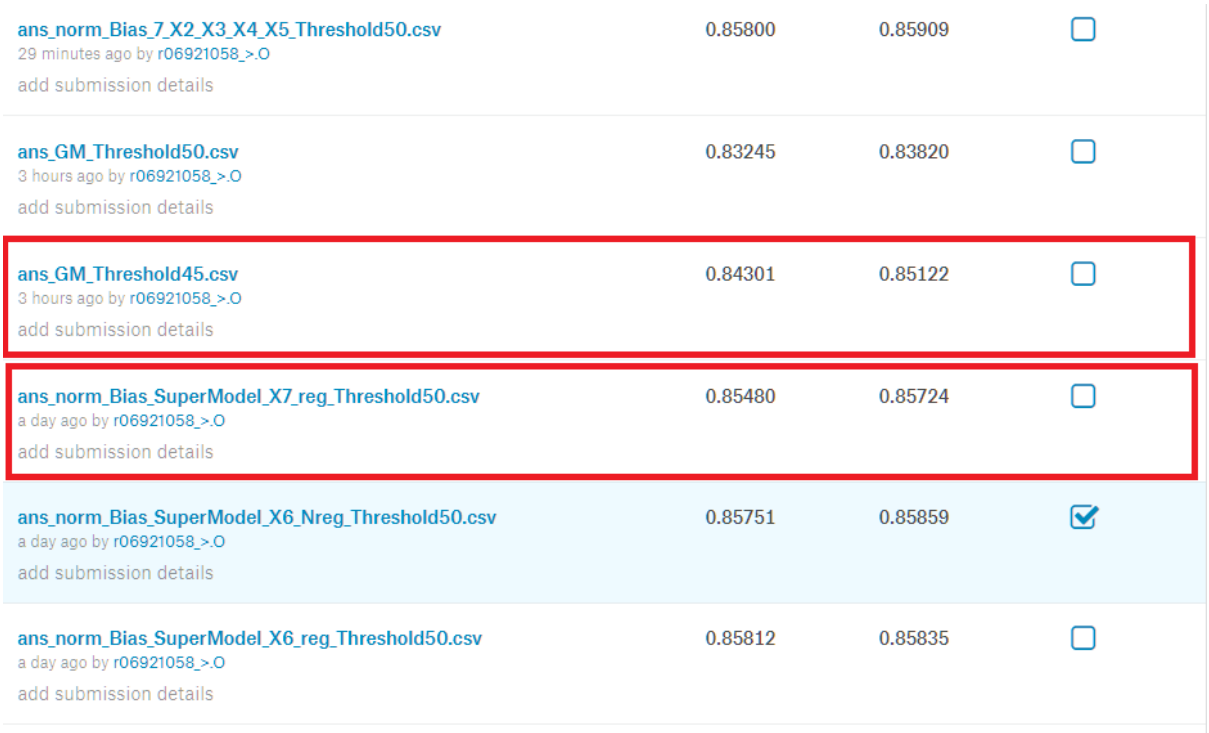
Homework 2 Report - Income Prediction

學號：r06921058 系級：電機碩一 姓名：方浩宇

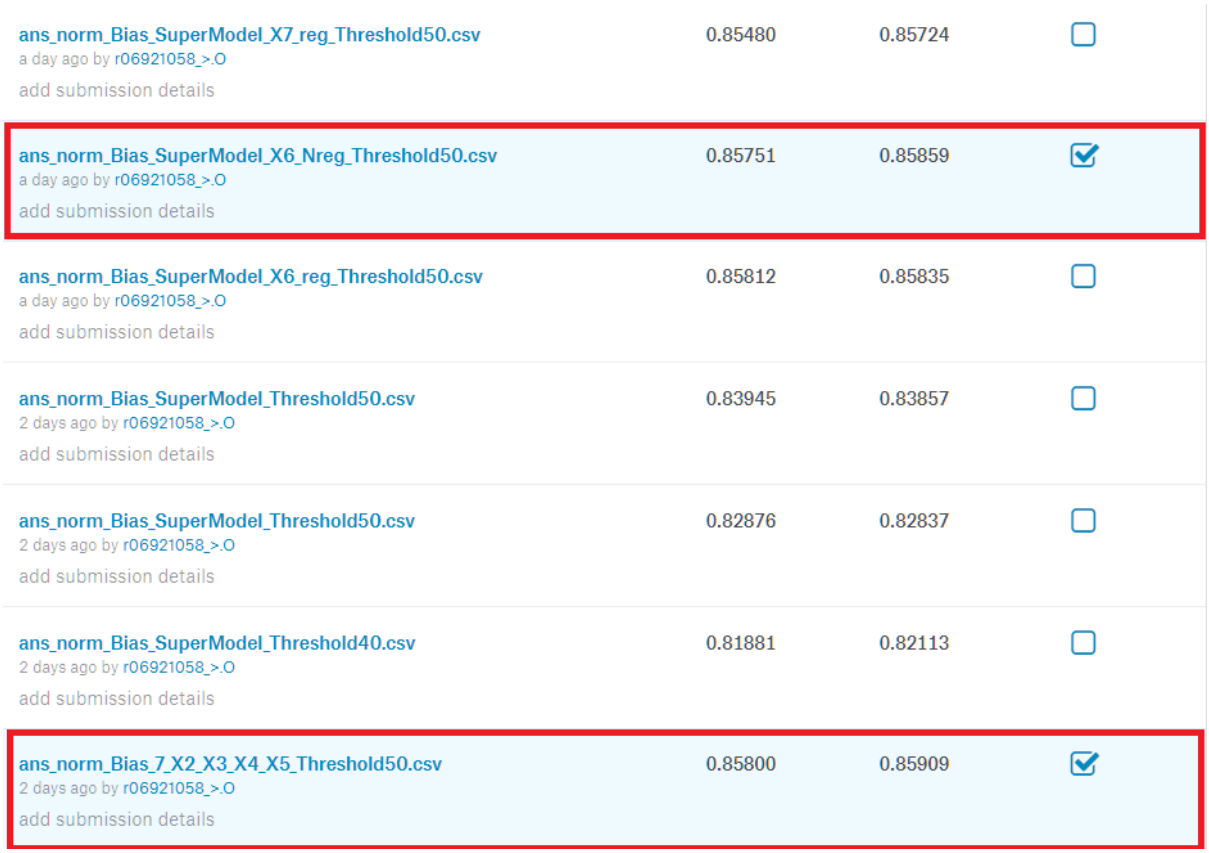
1. (1%) 請比較你實作的 generative model、logistic regression 的準確率，何者較佳？



圖中，上方的為 Generative model，下方為 Logistic Regression。

同樣為將原本的 TrainX 中的 Feature 增加到七次方(將連續性的參數增加高次方項)，並且在 Normalize 中將 fnlwgt 取到接近 0(不採用此變數)，並且根據 Validation(Training Data 直接拿來 Validation)設定 Threshold，Generative model 為 0.45，Logistic Regression 為 0.5。從圖中可以發現 LR 的結果比 GM 好不少(大約好 0.06)，且 private score 更明顯(大約好 0.1)。我認為原因有可能是因為，GM 較適合 Training Data 較少，或者是有較多極端值的情況，然後可能這次作業中，Data 的數量算是足夠，或者是這次作業中極端值較少，因此 LR 占了不少優勢。

