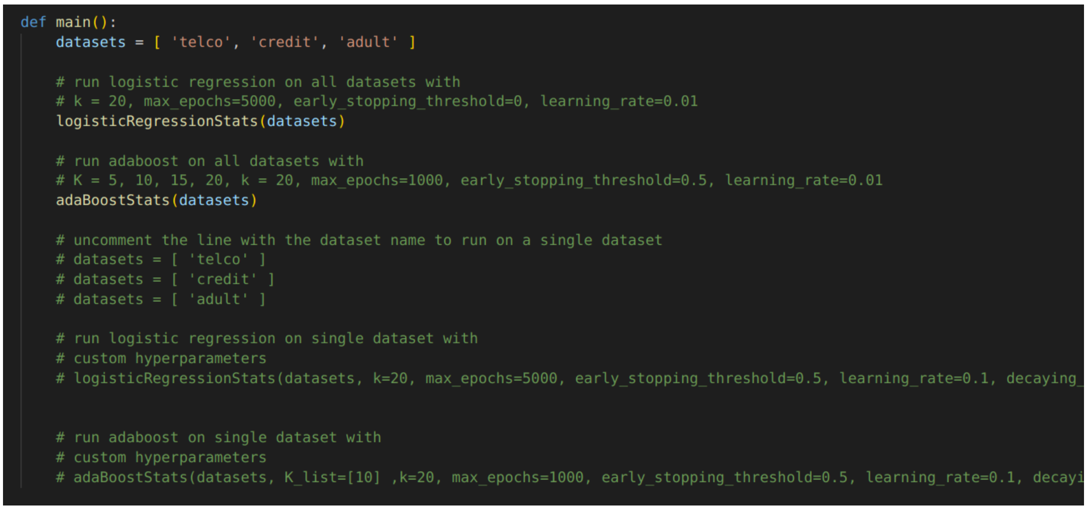
CSE472 : Machine Learning

Date: 9 December 2023

Submitted by:
Kazi Ababil Azam
Student No: 1805077
CSE, L-4/T-2, B1

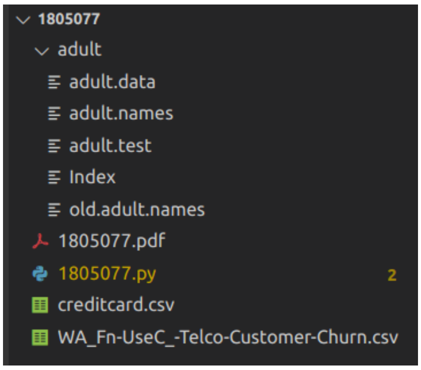
Instructions to train and test the models:

1. Please run **1805077.py** to run all experiments and get required metrics as mentioned in specifications. It will print the metrics of the datasets first trained and tested using only logistic regression and AdaBoost sequentially. It may take some time.



```
def main():
    datasets = [ 'telco', 'credit', 'adult' ]

    # run logistic regression on all datasets with
    # k = 20, max_epochs=5000, early_stopping_threshold=0, learning_rate=0.01
    logisticRegressionStats(datasets)

    # run adaboost on all datasets with
    # K = 5, 10, 15, 20, k = 20, max_epochs=1000, early_stopping_threshold=0.5, learning_rate=0.01
    adaBoostStats(datasets)

    # uncomment the line with the dataset name to run on a single dataset
    # datasets = [ 'telco' ]
    # datasets = [ 'credit' ]
    # datasets = [ 'adult' ]

    # run logistic regression on single dataset with
    # custom hyperparameters
    # logisticRegressionStats(datasets, k=20, max_epochs=5000, early_stopping_threshold=0.5, learning_rate=0.1, decaying_


    # run adaboost on single dataset with
    # custom hyperparameters
    # adaBoostStats(datasets, K_list=[10] ,k=20, max_epochs=1000, early_stopping_threshold=0.5, learning_rate=0.1, decayi
```

2. If it is required to run on a specific dataset with specific hyperparameters, some functions are commented out in the **main()** function [Line 416]. Please uncomment according to whichever dataset is necessary and whichever model is necessary to run any of the models (LR, AdaBoost) on that specific dataset and view performance metrics.
3. Please uncomment only the required function calls for decreasing runtime.
4. Please keep the dataset files in the same directory as the 1805077.py. The file names should be the original file names.



```
1805077
  adult
    adult.data
    adult.names
    adult.test
    Index
    old.adult.names
  1805077.pdf
  1805077.py                          2
  creditcard.csv
  WA_Fn-UseC_-Telco-Customer-Churn.csv
```

Performance measure of logistic regression:

Feature Selection = 20, Epochs = 5000, Learning Rate = 0.01 (Constant), No early stopping

Telco Churn Dataset:

Performance Measure	**Training**	**Test**
Accuracy	0.7949946751863685	0.78708303761533
Recall	0.5204013377926422	0.49732620320855614
Specificity	0.8941773375211404	0.8917874396135266
Precision	0.6398026315789473	0.6241610738255033
False Discovery Rate	0.36019736842105265	0.37583892617449666
F1 score	0.5739579490962745	0.5535714285714286

Credit Card Fraud Dataset:

Performance Measure	**Training**	**Test**
Accuracy	0.9943878484719088	0.9978043425225664
Recall	0.7893401015228426	0.9081632653061225
Specificity	0.9994374648415526	1.0
Precision	0.971875	1.0
False Discovery Rate	0.028125	0.0
F1 score	0.8711484593837535	0.9518716577540107

Adult Census Dataset:

Performance Measure	**Training**	**Test**
Accuracy	0.8365824547443803	0.8356573705179283
Recall	0.5543420351624934	0.547027027027027
Specificity	0.9301227156352079	0.9296654929577465
Precision	0.724456048738033	0.7169677647892313
False Discovery Rate	0.2755439512619669	0.2830322352107687
F1 score	0.6280842073492794	0.6205733558178752

Performance measure of AdaBoost implementation:

Feature Selection = 20, Epochs = 1000, Learning Rate = 0.01 (Constant),
Early Stopping Threshold = 0.5

Telco Churn Dataset:

Number of boosting rounds	**Training**	**Test**
5	0.792332268370607	0.7977288857345636
10	0.7754703585374512	0.7913413768630234
15	0.77209797657082	0.7913413768630234
20	0.7669506567270146	0.7806955287437899

Credit Card Fraud Dataset:

Number of boosting rounds	**Training**	**Test**
5	0.9941438418837308	0.9975603805806295
10	0.9697431830659428	0.9731641863869236
15	0.9164277435490759	0.9119297389607222
20	0.863905325443787	0.8665528177604294

Adult Census Dataset:

Number of boosting rounds	**Training**	**Test**
5	0.8268350905112393	0.8275564409030545
10	0.8233538889994032	0.8214475431606906
15	0.8271666335123665	0.8262284196547145
20	0.8301836748226245	0.8275564409030545

Observations:

1. The AdaBoost accuracy declines with the increase of the boosting rounds. This may be due to the fact that the ensemble overfits with more hypotheses, as the datasets are all negatively biased.

2. The negative bias of the datasets also impact the specificity of the models, as can be seen in the logistic regression analysis.

3. Preprocessing with bins has been used for continuous data.