Abigail Smith

ARS190011

CS 4375.004

Dr. Mazidi

TA: Ouyang Xu

04/08/2023

ML with Sklearn

a. Which algorithm performed better?

While there exists several factors influencing how well the logistic regression model, decision tree model, and neural network performed on the given dataset, score is a quick evaluator for which algorithm performed better since score represents the test accuracy of each model. As shown in the table below, the logistic regression had the highest score and, therefore, arguably performed better.

| Score | | |
|---|---|---|
| Model | Rank | Score |
| Logistic Regression | 1st | 0.8974358974358975 |
| Decision Tree | 2nd | 0.8717948717948718 |
| Neural Network #1 | 3rd | 0.6441604220821195 |
| Neural Network #2 | 4th | 0.4979347132265437 |

b. Compare accuracy, recall, and precision metrics by class.

As stated before, there exists several factors influencing how well the logistic regression model, decision tree model, and neural networks performed on the given dataset. Precision is simply the true positives divided by the sum of the true positives and false positives. Additionally, recall represents the true positives divided by the sum of the true positives and false negatives. The closer accuracy, precision, and recall are to one, the better the model's accuracy, precision, and recall (respectively).

| Classification Report for Logistic Regression | | | | |
|---|---|---|---|---|
| | Precision | Recall | F1-Score | Support |
| 0 (not high mpg) | 1.00 | 0.84 | 0.91 | 50 |
| 1 (high mpg) | 0.78 | 1.00 | 0.88 | 28 |
| Accuracy | | | 0.90 | 78 |
| Macro Avg | 0.89 | 0.92 | 0.89 | 78 |
| Weighted Avg | 0.92 | 0.90 | 0.90 | 78 |

As seen in the classification report for logistic regression, the precision and F1-Score were higher for 0, or not high mpg, than 1, high mpg. In contrast, the recall was higher for 1 than for 0. As described above, the closer each of the values are to one, the better the model's accuracy, recall, and F1-Score (individually). Since logistic regression has a higher precision and F1-Score for 0, it's arguable that the model worked better with label 0. It is important to note,

however, that 0 has a higher support than 1, meaning there were more samples with label 0 than label 1. This factor may have influenced the outcome of this classification report and, as such, should be remembered when evaluating the performance of the logistic regression model (as seen with the higher F1-Score for label 0, this was most likely the case).

| Classification Report for Decision Tree | | | | |
|---|---|---|---|---|
| | Precision | Recall | F1-Score | Support |
| 0 (not high mpg) | 0.92 | 0.88 | 0.90 | 50 |
| 1 (high mpg) | 0.80 | 0.86 | 0.83 | 28 |
| Accuracy | | | 0.87 | 78 |
| Macro Avg | 0.86 | 0.87 | 0.86 | 78 |
| Weighted Avg | 0.87 | 0.87 | 0.87 | 78 |

For the decision tree, the precision was higher for 0, or not high mpg, than 1, or high mpg. Similarly, so was the F1-Score and recall. The output for the classification report for the decision tree resembles the distribution shown in the classification report for the logistic regression model, but with different values and proportions. The gap between precision and recall was lower between label 0 and label 1 in the classification report for the decision tree than in the classification report for logistic regression, despite having the same number of observations in each label. This occurred as a consequence of the innate difference between how

logistic regression and decision trees operate -- how they work to make predictions. Despite these differences, both the logistic regression and decision tree had very close accuracies with logistic regression having a slightly higher accuracy.

c. Give your analysis of why the better-performing algorithm might have outperformed the other.

There exists several characteristics and factors impacting each of the algorithm's performance, most of which work in logistic regression's favor in terms of performance.

The main reason why logistic regression performed better than the decision tree model and neural network comes from the relationship between the high mpg and not high mpg with the predictors. As seen in section 6 with the various catplots, majority of each catplot had a clear relatively distinction between high mpg and low mpg regardless of the predictors being compared. This visual exploration implies the existing relationships in the dataset may work better with logistic regression due to how logistic regression works. Logistic regression has an ideal approach as logistic regression develops a decision boundary to divide observations into regions where most observations are of the same class. As seen visually, most relationships showed a relatively clear boundary between high mpg and not high mpg for each predictor, meaning that using a decision boundary would have a better chance at correctly dividing the observations. In contrast, since decision trees work by iteratively separating the input observation into partitions until the observations in each partition are uniform, the decision tree may have misclassified observations by placing them into partitions as opposed to the simpler single division of logistic regression. Similarly, neural networks work by using a predefined number of hidden layers and nodes in each layer with a feed-forward network and back propagation to identify relationships and perform classification and regression; however, as stated, the relationships were much simpler as not only were the output was expected to be binary, but the relationships between each predictor and target were, at least visually, relatively clearly defined.

The second reason why logistic regression may have performed better than the decision tree model and the neural network comes from the size of the dataset. With only 389 observations, the dataset is relatively small, and because of this, lacks the size that would benefit decision trees and neural networks. While the dataset size was enough to show a relatively clear distinction between high mpg and not high mpg by each predictor (as seen in section 6), the dataset size lacked the size required to fully explore how logistic regression, decision trees, and neural networks perform on this type of data. Since decision trees iteratively partition observations into groups until each group is uniform, the tendency for more of the explorable relationship to have a clear decision boundary between them may be a result of the small data size (that is to say, with more observations, the relationships shown in section 6 may have changed enough for decision trees to gain a performance advantage). Similarly, neural networks have the capability to explore complex relationships that simpler algorithms cannot -- and in this regard -- may perform better on a larger, more complex dataset. Additionally, since neural networks find local optima and there is not a guarantee that it is the best, the size and complexity of the dataset influence the performance of the neural network. Finally, the dataset size undoubtedly has the chance to influence the proportion of high mpg and not high mpg observations in both the train and test data set.

Overall, several factors impact how well the logistic regression model, decision tree model, and neural network performed on the provided dataset.

d. Write a couple of sentences comparing your experiences using R versus Sklearn. Feel free to express strong preferences.

Personally, I enjoy using both Sklearn and R, but I do have a preference towards Sklearn. The main reason is that I have more experience with programming languages like Java and C++, so using Python is a tool I can easily comprehend, apply, and manipulate. Additionally, while the Jupyter Notebook and RStudio are very similar, I prefer Jupyter Notebook mainly because, like Sklearn, Jupyter Notebook feels more flexible and malleable. To me, R was a great introductory tool to machine learning, but I believe Sklearn will be the next step in my learning adventure.