Dalya Adams
MSDS 422 section 59

Boston Housing Price Prediction

The real estate industry is a very old industry, one which has been based on the intuition of those investing for hundreds of years. While the industry has done well based on intuition historically, machine learning can complement this intuition greatly. Utilizing machine learning to capture and quantify patterns and relationships between features, in conjunction with the knowledge of humans, results in an increased accuracy and knowledge of the real estate market in an area.

In looking at different features connected with 506 census tracts in the Boston metropolitan area, we aim to predict the median value of homes in each tract. The features utilized in this analysis are: air pollution, crime rate, percent of land zoned for lots, percent of business that is industrial/non-retail, whether the tract is on the Charles River (or not), average number of rooms per home, percentage of home built prior to 1940, the weighted distance to employment centers, accessibility to radial highways, the tax rate, pupil/teacher ratio in public schools and the percentage of the population of lower socio-economic status. In moving from intuition-based decisions to data driven decisions, the selection of the appropriate model is paramount. A poor model selection can lead to inaccurate predictions, which can undermine the confidence in utilizing machine learning in the future. In this study we evaluate 6 different modeling methods, while tuning the hyper parameters, to identify the most accurate model for complementing conventional methods for assessing the market value of Boston's residential real estate.

In building our models, we target the median value of each census tract as the variable we are predicting. We evaluate the accuracy of our models by the root mean squared error (RMSE) of the prediction and select the model with the smallest RMSE. The RMSE provides us an idea of the average error of the prediction from the actual value. By selecting the model with the smallest RMSE, we ensure that the predictions we provide to the real estate firm are the most accurate. In considering our target audience, which is a real estate firm which is only beginning to dabble in machine learning, we initially aim for linear models, which provide interpretability and move towards random forests, which provide increased accuracy. Since the dataset is complex, we also test models which prevent overfitting, and allow for easier generalization, an important feature in complex markets like real estate.

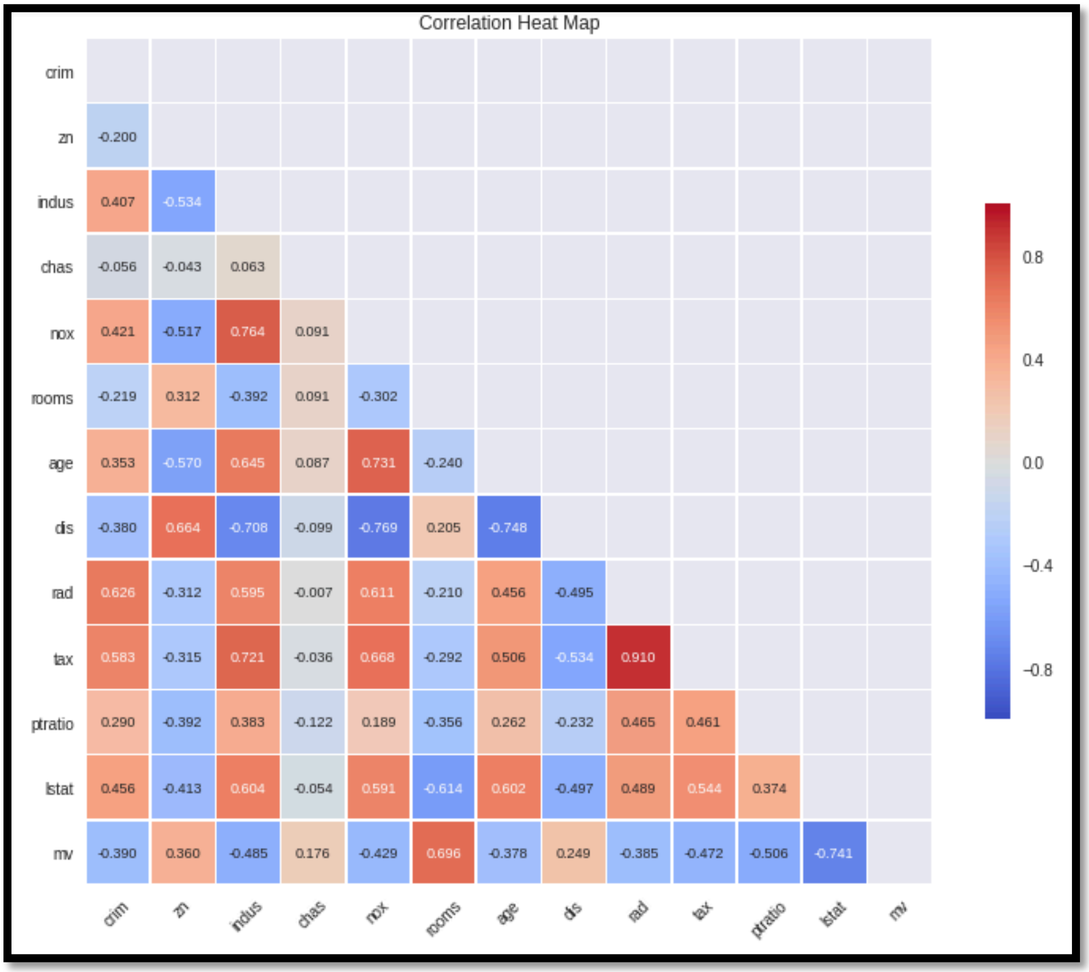
In modeling, we compare a linear model, ridge regression, lasso regression, elastic net, decision tree and random forest methods. The ridge, lasso and elastic net are each evaluated with the alpha hyper parameters of 0.1, 1.0 and 0.5. The random forest is evaluated with standard parameters and tuned hyper parameters. We compare all of these models against each other to determine which is the most accurate, based on the smallest RMSE. In utilizing machine learning in conjunction with conventional methods for assessing the real estate market, the tuned random forest results in the smallest RMSE.

