學號：R06922097 系級： 資工碩一 姓名：鄭雅文

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

答：

	Public set	Private set
Generative model	0.84582	0.84240
Logistic model	0.85552	0.85100

(Logistic第二欄因為會overflow的關係所以將第二欄刪除)

Generative model因為對features之間的共變數關係規定較嚴格，只使用同一個Σ ，所以準確率較Logistic model差。

2.請說明你實作的best model，其訓練方式和準確率為何？

答：

	Public set	Private set
xgboost	0.86842	0.86733

使用xgboost的package，實作gradient boosting decision tree，使用預設參數：max_depth=3, learning_rate=0.1, n_estimators=100, objective='binary:logistic', booster='gbtree', n_jobs=1, nthread=None, gamma=0, min_child_weight=1, max_delta_step=0, subsample=1, colsample_bytree=1, colsample_bylevel=1, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, base_score=0.5, random_state=0, seed=None, missing=None

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。

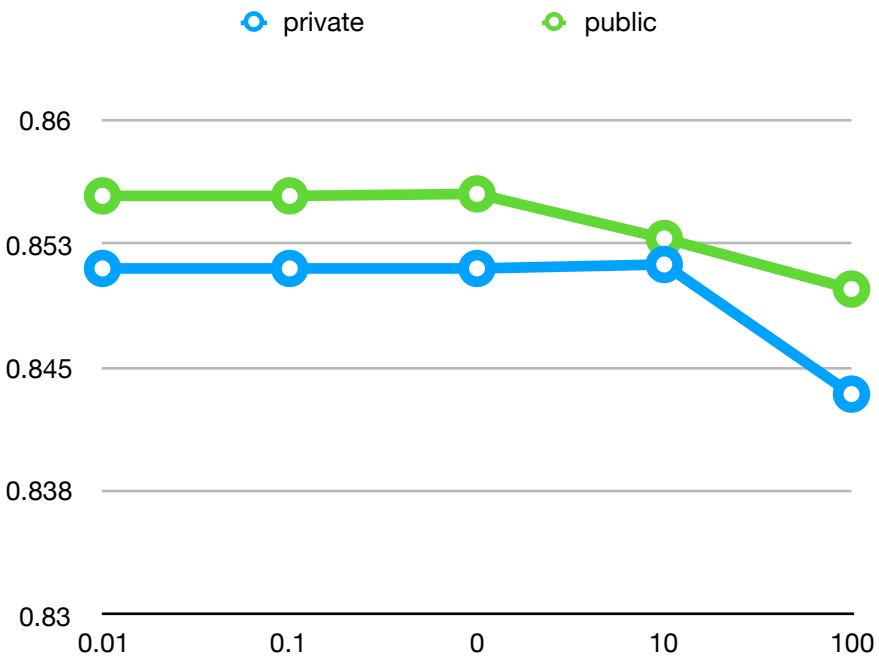
答：

	Normalize on Public set	Change	Normalize on Private set	Change
Xgboost	0.86535	▼0.00307	0.86561	▼0.00172
Logistic model	0.85282	▼0.0027	0.85014	▼0.00086

Normalize過後反而效果較差，可能因為轉成浮點數後會有精度的問題，而且X_train中的data已經有許多是0/1的項目，除了前幾欄data之外scale的變化不會太大，所以除了前幾個欄位有可能增加精確度之外，其他欄有normalize跟沒有normalize不會差太多，還有可能因為normalize後把原本的0/1改成-1~1之間的小數後性質沒有0/1那麼好，所以精準度下降。

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：



縱軸為kaggle上的準確率，橫軸為lambda值。lambda約在0時最佳。

因為此模型都只有用到一次方(X_train幾乎都是0/1，所以取次方沒有用)，所以regularization讓曲線變平滑的作用不大。

5.請討論你認為哪個attribute對結果影響最大？

答：

從logistic的model中得出Doctorate的項目係數最大，為1.1752982442078237；且generative的model中也是Doctorate的項目係數最大，為2.211309859828511。因此可以猜測學歷越高(博士學位)，年收入越高。