學號：R06922130　系級：資工碩一 姓名：葉韋辰

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

- *Generative：normalize 所有 features 一次項所得到的結果*
- *Logistic：normalize 所有 features 一次項（並將加入 age, capital gain, hours-per-week 之二次項）、batch=32、epoch =1000、SGD 所得到的結果*

類別	PRIVATE 分數	PUBLIC 分數	平均
GENERATIVE	0.84227	0.84508	0.843675
LOGISTIC	0.84719	0.85319	0.85019

從表中可知 logistic regression 在 private 與 public 分數皆優於 generative function，推測可能是 generative 並無 train 之過程，故效果稍差。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

- *下述實作採 Keras，normalize(min-max)所有 features 一次項、batch=128、epoch =1000、sigmoid、2 layers、12 neurons 所得到的結果：*

類別	PRIVATE 分數	PUBLIC 分數	平均
KERAS(一次項)	0.84633	0.85847	0.8524
KERAS(二次項)	0.84964	0.85552	0.85258

上述版本並沒有通過 private strong baseline，其實過程中其實有嘗試導入二次項，但 public 表現相對差決定捨棄二次項，deadline 後再度悲劇，體會 public 與 private 差異。

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。

- *Generative：normalize(std-mean)所有 features 一次項所得到的結果*

GENERATIVE	PRIVATE 分數	PUBLIC 分數	平均
NORMALIZATION	0.84227	0.84508	0.843675
WITHOUT	0.84191	0.84520	0.843555

從表中可以發現對於 Generative 而言，標準化有無並不會造成太大影響，畢竟求解時並非採 training 的方式，而是藉由平均值、變異數矩陣求解。

- *Logistic：normalize 所有 features 一次項（並將加入 age, capital gain, hours-per-week 之二次項）、batch=32、epoch =1000、SGD 所得到的結果*

LOGISTIC	PRIVATE 分數	PUBLIC 分數	平均
NORMALIZATION	0.84719	0.85319	0.85019
WITHOUT	0.77914	0.78378	0.78146

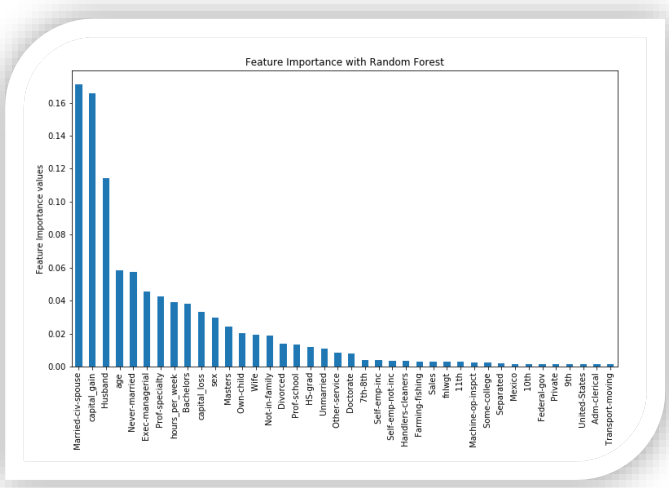
從表中可以發現對於 Logistic 而言，標準化有無影響準確率甚鉅，值域較大的 features 將會對結果有更直接的影響，這也是 training 典型的特性。

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。

LAMBDA	PRIVATE 分數	PUBLIC 分數	平均
0	0.84719	0.85319	0.85019
0.1	0.82520	0.82813	0.826665
0.01	0.84191	0.84324	0.842575
0.001	0.84633	0.85196	0.849145
0.0001	0.84891	0.85368	0.851295
0.00001	0.84879	0.84815	0.84847

總體而言，適當選取 lambda 值的情況下，增加 regularization 確實能有效提升準確度，需要透過數次測試，找出較佳的 lambda 值。

5. 請討論你認為哪個 attribute 對結果影響最大？



直觀地判斷，我認為 age, capital gain 與 hours-per-week 這三項 features (attribute)對結果影響較大：隨年齡增長收入將隨著上升；而資本所得愈高亦蘊含收入之增加。且 age 與 capital gain 之重要性亦與分析結果吻合(參見上圖)。