Christopher Rico
MSDS 422
Assignment 3
10/13/19

Introduction

To assess the value of real estate, many different factors must be taken into consideration. Several factors, such as home size, neighborhood, and school district are typically strong predictors of home value. However, many of these valuation metrics are time-consuming to research manually.

By leveraging a machine-learning backed model to predict home value based on data collected from home listings, it may be possible to reduce the amount of research needed to estimate home value. This would reduce the operational overhead of the brokerage firm and allow for more accurate home value predictions when machine learning predictions can complement more traditional valuation techniques.

Research Design and Statistical Methods

The dataset provided is a collection of about 500 census tracts within the Boston metropolitan area in 1978. The original objective of this dataset was to examine the effect of air pollution on home value while controlling for 13 explanatory variables such as proximity to the Charles river, home age, highway access, etc.

	log_median_value
avg_rooms	0.632536
pct_zoned_lots	0.363396
avg_commute	0.342527
is_waterfront	0.158569
pct_pre_war	-0.455029
highway_access	-0.486818
pt_ratio	-0.499433
air_pollution	-0.513431
crime_rate	-0.530001
pct_industrial	-0.543195
tax_rate	-0.566214
pct_poor	-0.809234

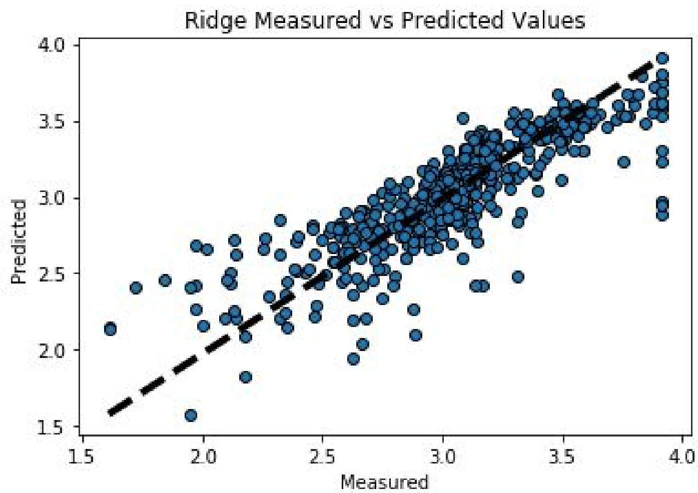
Explanatory vars and correlations with log median home value

First, we chose to remove the 'neighborhood' variable from the dataset, which leaves us with 12 explanatory variables from which to build and test several linear regression machine learning models. The ultimate goal of this work is to compare different machine learning models and recommend the best one to a real estate brokerage firm who wishes to augment their conventional methods for assessing market value of residential real estate.

Programming Work

This analysis was performed in the cloud using a Google Colaboratory notebook loaded with a variety of industry-standard data science packages: numpy, pandas, matplotlib, and scikit-learn. To run the analysis, the data were used to train different machine learning linear regression models from the scikit-learn library: linear regression, ridge regression, lasso, and elastic net. Each of these models have various characteristics that make

Christopher Rico
MSDS 422
Assignment 3
10/13/19



them suitable for different types of datasets, so to ‘level the playing field,’ we pre-processed the data by running it through a standard feature scaler that helped normalize the input variables.

Each regressor was then tested for its ability to predict home values, and validated using a multiple k-folds method. Root mean square error was used as a metric to compare the accuracy of the regressors. The model comparison was run twice: once using median home values as a response variable, and another time using the log of median home values.

Results and Recommendation

My recommendation to the brokerage firm is to use ridge regression as a market modeling technique, while using the log of the median home values as a response variable.

	Regressor	Train_RMSE	Test_RMSE	Diff
0	Linear	0.185761	0.223353	0.037592
1	Ridge	0.185761	0.223338	0.037577
2	Lasso	0.245910	0.286836	0.040926
3	Elastic Net	0.204515	0.235950	0.031435

This regressor strikes a good balance between tightness of fit, while still avoiding the predictive error that other models exhibit for higher-priced homes. The root mean square error of ridge regression is marginally better than those of other models when using log_median_value as a response variable, which indicates that the model is able to most accurately predict more home prices than other models.

On a different note, I would also suggest that the brokerage firm expand its dataset to be able to train a more accurate model. 500 points is, frankly, a rather small set of data to make meaningful predictions from.

A shared version of the interactive Jupyter notebook used to run this analysis can be found at:

https://colab.research.google.com/drive/1fFnwo2h-9qEEyWcMu-GeC9mutMrUUojs