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. Six different algorithms were used to create models for comparison: (1) linear regression; (2) ridge regression; (3) Lasso; (4) Elastic Net; (5) Random Forest; and (6) Gradient Boosting. The first four algorithms 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. The last two algorithms are ensemble methods, which function by utilizing an aggregation of classifiers. The Random Forest algorithm uses a random subset of features to generate an ensemble of decision trees. The samples for each tree are drawn with replacement (bootstrap=True) and the results are averaged together to form the final model. The Gradient Boosting algorithm performs a number of boosting stages while sequentially adding predictors. With each iteration, the algorithm will try to fit the new predictor to the residual errors made by the previous predictor. The ensemble models were evaluated utilizing the `Pipeline` function.

All four non-ensemble 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. The ensemble methods had higher scores with the train set versus test set, which indicates that the model should be pruned to reduce overfitting. 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 Random Forest model provided the lowest mean RMSE.

In order to determine the optimal values for the Elastic Net, Random Forest, and Gradient Boosting parameters, the `GridSearchCV` function was used. This function iterates over specified values for parameters and utilizes a cross-validation fold. `GridSearchCV` results were used to prepare the final models for k-fold cross-validation.

Finally, k-fold cross-validation function `KFold` was used to evaluate all six algorithms. The RMSE for each of the ten (10) cross-validation folds was computed and the mean for each method was determined. The Random Forest model had the lowest mean RMSE.

The Random Forest regression model is recommended for use in assessing market value of residential real estate. The model utilizes an ensemble of decision tress in determining the feature importance and had the lowest RMSE.