EE5373: Data Modeling Using R
Fall, 2019
Department of Electrical and Computer Engineering
University of Minnesota

Lab 7: Data segmentation for house price predictions project.

Due date: See the due date shown on the class web page.

Goal: This lab introduces you to data segmentation and the kaggle competition process.

What to do:

There is a link on web under this lab's tab for a kaggle competition to develop a regression model to predict house prices. You are to develop several different models using this house price data:

1. Use all of the available input parameters as predictors in your model. You should exclude predictors that are obviously not useful, such as the property ID number, and so forth. After partitioning the data into appropriate training and testing sets, show how well this model predicts the house prices.
2. Use the backward elimination process to develop your regression model using all of the available data, partitioned into appropriate training and testing sets. Show how well this model predicts the house prices for your testing data set.
3. Segment the data into three different sets based on zip code - one set for the zip code with the highest average house price, one set for the zip code with the median average house price, and one set for the zip code with the lowest average house price. Develop a regression model for each of these three zip codes separately using backward elimination. After partitioning the zip codes into appropriate training and testing sets, use each model to predict prices within its own zip code, and across the other two zip codes.
4. Develop another set of models using a data segmentation of your choice. For instance, you could try segmenting by price points (i.e., a different model for high, medium, and low priced houses), by distance from the highest density of houses (using the latitude and longitude data), by the condition of the house, etc. Show how well this model predicts the house prices within its own segment, and across segments.
5. Finally, use your best model(s) to predict the house prices for the test data set provided by the kaggle competition as explained in the kaggle invitation. Upload your best result to the kaggle site.

You must use the testing and training process to appropriately partition the data set to evaluate the predictive performance of all of your models. Measure the quality of your predictions on the test data set using the RMSE metric. You must also show how you cleaned the data to ensure that there were no obvious problems in the raw data.

What to turn in for grading:

Write a short lab report that shows all of your models, and explains the prediction results for each model. Include both the RMSE values you compute for your testing data, and the RMSE values reported by kaggle for your best model.