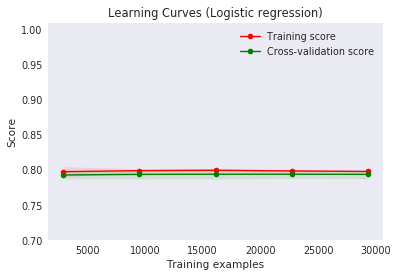
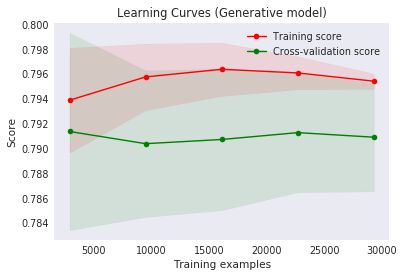
學號：R6946003 系級： 資料科學碩一 姓名：湯忠憲

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

答：



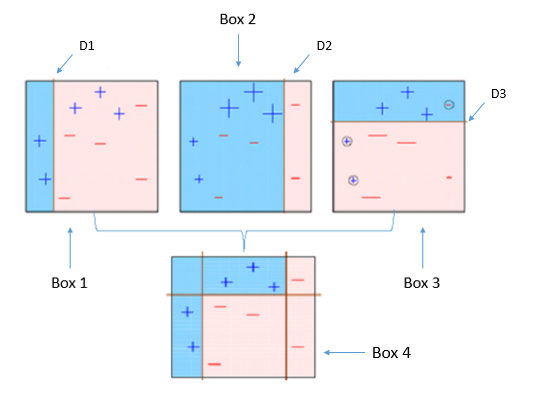
上圖是給定不同sample數量對兩種model作cross-validation後的結果。在特徵均未被篩選以及標準化的情況下，我們可以發現在給定不同sample數量下，logistic regression 有較佳得準確率，並且其表現也較generative model穩定。因為generative model的分佈參數容易受到資料點偏移產生變動，因此不同子集的資料會有較大的分佈、準確度差異。

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。

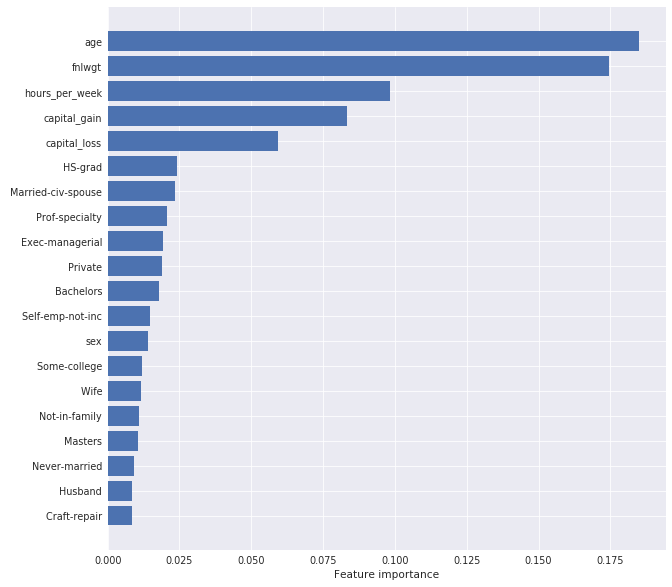
4. 請實作logistic regression的正規化(regularization)，並討論其對於你的模型準確率的影響。

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!