Deemal Patel
MSDS 422

Assignment 2 – Boston Housing

Residential real estate market value has been a difficult to assess and in order to get the values manually make it a very tedious process. There are so many factors that go in to the cost of a house and using every feature to calculate the prices requires too much work. Thanks to machine learning, there are four regression models that can assist us in assessing the market value of real estate. With the help of the Boston Housing Study, we have data that can help train our machine learning models in order to get strong predictions. All the features in the data set play a factor in to determining the market value and we will use these features with our machine learning model to train our model to predict housing prices.

We begin by looking at our data set and determine how it looks and what sort of data types are contained inside it. For the model purpose we must have numeric values only therefore we drop the neighborhood column and continue preparing, analyzing, and visualizing our data. We added a correlation heatmap to get an idea of what features play a big role in our housing values.

The four regression models that we may implement are Linear Regression, Ridge Regression, Lasso Regression, and Elastic Net Regression. Each method has it's pro's and con's but for the sake of determining housing prices, **I recommend using the Ridge Regression model in order to get the best accurate market value.** The reason behind my recommendation is that when analyzing the coefficient of determination and the root mean square error, their values came out to approximately 0.754 and 0.507, respectively. The coefficient represents a percentage of values that lie on the line that is generated from the model. 0.75 means that the predictors (features) explain about 75% of the variation in our response variable. Given this

logic, higher the coefficient, the more variation is explained and therefore is considered to be a better model. Linear, Lasso, and Elastic Net were close but Ridge had the best value. The root mean square error represents how spread out the data is from the regression line. The lower the error the better it is. Aside from the values, I also recommend Ridge because it is a good choice for a model when we want to use all our predictors. In this case we did use all of our predictors to train the model and every feature plays its role in getting the market value