HW5 - Model Comparison

Kayla Kippes

Task 1: Conceptual Questions

- What is the purpose of using cross-validation when fitting a random forest model?

Cross-validation is used when fitting a random forest model in order to split the data into multiple folds and only train the data on some of those folds while testing it on others. This helps pick the best model by seeing how the model will do on unseen data and it will help prevent overfitting.

- Describe the bagged tree algorithm.

Create many bootstrap samples of the data and then fit trees to each of those samples of the training data. For each tree, find a prediction value and take all predictions and either average (for regression) or chose the most common (for classification) to create one final prediction.

- What is meant by a general linear model?

General linear model is a group of models that allows for the response variable to taken on different distributions so the predictors could be continuous or categorical. One example would be modeling a response variable that is binary.

- When fitting a multiple linear regression model, what does adding an interaction term do? That is, what does it allow the model to do differently as compared to when it is not included in the model?

Adding an interaction term allows for the value of one predictor be dependent on the value of another predictor. Without it, the model assumes that each predictor affects the response independently.

- Why do we split our data into a training and test set?

We split our data into a training and test set so that the model can be fit on data it has never seen before. It can learn from the training data and from there we can evaluate it's performance on the test data.