Evan Mickler

Predictions:

The random forest classifier should preform better on unscaled data. The random forest classifier works by creating multiple decision trees from random subsets of the data and arriving at a final decision by a kind of majority rule. Using many other modeling techniques, the use of unscaled data can cause a few outliers to have too much influence over the model. However, since random forest averages multiple decision trees each based on only a subset of the data, outliers will only influence a few of these and will not have an outsized impact on the decision.

Scaling the data forces it to fit a normal distribution which will pull extreme outliers closer to the mean and will limit their impact on the overall model. This will greatly improve the performance of the logistic regression model but may offer small gains if any for the random forest model. This should mean that the two models should perform similarly with scaled data or logistic regression may even surpass random forest.

Reflections:

In the first round, random forest outperformed logistic regression as predicted scoring 0.65 compared to logistic regression at 0.51. However, in the second round, using scaled data, logistic regression outperformed random forest scoring 0.76 compared to random forest which remained relatively unchanged at 0.64. This too falls in line with the predictions.