HOME PRICE PREDICTION OBJECTIVES

EXECUTIVE SUMMARY

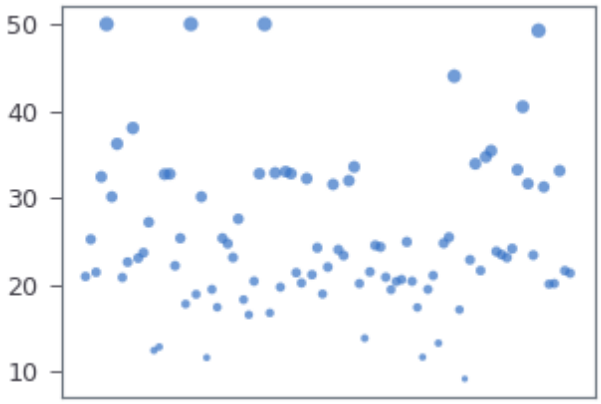
To gather an accurate prediction for a specific home valuation many variables must be taken into consideration. Traditionally, some of the higher impacting variables are the size of the home in question, which is typically a reflection of rooms in the house, the area in which it is located, and the school 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 six 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 analysis due to the continuous nature of our response variable, median house price. The specifics selected for this study are models in the linear, multiple, and colinearly regression algorithm space, as well as models in the tree-based algorithm space. Each of these models have their own nuanced data preprocessing requirements, as the non-tree based approaches perform better when the data is of a

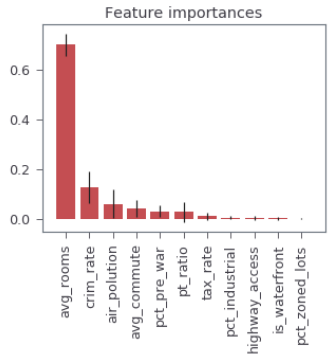
monotonic scale; however, the tree-based algorithms perform better with data of its naturally occurring scale. The non-tree-based methods will have a standard pipeline, and the tree methods will execute on the data as-is.

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.

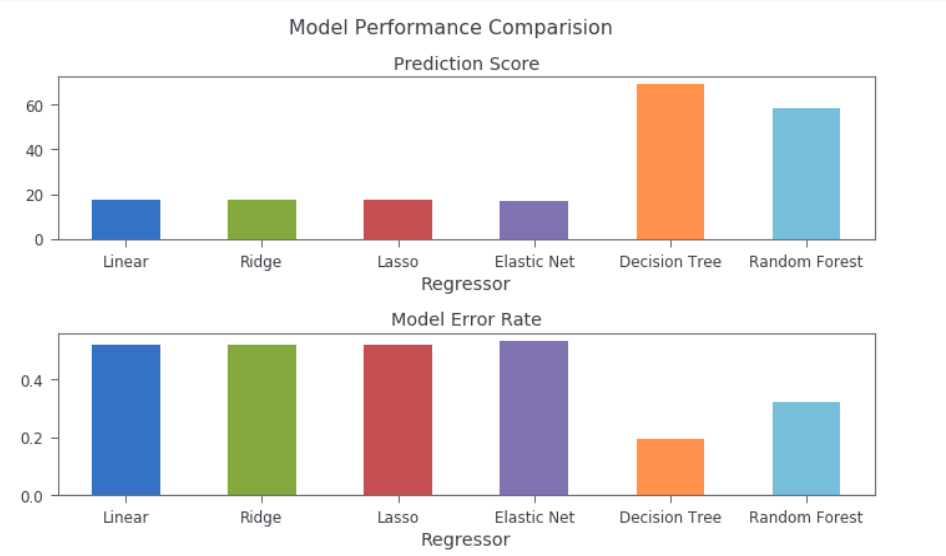
CONCLUSION

After execution and performance critiquing of the two classes of models in question, there is no doubt that this problem is best suited for a tree-based regression approach. The tree-based approaches outperformed the non-tree-based approaches by several orders of magnitude in overall prediction accuracy as well as smaller margins of error.



Having narrowed our decision down to one of the two tree-based methods, we analyzed the importance of the features the model needs to perform well, and overwhelmingly the results were clear that the average rooms (70%), crime rate (13%) and air pollution (5%) were the most influencing factors in determining the median value of a given home.

For the sake of completeness, we should advise management that our extensive research in evaluating six different learning algorithms in two distinct classifications. This in-depth research has provided us great confidence that our recommendation to move forward with the decision tree algorithm. With this system in production as your primary source of housing price predictions, it is our belief is that management will see substantial reductions in operating expenses (therefor, increased revenues) in a short period of time given the accuracy of the system. For an overall breakdown of the model performance, please refer to the performance figure.



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