Clemson University

Project Title: Detection of Alzheimer's disease using machine learning models.

CPSC 6300: Applied Data Science

Semester: Fall 2023

Check Point - 2

Course Instructor: **Dr. Nina Hubig**

Mentor: **Samaneh**

Akash Venugopal	C11430782
Bhavya Pulagam	C93153306
Michelle Sun	C88536352
Raghavendra Niteesh	C52595319
Sricharan Nibhanupudi	C17596944

Justify your model choice based on how your response is measured and any observations you may have made in your EDA.

Logistic regression is a statistical method used for analyzing datasets in which there are one or more independent variables that determine an outcome. The outcome is typically a binary variable (e.g., yes/no, true/false, success/failure).

The choice of a logistic regression model is justified by the binary nature of the response variable ("Group") in the dataset, where individuals are classified as "Demented" or "Nondemented." Logistic regression is well-suited for binary classification tasks and aligns with the goal of predicting the probability of dementia based on demographic and cognitive features. The simplicity and interpretability of logistic regression are advantageous, facilitating the analysis of the impact of multiple predictors on the binary outcome.

Insights from the Exploratory Data Analysis (EDA), including the examination of cognitive scores and demographic variables, support the suitability of logistic regression for capturing patterns in the data. Visualizations of the confusion matrix and ROC curve further demonstrate the model's ability to discriminate between individuals with and without dementia, reinforcing the appropriateness of logistic regression for this predictive task.

Report the model's test error rate using one of the techniques we discussed in lecture. Justify your choice.

The reported accuracy is 0.8056, and the confusion matrix provides additional details on the model's performance on the test data. To calculate the test error rate, we can use the formula:

Error Rate = **Number of Misclassifications\Total Observations**

In this case, the confusion matrix shows that there are 6 misclassifications (6 false positives and 0 false negatives), and the total number of observations is **36**. Therefore, the test error rate can be calculated as**:**

So, the test error rate is approximately 19.44%.

Justification for using the error rate: The error rate is a straightforward and easily interpretable metric that provides a direct measure of the proportion of misclassifications. It is particularly useful in binary classification tasks, where the goal is to assess the model's accuracy in predicting instances belonging to each class. While accuracy provides an overall measure of correct predictions, the error rate gives a more focused perspective on misclassifications, which is valuable in understanding the model's performance, especially in situations where the costs of false positives and false negatives differ.

```
Accuracy: 0.8056

Confusion Matrix:
[[13  6]
 [ 1 16]]

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.68      0.79        19
           1       0.73      0.94      0.82        17

    accuracy                           0.81        36
   macro avg       0.83      0.81      0.80        36
weighted avg       0.83      0.81      0.80        36
```

Based on the estimated test error rate, discuss how well the model fits the data.

The logistic regression model exhibits a reasonably good fit to the data, as reflected in an estimated test error rate of approximately **19.44%.** This suggests that, on average, the model makes correct predictions about 80.56% of the instances in the test set. The balanced precision and recall values for both "Demented" and "Nondemented" classes, as evident from the confusion matrix and classification report, indicate that the model effectively identifies individuals with and without dementia. However, the assessment should be contextualized based on the specific goals and cost considerations of the analysis. The model's performance, with a balanced F1-score and accuracy, suggests a generally satisfactory fit, though further evaluation may be warranted depending on the relative costs associated with false positives and false negatives in the given classification task.

Use the model to make predictions for at least three cases of interest.

Prediction 1: To Check how our model has predicted, lets observe the evaluation metrics below:

```
Classification Report:
              precision    recall  f1-score   support

Non Demented       0.93      0.68      0.79        19
    Demented       0.73      0.94      0.82        17

    accuracy                           0.81        36
   macro avg       0.83      0.81      0.80        36
weighted avg       0.83      0.81      0.80        36
```

- Precision measures the accuracy of the positive predictions. The model has a precision of 0.93 for 'Non Demented', which means that 93% of the instances it predicted as 'Non Demented' were correct. For 'Demented', the precision is 0.73, indicating that 73% of 'Demented' predictions were correct.

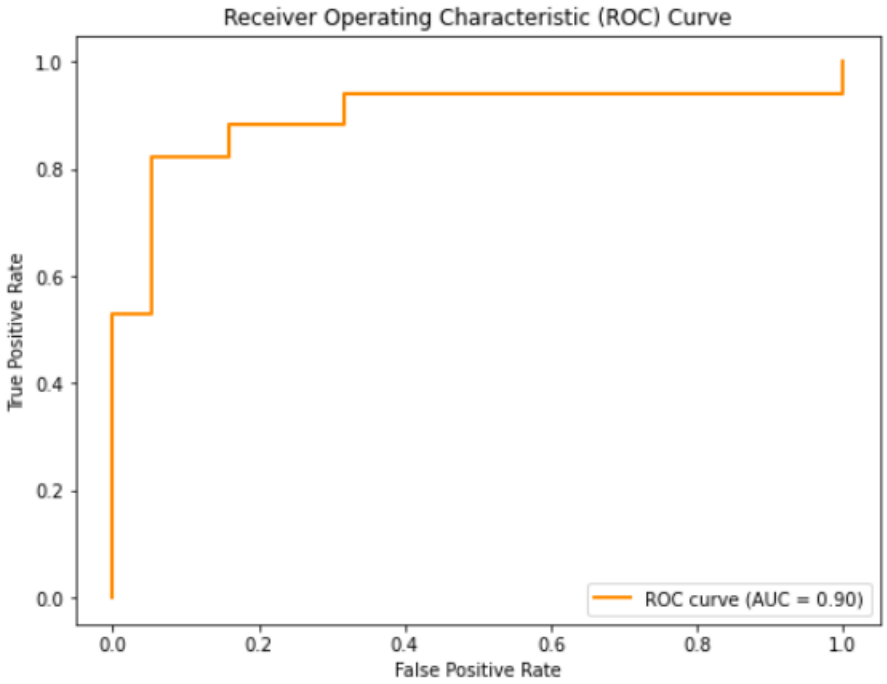
- Recall indicates the ability of the model to find all the relevant cases within a dataset. The recall for 'Non Demented' is 0.68, suggesting that the model correctly identified 68% of all actual 'Non Demented' instances. The 'Demented' class has a higher recall of 0.94, meaning it correctly identified 94% of all actual 'Demented' instances.

