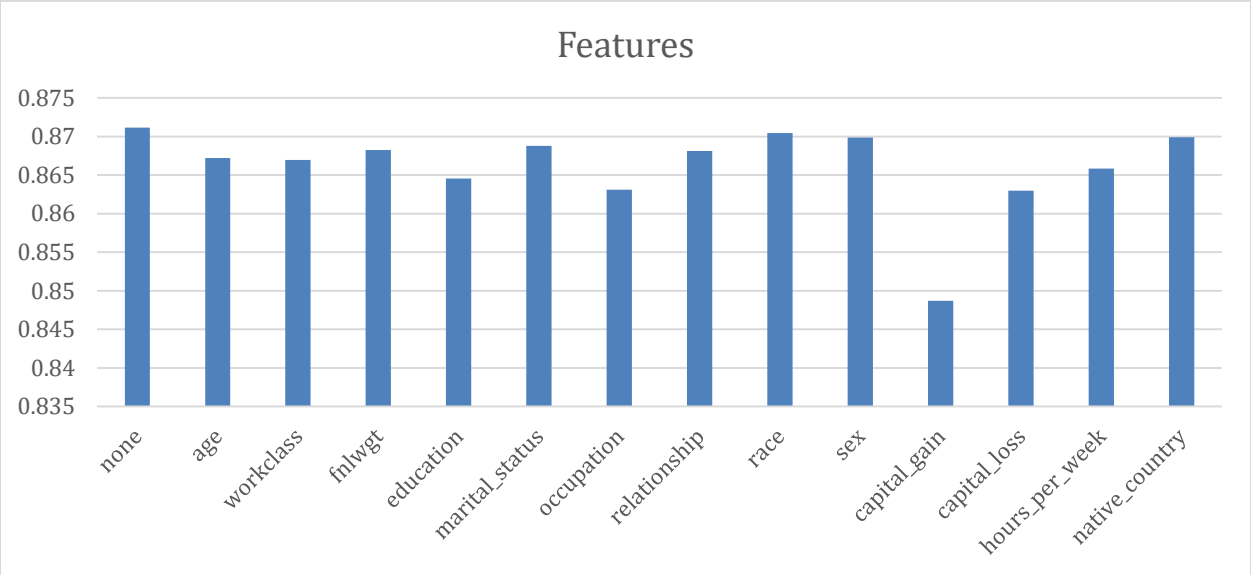
但我只 normalize 一部分資料，因為有些 feature 像是 gender 在 normalize 之後，會讓 model 準確率下降，所以最後只 normalize “age”、” fnlwgt”、” capital gain”、” capital loss”、” hours per week”，其他則是 one hot encoding。

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。



當λ太小對 regularization 沒有影響，太大又讓 model underfit，所以選介於中間的 1 似乎是最好的選擇。

5. 請討論你認為哪個 attribute 對結果影響最大？



(none 為使用完整 dataset 的情況，其餘的都是拿掉該 feature 後再重新 train 的準確率)

(本實驗使用 gradient boosting 來比較)

結論是，使用所有的 feature 效果是最好的，然而拿掉 capital gain 之後準確率大幅降低，可見 capital gain 和一個人的年收入有很大的關係。事實上，capital gain(資本收益)就是直接影響收入多寡的一個 feature，因此拿掉這個項目會對於預測年收入影響相當大。