GSB 544: Machine Learning

**Chapter 10: Introduction to Predictive Modeling – Not on Midterm**

* In predictive modeling, there is **one target variable** that we are hoping to predict in the future.
* The target variable could take many forms:
  + Price of a particular house
  + Profits of a company next year
  + Whether or not a person will click a targeted ad on a website
* Predictors/features – Information we choose to use in our prediction strategy.
  + Variables that will help predict the target variable

Model Specifications

* What is the procedure for coming up with these predictions?
  + Make an equation?
  + Look at the prices of similar houses

Choosing a final model

* “Choosing the final painter”
* Use the training data to nail down the details of our process
* Collecting information about house sizes and prices in the same neighborhood
* Determining which exact values of a, b, c, d do a good job predicting for known house prices.
* Deciding on categories of houses that seem to be priced similarly.

Test Data and Metrics

* We need to settle on only one procedure to use to come up with our predictions of unknown target values.

Use Mean Square Error (MSE) to measure how good your model is at predicting. (Lowest MSE)

Once we decide what the best model is (the best predicting process) we need to fit the final model – Use all the data we have, test and training, to re-train the winning model. This creates the fitted model we will use on our actual future unknown data.

Modeling with Scikit-learn

* Step 1: Determine response and predictor variables
* Step 2: Choose which model specifications we will consider as possible prediction procedures
* Step 3: Split data into training and tests datasets – This is a RANDOM Function!
  + Test set data will be hidden from us for future use.
* Step 4: Fit all 3 of our models on the training data
  + Save this as a new variable
  + Then, we use the fitted models to get predicted values for the test data.
* Step 5: Compare the predicted values from the test data to the actual true values for the test data.
  + Calculate the MSE for all three models
  + The smallest MSE will be the final model that we use!