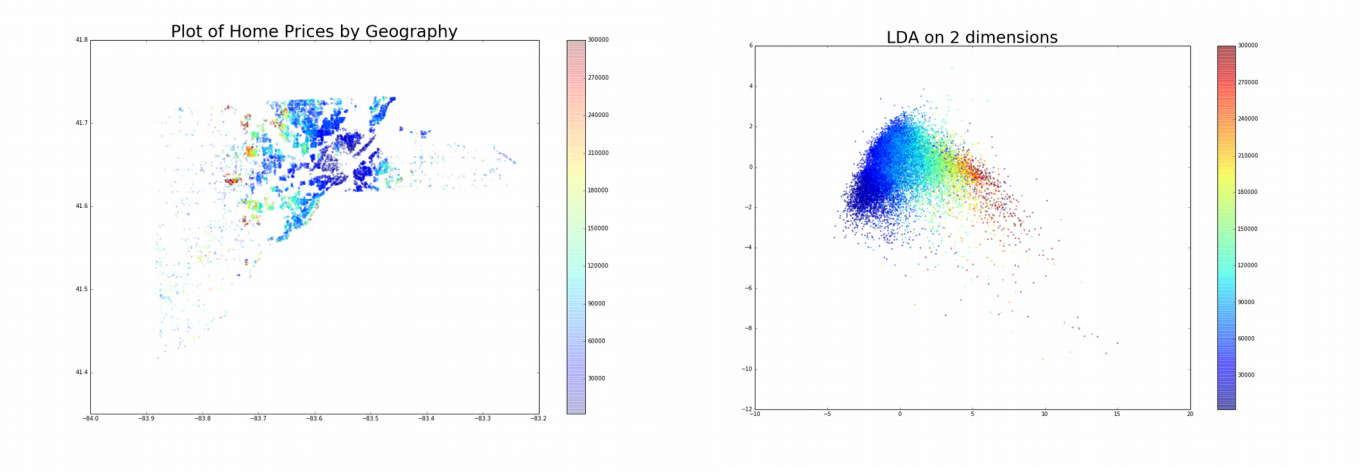
Christopher Tong
CE 263N
Fall 2016

Midterm Writeup

Intro:

My approach to modeling this midterm problem is based on my belief that location is the most important indicator of actual home prices. By looking at the map of housing prices as plotted by locations, we can see areas of high values homes clustered around each other. I initially wanted to capture cluster these areas of high value by using a Decision Tree Classifier. Why my eventual submission does not include this clustering step will be explain in the end.



Why Gradient Boosting Regression Trees:

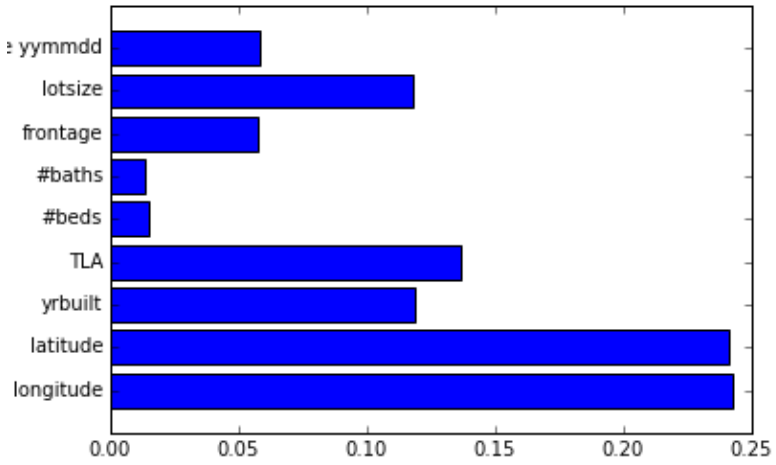
My first attempt at prediction reduced dimensionality using LDA. I chose this method as it allowed me to visualize the data using only a 2-d plot (above). Using LDA, cross validated prediction values resulted in a RMSE of 28,000.

One main benefit of using Gradient Boosting Regression Trees is the algorithm works well with data with multiple scales. In our dataset, latitude-longitude , wall codes, and prices all have different scales for their respective dataset. This advantage is important for our dataset more-so than datasets such as facial recognition (where all features are frequencies of some kind). Secondly, GBRT algorithm can also detect the non-linear relationships between features, which again is useful to model the relationships between are multi scaled dataset.

Boosting as a technique weights errors with greater emphasis on each iteration during training by fitting each sequential tree to the residuals of the prior tree. This characteristic allows us to further improve predictions by specifying a loss function that best fits our dataset.

Results:

GBRT's main drawback is the time it takes to tune parameter values. My computer spent nearly 30 minutes on the tuning step. After tuning the results show a large importance of the location features in the estimation.



Expanding to GBRT with prior classification:

Due to the shown importance of geography on price, I planned to improve predictions by first binning houses to different wealth levels and running separate models on each bin. Initial results proved encouraging as house prices for all houses under the 75th percentile had a prediction RMSE of 10,000. During the cross validation step, however, I was not able to tune parameters to an error lower than 18,000 – above what I have achieved with simple GMRT. This method of bining also required the explicit storage of the models I will use for each bin during the prediction step. This problem along with indexing considerations complicated implementation of the model in prediction of values. From further consideration, I realize GBRT may have already accounted for clustering in their algorithm. Nonetheless, I am confident if given enough time and resources, this method of classification then prediction through multiple models can lower the error of the model.