ML Assignment 1 – Data Preprocessing

309709022 陳政廷

1) Null (or Missing) Values Estimation

首先針對兩資料集當中每個欄位進行缺失值探找並個別統計總和，接著計算缺失值欄位個別占總資料的筆數，若超過 50%將刪除該欄位，表示缺失值過多，進行補值動作將可能誤導後續模型之結果，所幸兩資料集缺失值欄位占比平均落在 1.5%左右。在 diamonds 資料集當中發現目標變數”price”亦有缺失之情形，由於筆數約占總資料集 1%左右故在此將缺失之樣本移除後，重新進行針對其他欄位補值。兩資料集在補值方法均使用 miceforest，而 MICE(Multiple Imputation by Chained Equations)，為多重插補的方法之一，MICE 假設缺失的資料是隨機的，接著針對每個缺失值產生一個插補資料集，接著整合各資料集之結果形成完整資料集，透過結合隨機森林減少過多 hyperparameters 之調整並能處理資料中的非線性關係，達到更佳效果。

2) Data Balance

進行 resample 前應切分訓練與測試，而資料平衡將針對訓練資料集使用，在此使用 SMOTE+ENN 的方式解決原本資料不平衡的現象(原先分別為 0 為 5277 筆、1 為 179 筆，經調整後 0 為 4359 筆、1 為 5011 筆)

3) Feature Selection

因不希望主觀認定留存個數，故使用 Selectfrommodel 進行此任務，其中

diamonds 資料集 model 為 LassoCV，透過 Lasso 採用 L1 正則化結合交叉驗證，解決資料稀疏性的問題。最後 diamonds 資料集當中剩餘 8 個 feature(原本 9 個)，其中 threshold 由於資料本身欄位較少不希望刪除過多，故採用預設值，而 bankruptcy 的 model 則使用 Logistic regression 剩餘 13 個 feature(原本 95 個)，其中 threshold 設定為 mean，希望取出其參數值大於整體欄位之平均以達到較好的篩選效果。

4) Export

分別將 X_train 與 y_train 合併，X_test 與 y_test 合併並分別輸出成 excel 與 csv。