Report on CSE472 Assignment 2
Logistic Regression & AdaBoost For Classification

Md Asif Shahriar, #1805040

1 How to Run

To run the program for any of the existing datasets:

- First the dataset must be provided
- In the ‘root’ directory, create a folder named ‘data’
- Inside the ‘data’ folder place the csv file
- Now, inside the **main** function, store the file path inside a variable
- Tweak the parameters if needed
- Telco dataset : 1, Adult dataset : 2, Credit card dataset : 3
- For logistic learning, run **logistic_learning_test** for corresponding dataset
- For adaptive boosting, run **adaptive_boosting_test**

2 Dataset 1: Telco Customer Churn

- Learning rate: 0.01
- iterations/epochs: 5000
- Top features using information gain: 15

Logistic Regression

Performance Measure	Train Data	Test Data
Accuracy	0.7966	0.8119
Recall (True Positive Rate)	0.9014	0.9102
Specificity (True Negative Rate)	0.5067	0.5389
Precision (Positive Predictive Rate)	0.8348	0.8457
False Discovery Rate	0.1652	0.1543
F1 score	0.8668	0.8768

Adaptive Boosting

Number of Boosting Rounds	Train Data		Test Data	
	Accuracy	F1 score	Accuracy	F1 score
5	0.7817	0.8561	0.7991	0.8674
10	0.7907	0.8627	0.8048	0.8725
15	0.7911	0.8636	0.8001	0.8697
20	0.7950	0.8648	0.8084	0.8739

3 Adult Dataset

- Learning rate: 0.01
- iterations/epochs: 5000
- Top features using information gain: 15

Logistic Regression

Performance Measure	Train Data	Test Data
Accuracy	0.7979	0.8043
Recall (True Positive Rate)	0.3303	0.3367
Specificity (True Negative Rate)	0.9462	0.9489
Precision (Positive Predictive Rate)	0.6607	0.6710
False Discovery Rate	0.3393	0.3290
F1 score	0.4404	0.4484

Adaptive Boosting

Number of Boosting Rounds	Train Data		Test Data	
	Accuracy	F1 score	Accuracy	F1 score
5	0.8198	0.5295	0.8238	0.5319
10	0.8218	0.5589	0.8268	0.5652
15	0.8180	0.5560	0.8198	0.5539
20	0.8127	0.5917	0.8143	0.5885

4 Credit Card Data set

- Learning rate: 0.01
- iterations/epochs: 20000
- Top features using information gain: 15

Logistic Regression

Performance Measure	Train Data	Test Data
Accuracy	0.9793	0.9795
Recall (True Positive Rate)	0.1399	0.1515
Specificity (True Negative Rate)	1.0	1.0
Precision (Positive Predictive Rate)	1.0	1.0
False Discovery Rate	0.0	0.0
F1 score	0.2455	0.2631

Adaptive Boosting

Iterations: 5000

Number of Boosting Rounds	**Train Data**		**Test Data**	
	Accuracy	**F1 score**	**Accuracy**	**F1 score**
5	0.9940	0.8627	0.9956	0.9010
10	0.9939	0.8615	0.9956	0.9010
15	0.9939	0.8615	0.9958	0.9071
20	0.9939	0.8612	0.9956	0.9910