Tanyag Johann Francois | BSCS3B | CSST102

Interpretation & Discussion

For this lab activity, I created a model that predicts the test preparation course, whether a student has completed the prep course or did not. By using the scores and demographics provided by the dataset, the model can classify the students into two categories; completed or none. This allows individuals to see that correlation does not imply causation.

The public dataset used for this model is named as “StudentPerformance.csv”. This dataset provides various raw information such as; gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, and writing score. With the data in hand, various predictions can be made. However, for this activity, the test preparation column is specifically targeted.

The results of the confusion matrix clearly indicate that the model is effective at prediction. The model predicted 40 True Positives (TP), 32 False Negatives (FN), 16 False Positives (FP), and 112 True Negatives (TN). Based from the output, the model is better at identifying “none” (112 correct) than “completed” (40 correct).

In the 5-Fold Cross Validation phase, the model resulted with consistency as most fold outputs has an approximate ~0.7 score. The mean CV accuracy of the model is 0.7300 and its standard deviation is 0.0249. This means that the model consistently achieves 73% accuracy across all five folds, with only a small variation (~2.5%). The model is stable and can generalize well. Moreover, it isn’t overly sensitive to which data is in the training vs validation splits.

The last phase is the learning curve. The graph presents two curves: the blue curve which indicates the training score and the green curve that presents the cross-validation score. The blue curve starts high with a ~0.90 score but drops as more data is added, stabilizing around 0.78. In contrast, the green curve started low with a score of ~0.68, but it rises as more data is added. This curve stabilizes around 0.74. This indicates a small but consistent gap between the curves. Which means that the model is not overfitting and not severely under fitting.

The model can be improved by having a better dataset. The dataset that I’ve used holds back the model that I trained making its predictions lackluster. As the confusion matrix had presented, the model is better at predicting students that have not completed the test preparation course. This is because the dataset has more students that have a “none” value in their test preparation course. Thus, it is the cause of the imbalance between the ratio of the prediction. Having a better dataset can improve the model’s prediction to prevent bias. Moreover, it can make the system more reliable and accurate for predicting the test preparation course.