ML Project

電機三 B04901081
鄭元嘉

Algorithm

本次採用「RandomForest」作為分類器。原因如下：

(1)利用人力方式閱覽Train.csv中的資料，原先想利用NN Model作為演算法，但是由於欠債的數據如果用normalization的話非常難以調整，再加上金融或是業界資料，因為是人性化的資料，一般會認為其分類是由較為簡單分割方式，不需要用到NN Model這種對於抽象或難以分割的資料較為有效的演算法。

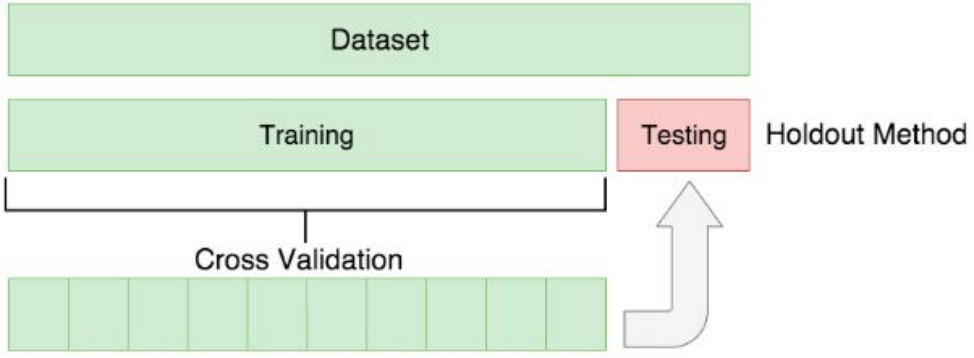
(2)在非NN Model的分類器，我直觀認為此數據必定存在某人容易切割的組合，所以選擇Decision Tree的強化版RandomForest。

(3)在非NN Model的分類器，選擇RandomForest還有另一個重要原因，那就是該演算法不需要事先的normalization，可以讓我省去優化normalization的時間去調配其他參數。

Validation

由於這次的評分標準是採用「Average precision」，所以在做模型hyperparameters校正時，是手動刻了一個評分器，而不是用一般的accuracy。

本次的cross validation分兩個部份



在整個Dataset中取出一份最為最終Testing評估用，剩下的部份再下去做k-folds cross validation找出最合適的hyperparameters，調整完畢後，再利用全體的Dataset用已經調整好的hyperparameters train出最終的model。

Important Parameters

(1)Information Gain Criterion :

定義impurity measurement有兩大類gini還有entropy。

在實務上，並沒有孰優孰劣，所以利用暴力測試，丟各種相同參數並且只更換Criterion，再利用上述的cross validation找出兩者最後entropy在較多case中有較好的表現。

(2)Number of trees :

利用網格搜尋法把數量從10~600每次增加10的調高並且測試不同參數下的結果，可以發現約在30~60左右會有最好的效果，可以掌握一定的票數比例，又可以避免overfitting的問題，算是常見的中庸選擇。

(3)Minimun impurity to split :

此參數表示，如果某個node的impurity小於該參數那麼就不要在這個node上繼續切割，可以避免overfitting。這個值非常難抓，會因為training data的分佈不同而有很大差異，所以我用cross validation網格搜尋大範圍的調整在慢慢小範圍調整找到以此題目而言，0.0002算是不錯的值。

(4)Minimun samples to split :

此參數表示，如果某個node的samples小於該參數那麼就不要在這個node上繼續切割，可以避免overfitting。抓取方式和上面相同。

(5)Maximun depth :

此參數表示，如果某個brach的depth大於該參數那麼就不要在增加深度，可以避免overfitting。抓取方式和上面相同。

Data Preprocessing

(1)去除sex :

針對sex，利用人力方式去評估，認為或許此條件不具有重要性，因此嘗試將sex去除，意外的發現確實結果大多有較好的結果，一方面可能這個feature本身的重要性或許真的不高，另一方面也有可能是減少feature可以減少overfitting的發生。

(2)增加PAY_1、P_AMT1和BILL各別的平均與標準差：

直觀上這3個原本的features本身的數值帶的意義有限，或許這些數值的波動性（標準差）和平均值能夠帶來更多的有效features。結果顯示，增加這些features確實能夠增加accuracy，但卻也帶來更強的overffiting，反而有時候validation cross的評分會變差，最終在經過非常多次的測試後取消使用。

(3)將Education改成one-hot encoding :

因為認為這個features非常的離散，數值間沒有太大關係，故使用one-hot encoding，結果顯示表現的確實好上許多。