HOME PRICE PREDICTION OBJECTIVES

EXECUTIVE SUMMARY

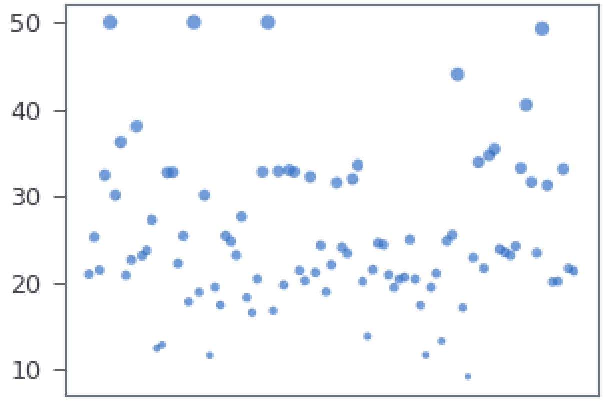
To gather an accurate prediction for a specific home valuation many variables must be taken into consideration. Traditionally, some of the higher impacting variables are the size of the home in question, which is typically a reflection of rooms in the house, the area in which it is located, and the school district in which the housing unit in question falls into.

These valuation metrics are time consuming when researched manually, and the results are often in a generous ballpark range that helps brokers advise clients on the price range to list their house, so they can both maximize value out of the home and minimize the time on the market. To reduce the manual effort involved in such tasks, we can employ machine learning techniques that can predict the value of a home based on the data collected from the housing areas brokers are targeting.

By employing a modern machine learning approach to this classical manual task, we can reduce the overall operational overhead of the firm in terms of staffing manual valuation agents, and additionally our model will improve over time as we collect more data around the housing market in general, the areas in question, we can calibrate the accuracy of our predictions on a forward-looking basis so we are constantly one step ahead of the competition.

RESEARCH DESIGN

For this problem we use as our baseline dataset that tracked the median value of homes in the Boston area in several neighborhoods. The first order of business in this study was to remove the neighborhood variable from the dataset, as if we know there are tight correlations between the average value, as we can see in the following plot where the size of the dot is the mean home value and the location is by area.



After removing neighborhood from the dataset, we are left with tweleve additional explanatory variables to build our valuation model. We will implement four distinct machine learning models on the same dataset then evaluate the overall performance in terms of root-mean squared error, and recommend the best performer to management.

TECHNICAL OVERVIEW

This analysis was performed purely in the cloud, leveraging a pre-canned environment from a world-class provider of machine learning solutions, Google, called Colabratory. This cloud environment enables us to not only conduct research in a more efficient way, it also allows us to share the details of this research to the astute reader who may be interested in the details of the results, as well as a deep dive into our work.

In this environment, we were able to load our Boston housing dataset, clean it, and execute several exploratory data analysis techniques to help us derive a cleansed and sensical dataset to our machine learning algorithms.

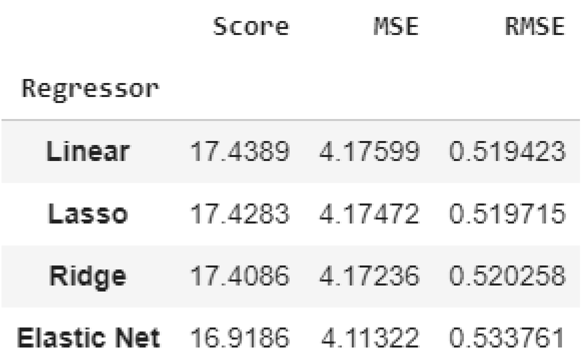
The algorithms in question are all in the category of Regression due to the continuous nature of our response variable, median house price. The specifics selected for this study are Linear, Ridge, Lasso and Elastic Net. Each of these models have various characters that make them suitable for specific datasets, so we will level the playing field by running our data through a standard pipeline consisting of a standard feature scaler that helps us normalize the inputs of the model.

The first exploratory data analysis technique we employed was a full scatter-plot matrix of all the variables for the model that give us a high-level overview of their respective distributions and their correlations to the other variables in the dataset.

After we concluded the dataset was suitable for such a research project, we employed a standard test harness for each of the models, so that they are evaluated on an apples-to-apples basis.

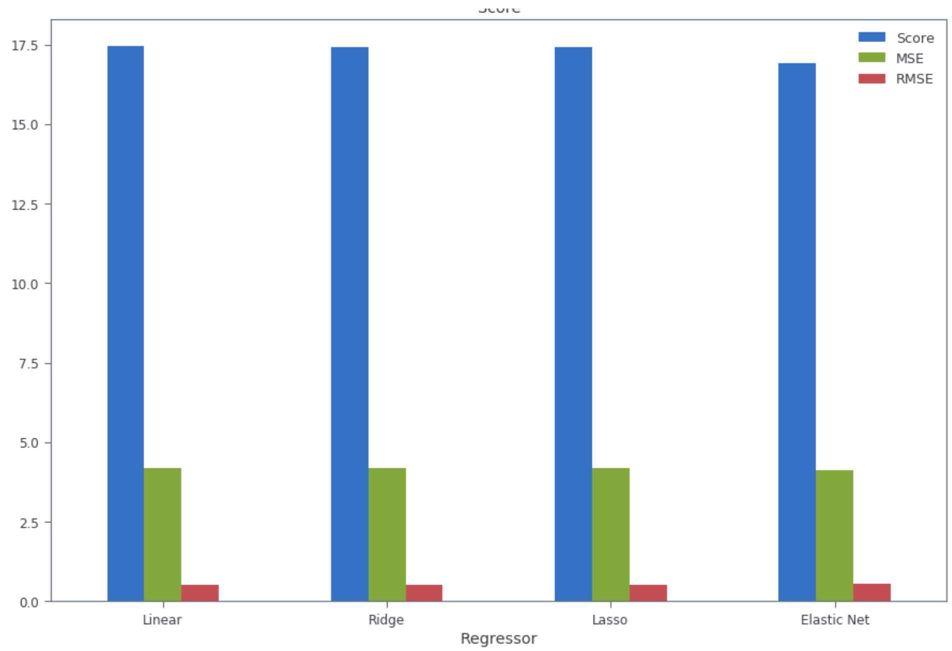
CONCLUSION

After execution and performance critiquing of the four models in question, there were two models that stood out to us as excellent and equal performers, Linear and Lasso regressors. These had performance indicators that vary at the ten-thousandth decimal place, which is a tiny difference. The Ridge and Elastic Net models did not perform as well as we hoped, although a relative comparison shows how tightly all four of these models were in terms of overall performance.



	Score	MSE	RMSE
Regressor			
Linear	17.4389	4.17599	0.519423
Lasso	17.4283	4.17472	0.519715
Ridge	17.4086	4.17236	0.520258
Elastic Net	16.9186	4.11322	0.533761

Our recommendation to management is to move ahead with productizing the version of the Linear model developed in this study to start tracking the predictions, collecting more data that we can feed into the model to drive accuracy up further, as well as serve as a baseline of comparisons for manually generated housing valuations. We would be well served to start assessing the models accuracy relative to a human doing the same task and compare the overall expenditure of the model in production versus keeping a full in-house staff to produce these results. Our belief is that management will see substantial reductions in operating expenses (therefor, increased revenues) in a short period of time.



For further information and details on this study was conducted, please visit the Colabratory Notebook.