Dataset 1 - Churn Prediction

Logistic Regression Model

Performance Measure	Training	Test
Accuracy	79.961 %	79.986 %
True positive rate (sensitivity, recall, hit rate)	54.118 %	53.687 %
True negative rate (specificity)	89.596 %	88.318 %
Positive predictive value (precision)	65.976 %	59.283 %
False discovery rate	34.024 %	40.717 %
F1 score	59.461 %	56.347 %

Adaboost Model

Number of boosting rounds	Training	Test
5	78.772 %	77.928 %
10	78.665 %	77.644 %
15	78.594 %	78.070 %
20	78.559 %	77.928 %

Dataset 2 - Adult Income Dataset

Logistic Regression Model

Performance Measure	Training	Test
Accuracy	84.985 %	85.013 %
True positive rate (sensitivity, recall, hit rate)	59.023 %	58.580 %
True negative rate (specificity)	93.220 %	93.189 %
Positive predictive value (precision)	73.414 %	72.677 %
False discovery rate	26.586 %	27.323 %
F1 score	65.437 %	64.872 %

Adaboost Model

Number of boosting rounds	Training	Test
5	82.940 %	82.642 %
10	82.912 %	82.630 %
15	82.949 %	82.673 %
20	83.093 %	82.747 %

Dataset 3 - Credit Card Fraud Dataset

Logistic Regression Model

Performance Measure	Training	Test
Accuracy	96.813 %	95.814 %
True positive rate (sensitivity, recall, hit rate)	62.933 %	60.684 %
True negative rate (specificity)	99.975 %	100.000 %
Positive predictive value (precision)	99.578 %	100.000 %
False discovery rate	0.422 %	0.000 %
F1 score	77.124 %	75.532 %

Adaboost Model

Number of boosting rounds	Training	Test
5	98.316 %	98.362 %
10	98.361 %	98.453 %
15	98.429 %	98.544 %
20	98.475 %	98.544 %

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

1. **Logistic Regression**
    a. **Learning Rate:** 0.1
    b. **Max Iteration:** 1000
    c. **Early Stop:** 0.0
2. **Adaboost**
    a. **Weak Learning Rate:** 0.1
    b. **Weak Learner Iteration:** 1000
    c. **Early Stop Error Threshold:** 0.5

- In all cases, the models were trained with top **10** features after calculating **Information Gain** of all the features.

The following Packages were used to run the script:

- pandas
- numpy
- matplotlib
- sklearn
- argparse
- pprint

How to Run the scripts

The script takes in command line arguments. The arguments are taken by using the "**--arg_name**" flag after the script. When the script runs, it runs both training and evaluation together on the specific dataset or on all the datasets together. The description of the arguments are given below:

- **--churn**: Full path of the churn CSV File to be given after the flag separated by a space.
    - Example: **--churn /path/to/churn.csv**
- **--adult**: Full path of the Adult data **folder(Not File)** containing the adult.data and adult.test files to be given after the flag separated by a space.
    - Example: **--adult /path/to/data/adult-dataset**

- **--fraud**: Full path of the Credit-Card Fraud CSV File to be given after the flag separated by a space.
    - Example: **--fraud /path/to/fraud.csv**
- **--dataset:** Has Possible values of 0,1,2,3. Here 0 indicates we are to run training and evaluation on **all** the datasets and correspondingly **1** indicates **Churn dataset**, **2** indicates **adult dataset** and **3** indicates **credit card dataset**
    - Example: **--dataset 0**
- **--featnum**: The number of features to be trained on, The features are selected after computing Gain of the features and selecting the top **featnum** features.
    - **--featnum 10**
- **--brounds:** The number of boosting rounds of the adaboost classifier
    - **--brounds 5**
- **--lgr_miter:** The max number of iterations for logistic regression classifier.
    - **--lgr_miter 1000**
- **--lgr_lr:** The learning rate of logistic regression classifier
    - **--lgr_lr 0.1**
- **--ab_learner_miter:** The maximum iteration of adaboost classifier weak learners
    - **--ab_learner_miter 1000**
- **--ab_learner_lr:** The learning rate of adaboost classifier weak learners
    - **--ab_learner_lr 0.1**
- **--ab_learner_estop:** The early stopping value(in fraction) for the adaboost weak learner.
    - **--ab_learner_estop 0.5**

For clarity, an example of the command to run the script is shown below:

python 1605079.py --churn /home/akil/Work/Work/Academics/4-2/ML/Assignment-1/data/cust_churn.csv --adult /home/akil/Work/Work/Academics/4-2/ML/Assignment-1/data/adult-dataset --fraud /home/akil/Work/Work/Academics/4-2/ML/Assignment-1/data/creditcard.csv --dataset 0 --featnum 10 --brounds 5 --lgr_miter 1000 --lgr_lr 0.1 --ab_learner_miter 1000 --ab_learner_lr 0.1 --ab_learner_estop 0.5

Here all the arguments are to be given in a single line one after another.

Observations

- The third dataset was heavily skewed with only 492 positive samples. So 5000 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. But there is room for improvements if we use systems to fight data imbalances.
- Variable learning rate was used to test how logistic regression is performing, but no improvements are observed for using variable learning rate.