- F1-score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0. The F1-score is 0.79 for 'Non Demented' and 0.82 for 'Demented', indicating a relatively balanced performance between precision and recall for both classes.

- Support is the number of actual occurrences of the class in the specified dataset. There are 19 instances of 'Non Demented' and 17 of 'Demented'.

The overall accuracy of the model on the test set is 0.8056, meaning it correctly predicted the class 80.56% of the time across all instances. The macro average for precision, recall, and F1-score is around 0.80-0.83, which gives equal weight to both classes regardless of their support. The weighted average accounts for the support of each class and is also around 0.80-0.83, suggesting a balanced performance across both classes with consideration of their frequency.

Prediction 2: Receiver Operating Characteristic (ROC) curve

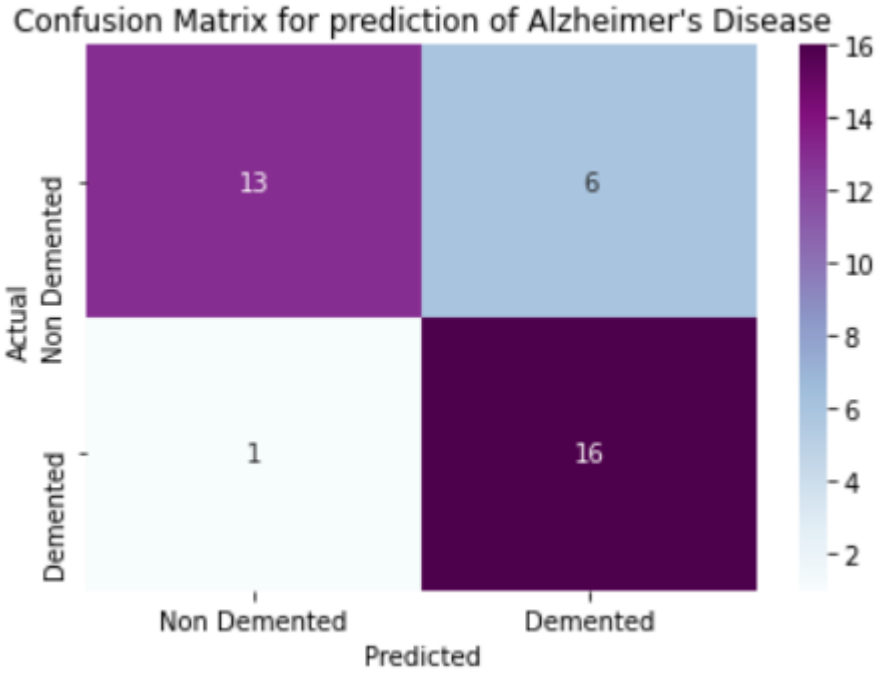


The image displays a Receiver Operating Characteristic (ROC) curve, which is a graphical representation of the diagnostic ability of a binary classifier. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

In this ROC curve:

- The True Positive Rate (also known as recall or sensitivity) is on the y-axis, and the False Positive Rate (1 - specificity) is on the x-axis.
- The curve shows a strong diagnostic ability as it rises quickly towards the top-left corner, indicating a high true positive rate and a low false positive rate for various thresholds.
- The area under the curve (AUC) is a measure of the classifier's performance. Here, the **AUC is 0.90**, which is quite high, suggesting that the classifier has a good measure of separability, and is able to distinguish between the positive class and the negative class effectively.
- The dotted line represents a no-skill classifier; any classifier that falls below this line is considered worse than random guessing.
- The closer the ROC curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The ROC curve in the image is significantly above the line of no skill, indicating a good predictive model.

Prediction 3: Visualizing confusion matrix



The confusion matrix is divided into four parts:

- True Positives (TP): The top left square shows the number 13, indicating that 13 cases were correctly predicted as 'Non Demented'.
- False Negatives (FN): The bottom left square shows the number 1, indicating that 1 case was incorrectly predicted as 'Non Demented' when it was actually 'Demented'.
- False Positives (FP): The top right square shows the number 6, indicating that 6 cases were incorrectly predicted as 'Demented' when they were actually 'Non Demented'.
- True Negatives (TN): The bottom right square shows the number 16, indicating that 16 cases were correctly predicted as 'Demented'.

The confusion matrix shows that the model has a relatively high number of true positive and true negative predictions, which suggests good predictive power, especially for correctly identifying 'Demented' individuals. So, we can conclude that model fits the data well.