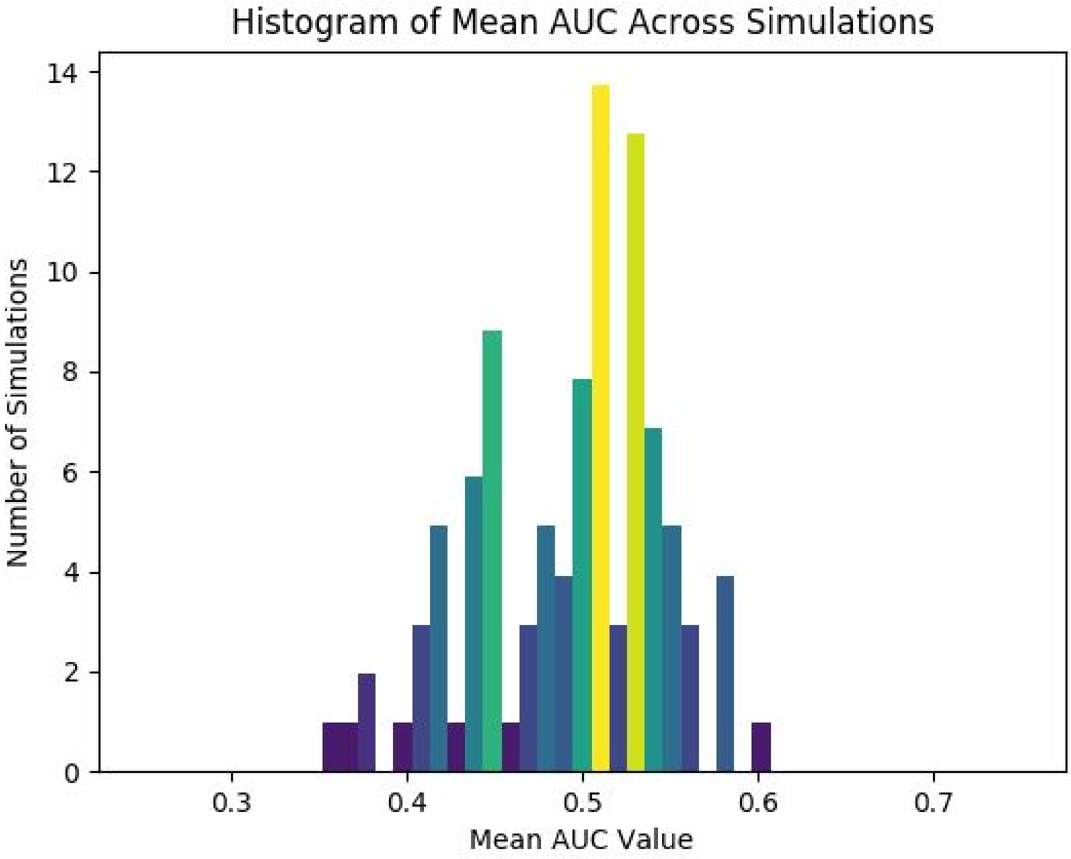
Avery Simon
April 28, 2020

Problems for Data Scientists: Report

My analysis indicates that the classifier used during cross validation proved indiscriminate, meaning that the model was unable to significantly differentiate between the data and the simulations. Below is a histogram of the mean area under the curve across all of the provided simulations color coded by density. The highest number of simulations had a mean AUC value of or around 0.5, while many others had a value less than that. Because AUC is the measure of true positive rate in relation to false positive rate, this indicates our model was unsuccessful.

The next step I would take to address the central question would be to reevaluate the simulation generator to see if it can be changed to produce more accurate results. Additionally, I would think about what other classifiers could be used to better differentiate between the data and simulations. Another idea would be to restructure the simulations so that they are organized by subject rather than by simulation and then continuing to use the data versus simulation classifier.


Histogram of Mean AUC Across Simulations
Number of Simulations
Mean AUC Value