Dataset 1 - Churn Prediction

Logistic Regression Model

Performance Measure	Training	Test
Accuracy	80.547 %	79.276 %
True positive rate (sensitivity, recall, hit rate)	52.949 %	53.553 %
True negative rate (specificity)	90.334 %	89.261 %
Positive predictive value (precision)	66.018 %	65.937 %
False discovery rate	33.981 %	34.063 %
F1 score	58.766 %	59.104 %

Adaboost Model

Number of boosting rounds	Training	Test
5	80.334 %	80.057 %
10	80.334 %	79.985 %
15	80.331 %	79.985 %
20	80.334 %	79.985 %

Dataset 2 - Adult Income Dataset

Logistic Regression Model

Performance Measure	**Training**	**Test**
Accuracy	85.028 %	85.111 %
True positive rate (sensitivity, recall, hit rate)	57.888 %	57.514 %
True negative rate (specificity)	93.637 %	93.647 %
Positive predictive value (precision)	74.264 %	73.684 %
False discovery rate	25.736 %	26.315 %
F1 score	65.061 %	64.603 %

Adaboost Model

Number of boosting rounds	**Training**	**Test**
5	82.943 %	82.489 %
10	83.468 %	82.673 %
15	83.465 %	82.673 %
20	83.468 %	82.674 %

Dataset 3 - Credit Card Fraud Dataset(10k negative samples and all positive samples)

Logistic Regression Model

Performance Measure	**Training**	**Test**
Accuracy	96.949 %	96.903 %
True positive rate (sensitivity, recall, hit rate)	35.051 %	67.500 %
True negative rate (specificity)	99.950 %	100.000 %
Positive predictive value (precision)	97.143 %	100.000 %
False discovery rate	2.857 %	0.000 %
F1 score	51.515 %	54.545 %

Adaboost Model

Number of boosting rounds	**Training**	**Test**
5	95.686 %	95.188 %
10	95.686 %	95.188 %
15	95.686 %	95.188 %
20	95.686 %	95.188 %

Note: The performance evaluation were reported with the following parameters:

1. **Logistic Regression**
    a. **Learning Rate:** 0.05
    b. **Max epochs:** 1000
    c. **Early Stop:** 0.0
2. **Adaboost**
    a. **Weak Learning Rate:** 0.5
    b. **Weak Learner epochs:** 100
    c. **Early Stop Error Threshold:** 0.5

Script Run Instruction:

1. Before running the script the dataset file path should be specified in the main function of the code.
   For example:

```
if __name__ == '__main__':
   # dataset filepath needs to be specified here
   dataset_1_file_path = "dataset1/WA_Fn-UseC_-Telco-Customer-Churn.csv"
   dataset_2_train_file_path = "dataset2/adult.data"
   dataset_2_test_file_path = "dataset2/adult.test"
   dataset_3_file_path = "dataset3/creditcard.csv"
```

2. Then for preprocessing different dataset just comment out the specified dataset preprocessing function
   For example if you want to run the dataset1 then just comment out the below line

```
# preprocess working here
X_train, X_test = preprocess_dataset1(dataset_1_file_path)
# X_train, X_test = preprocess_dataset2(dataset_2_train_file_path,
dataset_2_test_file_path)
```

3. Lastly run **python3 1605084.py** and you can see the results printed on the console also

Observations

- The third dataset was heavily skewed with only 492 positive samples. So 10000 negative samples are taken since the dataset is heavily imbalanced. Else the result is too underperforming as observed in experiments.

- The first two datasets are also imbalanced but not as much as the third one.