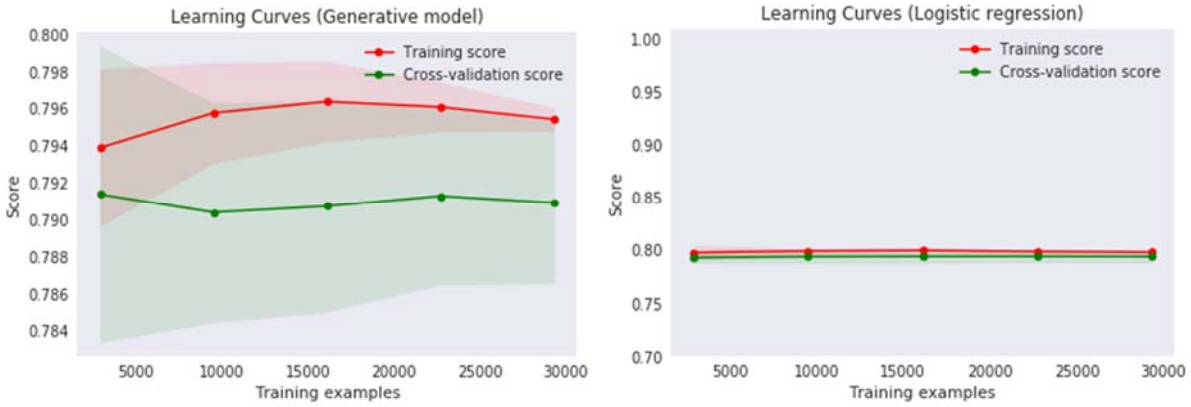
學號：R6946003 系級： 資料科學碩一 姓名：湯忠憲

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

答：



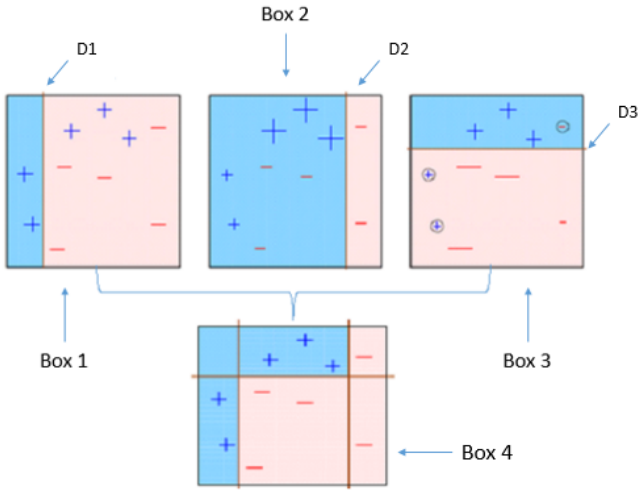
上圖是給定不同 sample 數量對兩種 model 作 cross-validation 後的結果。在特徵均未被篩選以及標準化的情況下，我們可以發現在給定不同 sample 數量下，logistic regression 有較佳得準確率，並且其表現也較 generative model 穩定。因為 generative model 的分佈參數容易受到資料點偏移產生變動，因此不同子集的資料會有較大的分佈、準確度差異。

實際在 Kaggle 上的表現中，logistic model 大約到 0.80068，generative model 則是約 0.84533。Kaggle 的部分 generative model 得到較好的成績。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

我這次使用的 best model 是 Gradient Boosting Tree 的一種 implementation，XGBoost。Kaggle 上蠻多人用的。該算法類似 Random Forest，不過他透過 sequential learning 的方式，補足上一次分類的錯誤(如下圖，Box 2 修正 Box 1，Box 3 修正 Box 2，以此類推)。同時由於該模型實現樹的 Regularization，即使樹的深度加大，他也能夠一定程度的防止 overfitting。



https://www.analyticsvidhya.com/blog/2015/11/quick-introduction-boosting-algorithms-machine-learning/

準確度的部分可以來到 0.8765。Fine tune 參數(depth = 5 or 6, estimators = 100, col_sample = 0.5 or 0.8) 可以拿到更好的成績。最終 private 準確度為 0.87923。

Reference : https://xgboost.readthedocs.io/en/latest/

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：

	標準化前	標準化後
Logistic Regression	0.80068	0.85393
Generative Model	0.84533	0.84227

上表為對兩個模型做標準化後的準確率，我們可以發現 Logistic Regression 的準確率得到了很大的進步，並且在 training 的過程中也不易發生 overflow 的情況。Generative Model 的部分則沒有明顯的差異。

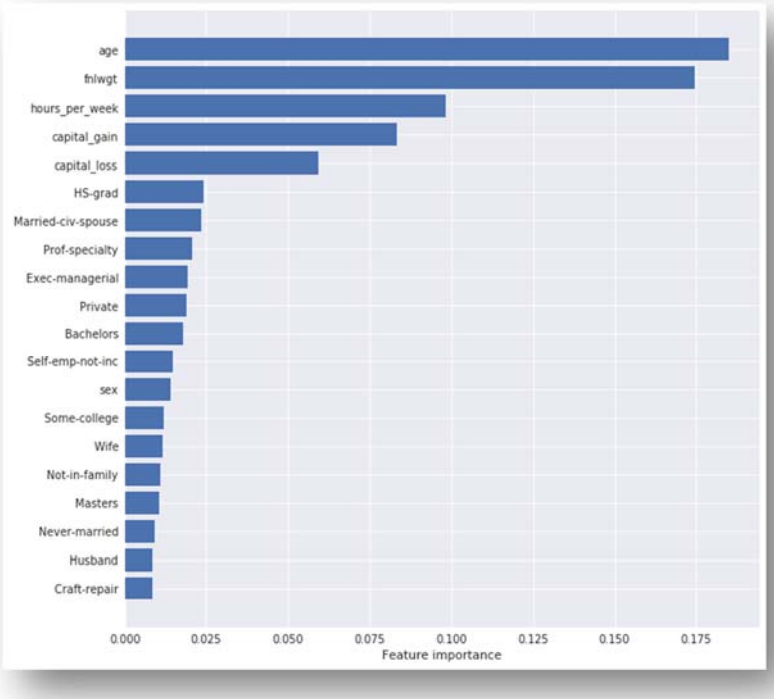
4. 請實作 logistic regression 的正規化(regularization)，並討論其對於你的模型準確率的影響。

答：

Regularization	1	0.1	0.01	0.001	0
Accuracy	0.85073	0.85405	0.85343	0.85393	0.85393

從上表我們可以看到對 Logistic regression 在這個 dataset 中並沒有比較好的效果。若仔細去看，可以發現在沒有做 regularization 的情況下 weight 的其實都沒有到很大。我想這是為什麼 regularization 在這裡沒有特別顯著的差異。若給定太大的 penality，反而還會讓準確率下降。

5.請討論你認為哪個 attribute 對結果影響最大？



由於 XGBoost 是一種 tree base 的演算法，因此我們可以觀察 information gain 來判斷哪些特徵是重要的。從上圖我們可以發現年齡跟 fnlwgt 十分有用。其餘像是 hours_per_week、capital_gain、capital_loss 等特徵也很重要!