學號：R06922129 系級： 資工碩一 姓名：丁縉楷

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

答：

	Private Score	Public Score
generative model	0.84252	0.84533
logistic regression	0.84657	0.83864

2.請說明你實作的best model，其訓練方式和準確率為何？

答：

```
model = Sequential()
model.add(Dense(12, activation='sigmoid', input_dim=x_normed.shape[1]))
model.add(Dense(output_dim=2, activation='sigmoid', input_dim=12))
model.compile(optimizer='rmsprop', loss='binary_crossentropy', metrics=['accuracy'])
model.fit(x_normed, y_encoded, batch_size=128, nb_epoch=100)
```

使用兩層，沒有切validation，epoch = 100

有實作試試看取選取feature但準確率還是把全部的feature餵進去比較高

Private Score	Public Score
0.85198	0.85368

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：

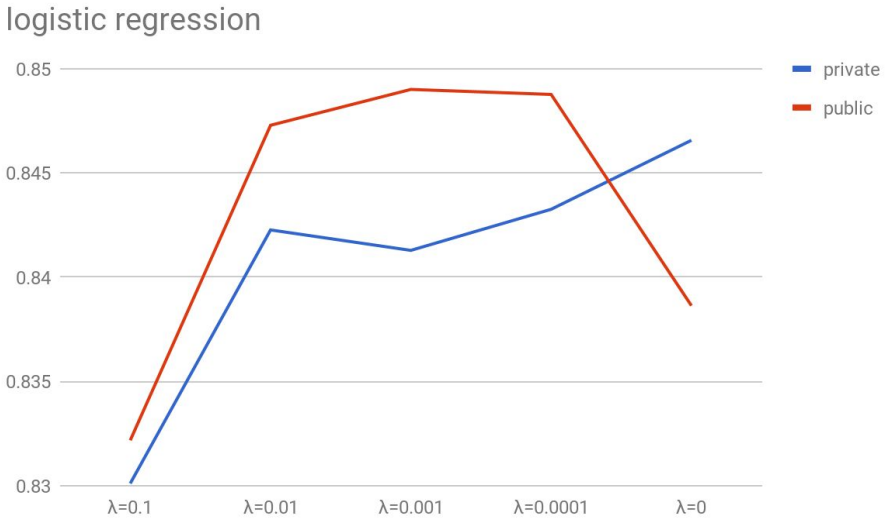
	Private Score	Public Score
generative (w/ normalization)	0.84252	0.84533
generative (w/o normalization)	0.84240	0.84520
logistic (w/ normalization)	0.84657	0.83864
logistic (w/o normalization)	0.78589	0.78968

可以看到normalization對generative model影響不大，因為重新scaling對算高斯分布的平均以及標準差沒有太大影響

但有無normalization就對logistic regression影響很大，因為gradient descent容易受到feature數值大小的影響，在epoch數一樣的情況下，normalization對training速度以及performance都有顯著的幫助

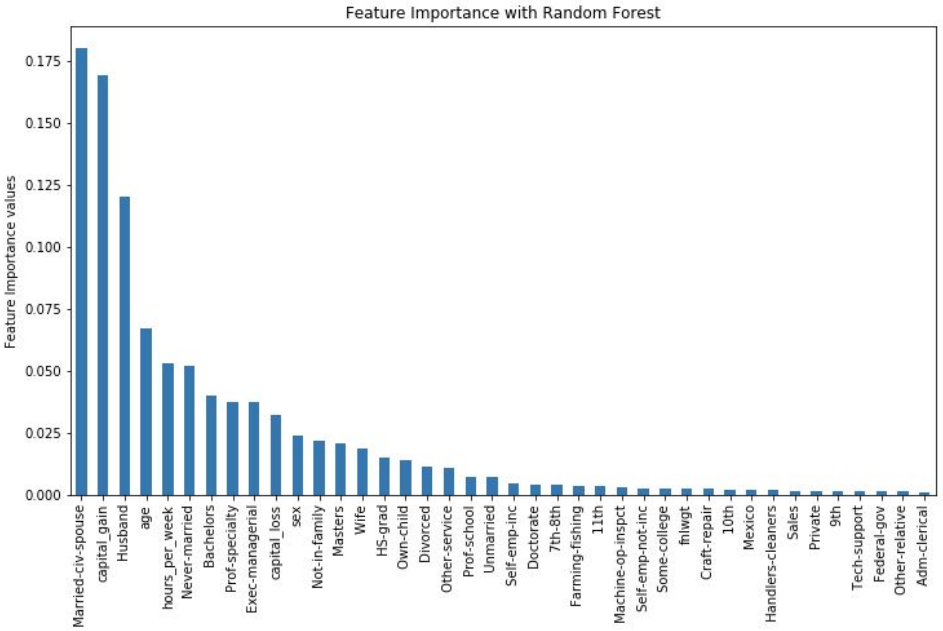
4. 請實作logistic regression的正規化(regularization)，並討論其對於你的模型準確率的影響。

答：



可以看到相較於沒有做regularization，在λ = 0.0001時大概可以得到最好的結果，也就是達到我們想要smoother的目標，而λ更大時準確度開始下降。

5.請討論你認為哪個attribute對結果影響最大？



圖為使用sklearn的RandomForestClassifier做出的feature importance，可以看到capital gain, age有相對大的importance。
所以有試試看取其100個feature以及加上capital gain, age的二次項來做predict，但結果差強人意，在對random forest的性質不太熟悉的情況下，直接取全部的feature還是有最好的performance.