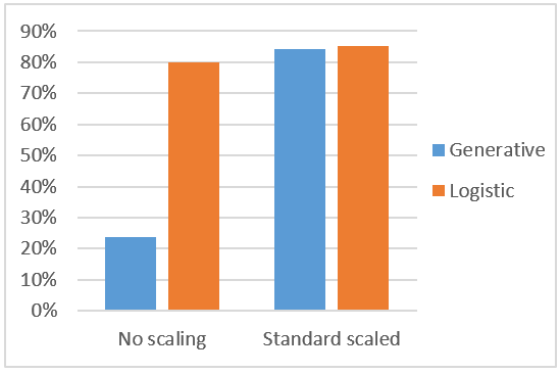
學號：B04901060 系級： 電機三 姓名：黃文璁

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

	Scaled (normalized)?	public	private	average
Generative	No scaling	0.23476	0.23768	0.23622
	Standard scaled	0.84520	0.84252	0.84386
Logistic	No scaling	0.80233	0.79695	0.79964
	Standard scaled	0.85393	0.85112	0.85253



在沒有 Standard scale (normalize) 的情況下，logistic 的準確度遠高於 generative model。

在有 Standard scale (normalize) 的情況下，logistic 和 generative model 的準確度較接近，但 logistic 還是稍微好一些。故**整體來說，logistic regression 的準確度較好**

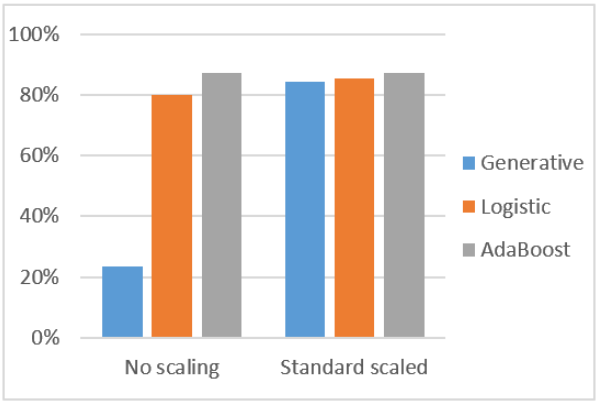
2.請說明你實作的 best model，其訓練方式和準確率為何？

本次 best model 使用 scikit-learn 中的 AdaBoostClassifier。在 Kaggle 上的 public 和 private 分數分別為：0.87174 和 0.87163，最終排名為：59/402。

AdaBoost 是 Adaptive Boosting 的縮寫，是一種 meta-algorithm 故可以套用在許多 ML model 上，一般來說套用 decision tree。此演算法基本概念是用前一個 classifier 分錯的樣本來訓練下一個 classifier。AdaBoost 對離群值較敏感，某些情況下比一般演算法更不容易 overfitting。

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。

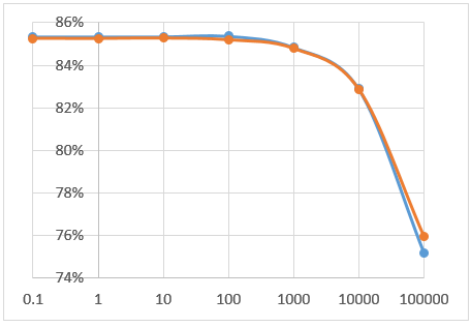
	Scaled (normalized)?	public	private	average
Generative	No scaling	0.23476	0.23768	0.23622
	Standard scaled	0.84520	0.84252	0.84386
Logistic	No scaling	0.80233	0.79695	0.79964
	Standard scaled	0.85393	0.85112	0.85253
AdaBoost (n_estimators = 1000)	No scaling	0.87272	0.86893	0.87083
	Standard scaled	0.87272	0.86893	0.87083



Generative model 在沒有 normalize 的情況下準確度非常差，但 normalize 後準確度就上升到和 logistic 接近的水準。整體來說 **generative model 和 logistic regression 在 normalize 後準確度都有提升。**而本次在 best model 中使用的 **AdaBoost 則不受 normalize 影響**，這是由於 scikit-learn 中的 AdaBoostClassifier 預設使用的 estimator 是 DecisionTreeClassifier，而 decision tree 較不受 normalize 影響結果。

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。

λ	training	public	private	average
0.1	0.85332	0.85393	0.85124	0.85259
1	0.85335	0.85393	0.85112	0.85253
10	0.85338	0.85442	0.85100	0.85271
100	0.85366	0.85429	0.84952	0.85191
1000	0.84838	0.84926	0.84645	0.84786
10000	0.82884	0.83292	0.82446	0.82869
100000	0.75200	0.76326	0.75592	0.75959



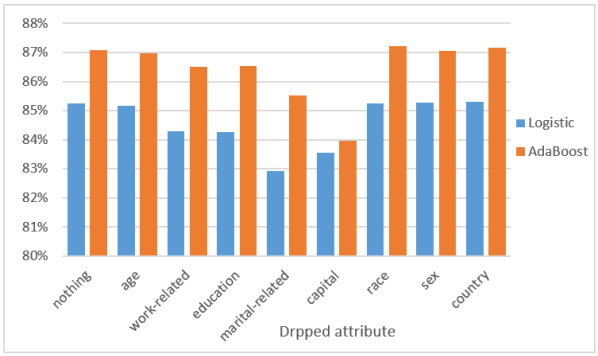
由圖表觀察出 regularization 的效果不佳。

注意到 logistic regressoin 的 Kaggle 分數和 training 時的 accuracy 都很接近，也就是 logistic regression 並沒有嚴重的 overfitting 狀況，故 regularization 的效果也就較不明顯。

5.請討論你認為哪個 attribute 對結果影響最大？

分別利用 logistic 和 AdaBoost 兩個模型，每次 drop 一個 attribute，比較 Kaggle 分數：

Drop	Logistic			AdaBoost		
	public	private	average	public	private	average
nothing	0.85393	0.85112	0.85253	0.87272	0.86893	0.87083
age	0.85233	0.85124	0.85179	0.87137	0.86819	0.86978
work-related	0.84336	0.84227	0.84282	0.86609	0.86426	0.86518
education	0.84692	0.83859	0.84276	0.87014	0.86033	0.86524
marital-related	0.82972	0.82901	0.82937	0.85859	0.85210	0.85535
capital	0.83660	0.83454	0.83557	0.84004	0.83945	0.83975
race	0.85343	0.85173	0.85258	0.87272	0.87163	0.87218
sex	0.85429	0.85124	0.85277	0.87137	0.86954	0.87046
country	0.85454	0.85149	0.85302	0.87334	0.87004	0.87169



其中 work-related 是同時 drop ‘workclass’ 和 ‘occupation’。

而 marital-related 是同時 drop ‘marital’ 和 ‘relationship’

兩種模型中 attribute drop 之後影響最大的都為 marital + relationship 和 capital gain + loss，故可以歸納出：marital + relationship 和 capital gain + loss 影響預測結果最大。

此外 workclass + occupation 和 education 也有較明顯的影響。