Analyzing Boston Property Assessments
STAT139 - Project Proposal

Samuel Daulton, Avery Faller, Fanny Heneine, Yuhan Tang

April 20, 2016

GitHub: https://github.com/sdaulton/stat139_final_project

The Data

We found a verwby interesting data set on the City of Boston's website that details the assessed property values for all property within the City of Boston for the year 2014. The data can be viewed here:
https://data.cityofboston.gov/dataset/Property-Assessment-2014/qz7u-kb7x

This data set is very extensive and includes 164,092 properties, with up to 55 variables for each property. Importantly, the data can be downloaded as a CSV and then can be imported into R. This let's us start analyzing the data immediately, and not have to spend time scraping data before we can even begin.

We have already begun exploratory data analysis on this data set and found some interesting points, and some rather unlikely points, i.e. one property that was purportedly constructed in the year 62,990. Obviously, we will need to watch out for these erroneous entries while we are constructing our models and running tests. To do this, we will pre-process the data to address any missing values and entry errors.

Below we outline the main research questions we will tackle and lay out the methodology we will follow to answer them.

Applied Research Question

How well can we predict property assessment value using predictor variables containing details about the property? Which predictor variables and what type of model give the most accurate predictions?

Models and Techniques

First, we will partition the data into training, validation, and test sets. The test set will only be used to compare finalized models. The training set will be used to train each model, and we will tune parameters and adjust the predictor variables using the validation set.

The baseline model will be a linear parametric model consisting of the original (significant) features; this means no interaction terms or transformed predictor variables. We will check for, and address if present, heteroskedasticity, multicollinearity, model bias, and non-linearities. We will construct additional models using interaction terms and predictor variable transformations to find the best fitting parametric model we can. If non-linearities are present, we will build semi-parametric models to account for the nonlinear relationships. Lastly, we test out other techniques including the following possibilities: k-Nearest Neighbors and Random Forests.