說明：請各位使用此template進行Report撰寫，如果想要用其他排版模式也請註明題號以及題目內容（請勿擅自更改題號），最後上傳至github前，請務必轉成PDF檔，並且命名為report.pdf，否則將不予計分。

----------閱讀完以上文字請刪除----------

學號：B04901000 系級：電機四 姓名：鄭老闆

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

	Private Score	Public Score
Generative	0.77680	0.78439
Discriminative	0.85345	0.85159

ps. Discriminative : normalize the continuous features. Generative : keep original features.

在這個case裡面，我實驗的結果是generative結果明顯比discriminative低，原因我認為是因為generative model是採用training data sample出來的結果去做，而training data數目並不足以讓generative model得approximation估計得很準。

2. 請說明你實作的best model，其訓練方式和準確率為何？

我實作的方式是使用gradient boost classifier, 並且將testing data unseen的category 刪掉，用了150 estimator 以及 90 feature，gradient boost classifier屬於ensemble得model，在這個case裡面其實我試過了sklearn幾乎所有的model，並且發現ensemble的model表現都特別好，例如random forest會到public 0.86*，而gradient boost classifier是我試過第一名好的model, 其他做過keras 的classification tune了很久但大概也只會到0.86左右，都遠低於random forest或者gradient boost這種ensemble的classifier，而gradient boost 在public 會到大概0.877左右、private會到0.874左右。

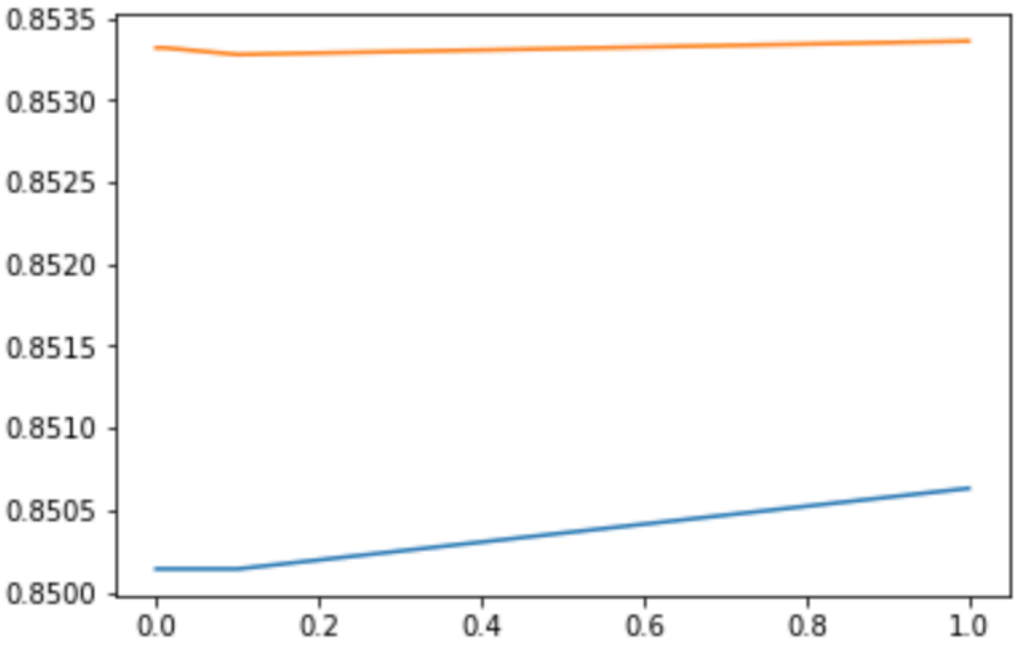
3. 請實作輸入特徵標準化(feature normalization)並討論其對於你的模型準確率的影響

做Discriminative 得normalization因為會通過sigmoid這樣的activation function，所以一開始沒有做會overflow，而做normalization的時候只能對continuous的feature做，不然會讓dummy variable失去原來的意義，而使得model的表現變差。

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。

實作regularization使用這些alpha list，在local端我做了train test split, 大概切了8:2左

右，以下這是我做實驗的結果，橘色是在training set的accuracy, 藍色是在testing set上面的accuracy，可以發現雖然在橘色沒有很明顯的影響，但在testing set上面可已很明顯的看到些微的改善，我們可以知道做regularization可以讓testing set上面的表現更好。



5. 請討論你認為哪個attribute 對結果影響最大？

我將logistic regression的weight依據絕對值得大小排序，在實驗結果裡，影響最大的因素是Separated, 有separated這個feature會讓label <= 50k的機率比較高。

	0	1
0	Separated	-6.068013
1	Prof-school	-3.361553
2	capital_gain	-2.757810
3	Prof-specialty	-2.699768
4	capital_loss	2.355605
5	?_workclass	-2.261133
6	HS-grad	2.248689
7	Married-spouse-absent	2.191312
8	Cuba	-2.171527
9	Married-AF-spouse	-2.101552