A housing study was conducted on 506 census tracts in Boston. A census tract is a small, relatively permanent statistical subdivision of a county and could be considered a neighborhood. Some of the variables in the study included air pollution (nitrogen oxide concentration), crime rate, location information, age of the home, tax rate, school and socio-economic status. The goal is to use this data set to assess the market value of residential real estate and predict the median value of homes.

The study variables were collected into a comma separated value file (.csv). The .csv file was opened in Python and saved into a data frame called `boston_input`. The data frame contains information on 506 homes on 14 variables. The neighborhood variable was dropped from the data frame. A correlation matrix of survey variables was constructed into a graphical display called a heat map when applicable. The correlation coefficient value is displayed, and color coordinated. Positive correlation is red and negative correlation is blue. The average number of rooms per home had a positive correlation to the median value, while the percentage of population of lower socio-economic status had a negative correlation to the median value. There was noted correlation between all variables except for the variable “chas” (on the Charles River). The data set was standardized by subtracting the mean and dividing by the variance for each data point. Standardization was performed to prepare the data for analysis with machine learning algorithms that do not perform well with data that contain observations orders of magnitude larger than others.

The data was split into 75:25 train:test sets. Four different algorithms were used to create models for comparison: (1) linear regression; (2) ridge regression; (3) Lasso; and (4) Elastic Net. All are similar to each other with differences in the type of regularization used.

Regularization is used to constrain a model to limit the weights used to create a simpler model and permit better generalization towards new data. Linear regression has no regularization. Ridge regression uses L2 regularization, which constrains weights to be as small as possible or be close to zero, and Lasso uses L1 regularization, which can set weights to zero to eliminate the least important features. Elastic Net uses a ratio of both L1 and L2 regularization.

All four algorithms scored higher with the test data set compared to the train set, which means that the models were not overfit to the training data. Models were scored utilizing 10 folds in cross-validation with the Root Mean Square Error (RMSE) performance metric. A small value for RMSE is preferred. The Elastic Net model provided the lowest mean and standard deviation for RMSE.

In order to determine the optimal values for the Elastic Net hyperparameters of alpha and L1 ratio, the `GridSearchCV` function was used. This function iterates over specified values for parameters and utilizes a cross-validation fold. `GridSearchCV` determined that the optimal parameters are an alpha of 0.5 and L1 ratio of 0.1.

Finally, k-fold cross-validation function `KFold` was used to evaluate all four algorithms. The RMSE for each of the ten (10) cross-validation folds was computed and the mean for each method was determined. As with the earlier evaluations, Elastic Net had the lowest mean RMSE.

The Elastic Net regression model is recommended for use in assessing market value of residential real estate. The model utilizes both L1 and L2 regularization in determining the model coefficients and had the lowest RMSE.