The untuned Random Forest has a RMSE of 4.14 which equates to an average error of $4,139.65 in 1970 dollars. While this seems like a small error in 2018 dollars, the median value of a home, based on the model, is $33,848.90 in 1970 dollars. With this intercept, an error of approximately $4,000 seems much more concerning. Despite this, the model provides us with great interpretability from the coefficients. In general, the median value of a home is decreased by higher: crime, pollution, age, distance from employment centers, tax, pupil/teacher ratio, percent of lower socio-economic status and percent of nonretail/industrial businesses. The median value of a home is increased by: being on the Charles River, having more land zoned for lots, more rooms, and better accessibility to radial highways. The Feature Importance from the Random Forest models can be found in the appendix, as well as the RMSE of all tested models.

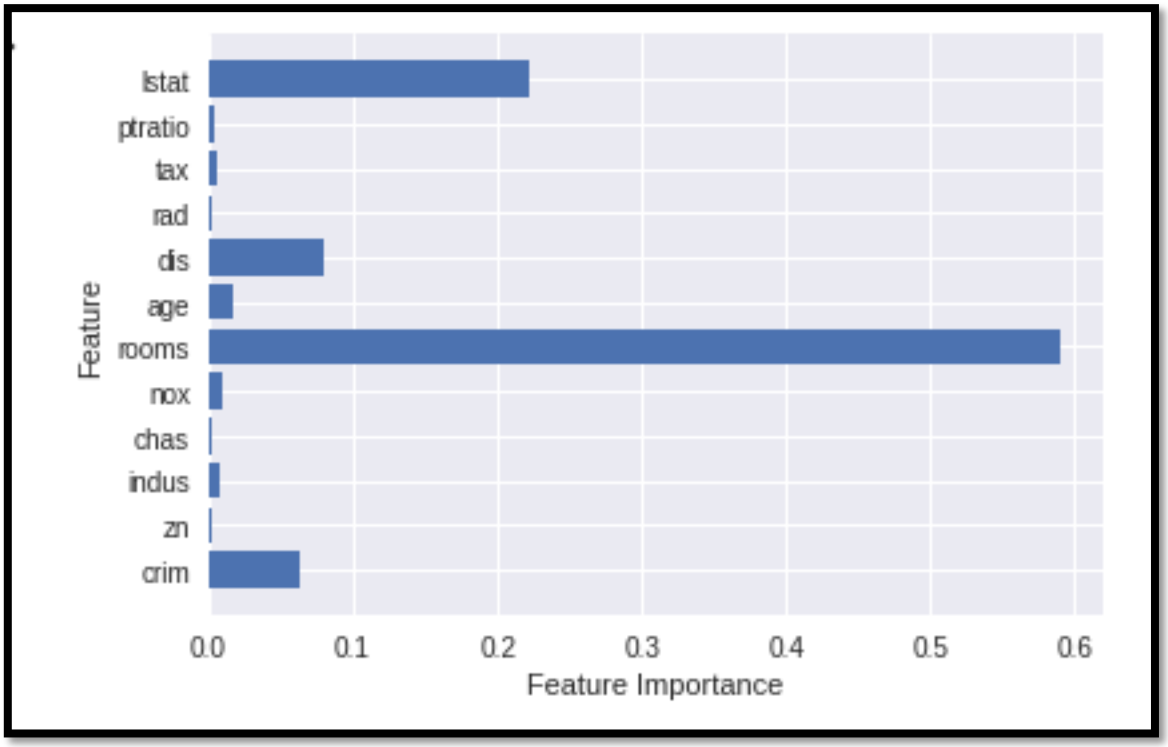
Appendix

Model	**Average RMSE**
Linear	5.15
Ridge (a=0.1)	5.14
Lasso (a=0.1)	5.13
Elastic Net (a=0.1)	5.08
Ridge (a=1)	5.085
Lasso (a=1)	5.52
Elastic Net (a=1)	5.29
Ridge (a=0.5)	5.10
Lasso (a=0.5)	5.26
Elastic Net (a=0.5)	5.15
Decision Tree	5.83
Random Forest (Untuned)	4.14
Random Forest (Tuned)	4.22

