1.
a.feature engineering

這個部分我觀察Readme文件的說明，把EDUCATION類的0,4,5,6都map成0，然候用xgboost套件做出feature importance，發現第一名是PAY_1，二三名分別是BILL_AMT1, PAY_AMT1, 最後一名是SEX，所以我先將PAY_1做one_hot_encoding，而其他feature的調整則根據model建完之後調整，調整使用的演算法為cross validation，使用套件為sklearn的StratifiedKFold。

b.建構model

model的部分我選擇的是keras的Deep neuron network，原因為我測試了randomforest, xgboost的結果都沒有keras來得優秀，只是keras的缺點就是每次跑的結果都會不太一樣，所以有時候會特別優秀但有時候就會特別差，為了降低隨機性我用了np.random.seed()和tensorflow的set_random_seed()，但還是無法使得每次的輸出結果都一樣。模型的建立上我使用GridSearchCV來調整我的參數，並使用keras套件裡面的ModelCheckpoint來輸出訓練中最好的model。

2.在輸入層傳入30個input(features)，第一層有200個units，第二層有175個units，第三層有75個units，最後是輸出層只有一個units(probability)，並在第二層和第三層添加Dropout(0.3)的項，其中hidden layer的activation function為relu，而輸出層使用sigmoid。這些參數的生成是由GridSearchCV得出來的，而model好壞的評估的標準為KFold得出來的。

3.從一開始的preprocessing到最後的feature engineering 中間做了很多feature的調整，應該說kaggle最花時間的就是這塊，最後得到的結果為將PAY_1的7,8,9都map到6之後做one_hot_encoding，做完之後再把PAY_1_6這欄刪除降低col之間的線性關係。之後再將所有的feature歸一化

4.使用的套件為keras, panda, numpy, sklearn