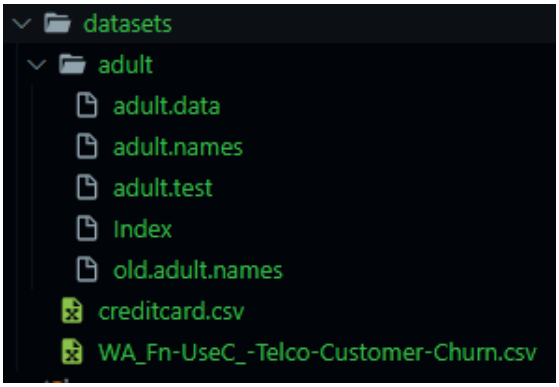
CSE472: Machine Learning

Date: 9 December 2023

Submitted by:
Hasan Masum
Student No: 1805052
CSE, L-4/T-2, A2

Instructions to train and test the models:

Step 1: Download the daset.
Download the datasets. Create a folder named datasets. Copy the datasets files in the datasets directory.(same directory as 1805052.py)



Step 2: Install Dependencies
pip install scikit-learn
pip install pandas

Step 3: Run python file

- **To run all the datasets**
  python 1805052.py

- **To run on telco dataset**
  python 1805052.py -dataset telco

- **To run on adult dataset**
  python 1805052.py -dataset adult

- **To run on credit dataset**
  python 1805052.py -dataset credit

Performance measure of logistic regression:

Epochs = 2000, Learning Rate = 0.05 (Constant), No early stopping

Telco Churn Dataset: Number of Features (80%)

Performance Measure	**Training**	**Test**
Accuracy	0.8039	0.8003
Recall	0.5473	0.5312
Specificity	0.8972	0.8960
Precision	0.6594	0.6447
False Discovery Rate	0.3406	0.3553
F1 score	0.5982	0.5825

Credit Card Fraud Dataset: Number of Features (80%)

Performance Measure	**Training**	**Test**
Accuracy	0.9953	0.9939
Recall	0.8165	0.8190
Specificity	0.9996	0.9990
Precision	0.9776	0.9596
False Discovery Rate	0.0224	0.0404
F1 score	0.8882	0.8837

Adult Census Dataset:

Performance Measure	**Training**	**Test**
Accuracy	0.8483	0.8471
Recall	0.6083	0.6001
Specificity	0.9244	0.9235
Precision	0.7184	0.7082
False Discovery Rate	0.2816	0.2918
F1 score	0.6588	0.6497

Performance measure of AdaBoost implementation:

Epochs = 1000,
Learning Rate = 0.05 (Constant),
Early Stopping Threshold = 0.5

Telco Churn Dataset:

Feature Selection = All

Number of boosting rounds	**Training**	**Test**
5	0.7845	0.7740
10	0.7705	0.7676
15	0.7714	0.7690
20	0.7641	0.7662

Credit Card Fraud Dataset:

Feature Selection = All

Number of boosting rounds	**Training**	**Test**
5	0.9706	0.9656
10	0.9900	0.9846
15	0.9894	0.9856
20	0.9881	0.9859

Adult Census Dataset:

Feature Selection = 20

Number of boosting rounds	**Training**	**Test**
5	0.8232	0.8235
10	0.8266	0.8284
15	0.8279	0.8272
20	0.8272	0.8256

Observations:

- Data preprocessing has a great impact on performance matrices
- Data normalization and one hot encoding are used to improve the performance of matrices
- Performance is improved with the increase in the round for Adaboost for the credit card and adult census datasets.