Machine Learning Sessional

Performance Evaluation of Logistic Regression and AdaBoost for Classification

Student Name: Kowsar Mahmud Pappu
Student ID: 1805075

Report Submission Date:

Saturday 9th December, 2023

Course Code: CSE 472
Department: Department of Computer Science & Engineering

1 Instructions To Run The Code

- These Python modules are needed to run this code

  - : Numpy: Install numpy module using the command

    ```
    $ pip install numpy
    ```

  - : Pandas: Install pandas module using the command

    ```
    $ pip install pandas
    ```

  - : Scikit-Learn: Install Scikit-Learn module using the command

    ```
    $ pip install scikit-learn
    ```

- Place the dataset files inside the workspace folder and place their relative filename inside the file.
  - Set the relative file-path of the *Telco Customer Churn Dataset* to the variable **churnDataFileName** at 11th line of the code
  - Set the relative file-path of the *Adult Salary Scale Training Dataset* to the variable **adultDataFileName_train** at the 12th line of the code
  - Set the relative file-path of the *Adult Salary Scale Testing Dataset* to the variable **adultDataFileName_test** at the 13th line of the code
  - Set the relative file-path of the *Credit Card Fraud Detection Dataset* to the variable **creditCardDataFileName** at the 14th line of the code
  - Set the *Number of Features* to the **nFeatures** variable at the 287th line of the code for Feature Selection.
- Type command

  ```
  $ python 1805075.py
  ```

  to run the script
  - **The preprocessing part of the Adult Salary Scale Dataset was not running in my local machine and was showing errors that I could not find solutions for. But it ran smoothly in Jupyter Notebook on Google Colaboratory, so I tested just the Adult Salary Scale Dataset part on Google Colaboratory to collect performance evaluation data. Rest of the experiments were done in my local machine.**

2 Performance Evaluation

2.1 Telco Customer Churn Dataset

Logistic Regression

Performance Measure	Training	Test
Accuracy	0.8021	0.7892
True positive rate	0.4582	0.4805
True negative rate	0.9251	0.9053
Positive predictive value	0.6862	0.6560
False discovery rate	0.3138	0.3440
F1 score	0.5495	0.5547

AdaBoost Accuracy

Number of boosting rounds	Training	Test
5	0.8019	0.7935
10	0.8001	0.7899
15	0.7950	0.7921
20	0.7959	0.7956

2.2 Adult Salary Scale Dataset

Logistic Regression

Performance Measure	Training	Test
Accuracy	0.8255	0.8249
True positive rate	0.4448	0.4368
True negative rate	0.9463	0.9449
Positive predictive value	0.7243	0.7104
False discovery rate	0.2757	0.2896
F1 score	0.5511	0.5410

AdaBoost Accuracy

Number of boosting rounds	Training	Test
5	0.8372	0.8380
10	0.8407	0.8382
15	0.8413	0.8421
20	0.8393	0.8401

2.3 Credit Card Fraud Detection Dataset

Logistic Regression

Performance Measure	Training	Test
Accuracy	0.9644	0.9593
True positive rate	0.9185	0.8276
True negative rate	0.9656	0.9621
Positive predictive value	0.4035	0.3214
False discovery rate	0.5965	0.6786
F1 score	0.5607	0.4630

AdaBoost Accuracy

Number of boosting rounds	Training	Test
5	0.9751	0.9580
10	0.9586	0.9700
15	0.9683	0.9673
20	0.9715	0.9612