2. (1%) 請說明你實作的 best model，其訓練方式和準確率為何？



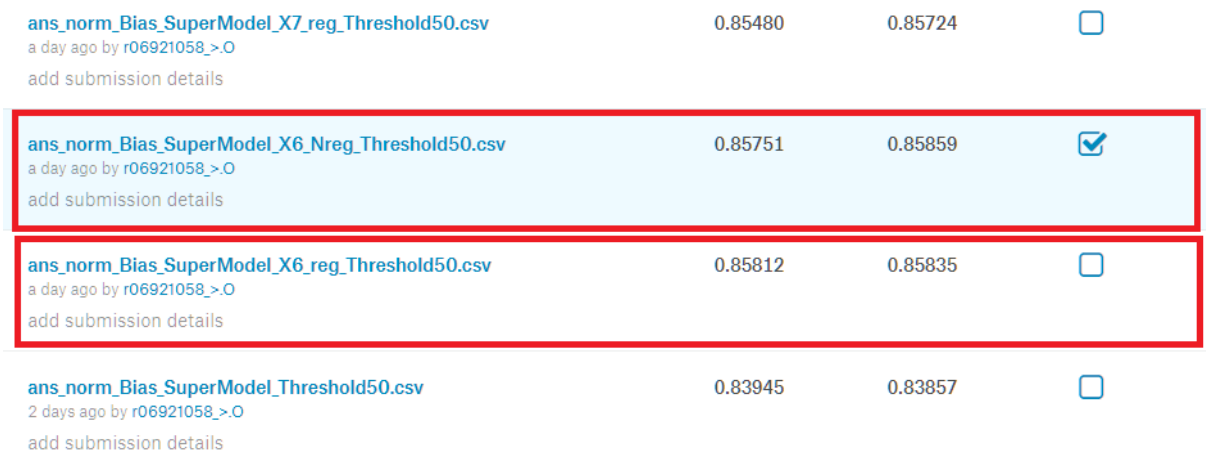
ans_norm_Bias_SuperModel_X7_reg_Threshold50.csv a day ago by r06921058_>.O add submission details	0.85480	0.85724	☐
ans_norm_Bias_SuperModel_X6_Nreg_Threshold50.csv a day ago by r06921058_>.O add submission details	0.85751	0.85859	☑
ans_norm_Bias_SuperModel_X6_reg_Threshold50.csv a day ago by r06921058_>.O add submission details	0.85812	0.85835	☐
ans_norm_Bias_SuperModel_Threshold50.csv 2 days ago by r06921058_>.O add submission details	0.83945	0.83857	☐
ans_norm_Bias_SuperModel_Threshold50.csv 2 days ago by r06921058_>.O add submission details	0.82876	0.82837	☐
ans_norm_Bias_SuperModel_Threshold40.csv 2 days ago by r06921058_>.O add submission details	0.81881	0.82113	☐
ans_norm_Bias_7_X2_X3_X4_X5_Threshold50.csv 2 days ago by r06921058_>.O add submission details	0.85800	0.85909	☑

我所實作的 Best Model 分數如上圖，下方的為 Best，上方為另外一個選擇的 Final Score。我的 Best Model 是採用 Logistic Regression 實作的。有使用 Normalize、Shuffle、Learning Rate 為 0.1 並且執行 20 萬個 Iteration，沒有使用 Regulation，如上圖所示，有 regulation 的效果比較差(上方紅框的下面一筆測資，不過只少大約 0.002，並不明顯)，然後 feature 使用到五次方項(將連續性的參數增加高次方項)，並且不採用 fnlwgt，我在低次方項時(2~3 次)有試著根據 fnlwgt 複製資料，由於 fnlwgt 這個參數代表這項資料所能代表的人數，因此我將 fnlwgt 除 5000 之後將資料複製這個次數，在低次方項時能夠大幅改善結果(大約改善 0.02)，但是在高次方時效果並不明顯，在五次方時分數反而會下降 0,002 左右。而且這個方法會大幅增加運算時間，由於將原本 3 萬 1 千多筆資料複製到 131 萬筆左右，因此運算時間會多至少三十倍，所以我最後並沒有使用這個方法。最後，Private 不管是哪個都有些微下降，推測是因為使用到 5/6 次方會有 Overfitting 的問題。

3. (1%) 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。(有關 normalization 請參考：https://goo.gl/XBM3aE)

如果不使用特徵標準化的話，在訓練方面會變得非常困難，由於 LR 會根據參數大小去移動，而且這次資料中有些特別大，所以會造成需要非常小的 Learning Rate，並且造成這些較大得資料在訓練過程中會佔有優勢，即使他並不是影響力的資料，因此這也會造成最後準確率的下降。

4. (1%) 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。(有關 regularization 請參考：https://goo.gl/SSWGhf P.35)



途中上方為沒有 Regulation，下方為有 Regulation 的結果，在我的模型中使用 Regulation 的結果的確有改善 public score。其中我使用的 Regulation Param 為 1，我有試過 0.1、1、10，不過 1 的結果較好。Regulation 會造成最後結果 private 準確率上升的原因，我推測是因為，根據上課的內容，Regulation 主要是讓曲線變得更平滑，降低 Model 對於 Training Data Overfitting 的情況，所以有可能是因為在我的 Model 中有一些 Over fitting 的情況，因此使用 Regulation 讓結果比較好。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

在這次的 Data Set 中，我認為 age 的影響最大。我有試著將資料逐一取消 Feature，並且和整筆 Training Data 比較，其中取消 Age 會對準確率造成最大的影響。另外，人種的部份只有少數幾個種族有影響，例如白人和黃種人，其他影響較少。學歷也對結果有不少的影響。