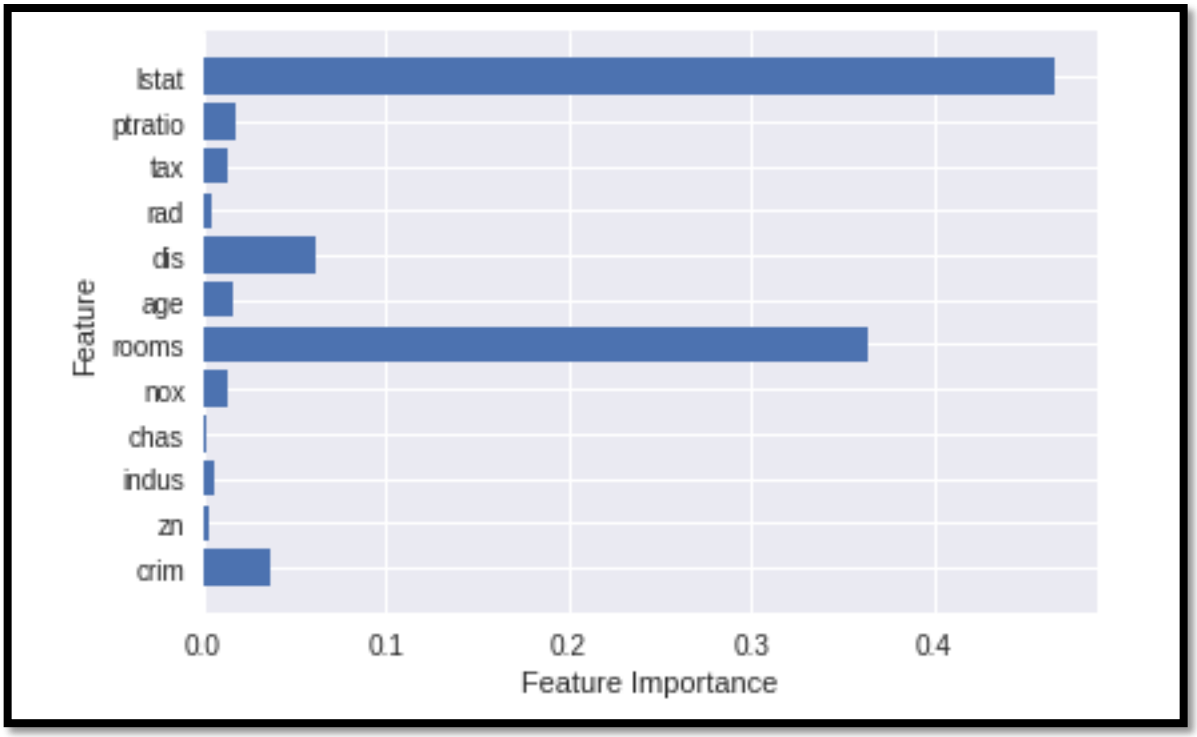
Correlation Heat Map



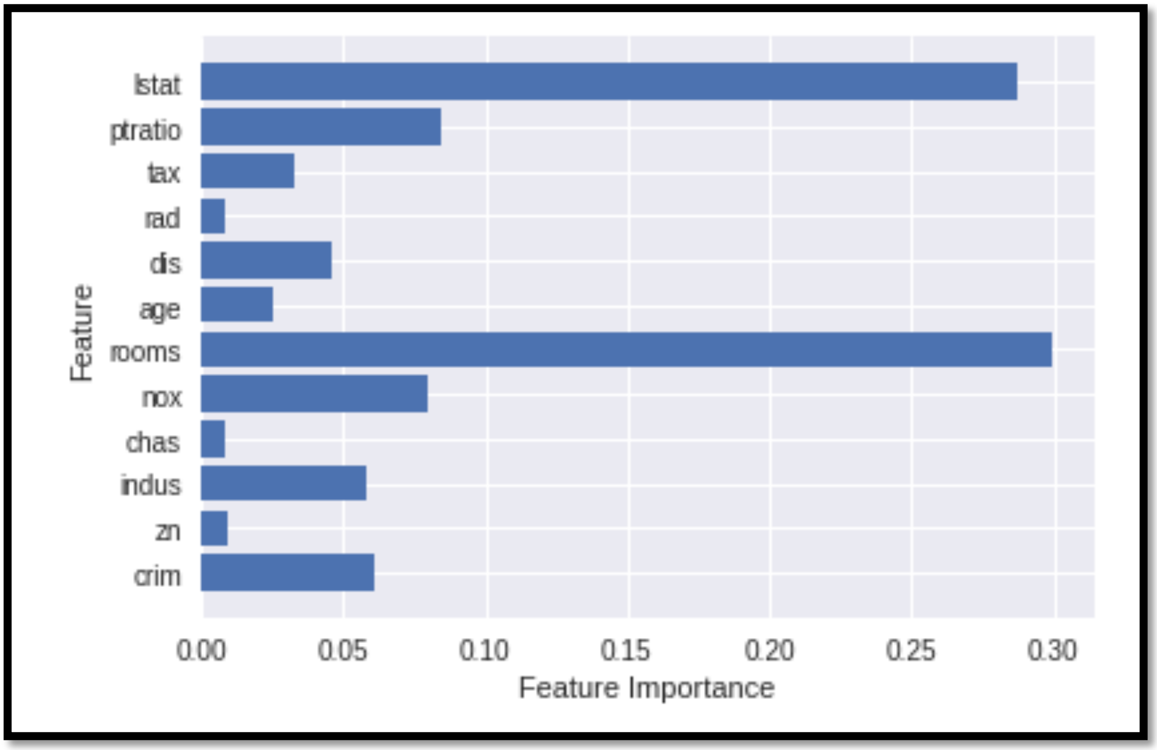
Feature Importance – Decision Tree



Feature Importance – Random Forest



Feature Importance – Tuned Random Forest


lstat
ptratio
tax
rad
dis
age
rooms
nox
chas
indus
zn
crim
Feature
Feature Importance