Intro

To generate a sufficiently accurate model in predicting home prices in Richmond and Pittsburgh, we first scrutinized the data set for any inconsistencies and trends. Several methods were then compared against one another based on accuracy and simplicity. We identified the most useful features from the given data set. A discussion of model selection and interpretation of the results concludes the analysis.

EDA

We will first identify issues with the data set. Many of the given houses had “NA” assigned under the “Fireplaces” category, and many others had “0”. We assumed these to be describing the same situation: a house without any fireplaces. Accordingly, any houses that initially had a value of “NA” for this feature were reassigned zero.

Another problem presented itself in the “Lot Area” data. A number of homes in both the training and test sets had a lot area of 0, which appeared to be quite strange. Upon further inspection, every home with zero lot area was also labelled as a condominium. Lot area indicates total ground space as well as any square footage contributed by a backyard or driveway. The term condominium sometimes refers to an apartment, in which case zero lot area would point towards stacked dwellings which contribute no additional lot area to a building. We ultimately made no adjustment to this feature.

We also made the decision to not consider zip code in any of our models. We figured that if a trend existed in the zip codes, like if higher numbers indicated pricier homes, it would be due to pure happenstance and not actually because of effects from the zip code number itself.

Upon correction of these problems, we analyzed the correlation matrix for all continuous features. The features that initially appeared to be most important in predicting price were Square Footage, Total Rooms, Bedrooms, and Bathrooms. We took into account the fact that there is certainly some collinearity between these last three predictors before performing our analysis.

Methods

We considered a number of methods to tackle this regression problem. For each model, we calculated the MSE on the test set. To establish a baseline, we predicted each home price in the test set based on the mean home price in the training set. We then applied models in three distinct groups: dimension reduction, linear, and tree-based. The first group consisted of LASSO, ridge regression, PCR, and PLS. The second group contained of only multiple linear regression. The third and final group applied a single regression tree, bagged trees, boosted trees, and a random forest to the data.

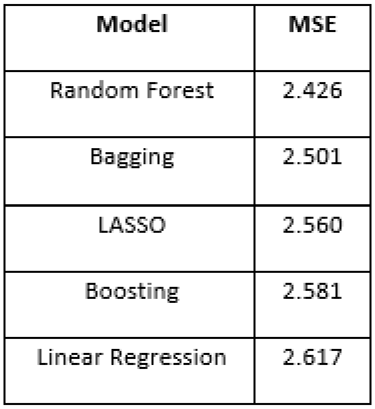
Upon applying these models, we chose three with the lowest MSE, and found the most important variables using out-of-bag estimates. From there, we trimmed down the “best” models to only use the most important features. We hoped that doing so would decrease reliance on less useful features and increase interpretability without sacrificing much accuracy.

Summary

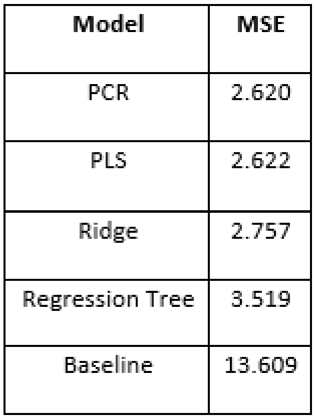
We found that three of the tree-based methods were among the most successful. The dimension reduction strategies were comparatively ineffective, with the exception of LASSO. The two leftmost charts at the end of this section detail the test MSE of each method using all the features, in tens of billions of dollars. Each of the models we used performed significantly better than the baseline model. Indeed, the eight best models were nearly identical in terms of MSE, with only the single regression tree lagging behind in any noticeable way.

With this in mind, we decided to pursue the bagging, boosting, and random forest methods in more detail. In the interest of simplicity, we identified the two or three most important features from each model. The rightmost table at the end of this section identifies these selections resulting from the out-of-bag estimates. Theoretically speaking, it makes perfect sense that bagging and random forest would have nearly identical MSEs when such a small subset of features is initially considered. In this instance, boosting is clearly worse off than these other two methods.

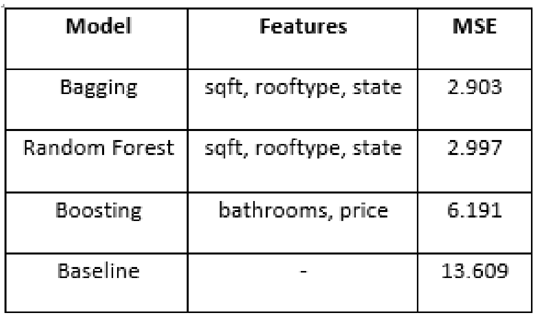
To generate our final predictions, we accordingly used bagging with only the following features: square footage, roof type, and state. The square footage feature is self-explanatory. Upon analysis of a single regression tree and linear regression, houses in Pittsburgh tend to be cheaper than in Richmond. Additionally, out of the four roofing materials, slate tends to indicate more expensive houses.



Model	MSE
Random Forest	2.426
Bagging	2.501
LASSO	2.560
Boosting	2.581
Linear Regression	2.617



Model	MSE
PCR	2.620
PLS	2.622
Ridge	2.757
Regression Tree	3.519
Baseline	13.609



Model	Features	MSE
Bagging	sqft, rooftype, state	2.903
Random Forest	sqft, rooftype, state	2.997
Boosting	bathrooms, price	6.191
Baseline	-	13.609

Conclusion

The final model appears to be quite effective, for a couple of reasons. Each feature has justification for its presence. The significant difference in prices between the two states could be attributed to cost of living. Square footage, which consistently appeared as a top predictor, likewise makes perfect sense to be part of the final model. All else equal, more space is going to cost more money. Roof type seems fair to include because a slate roof is likely an indicator of luxury, whereas the other three types are much more common.

By far the most difficult part of this analysis was discerning the most useful features. A few were excluded outright, but nearly any feature on its own could plausibly predict price. It of course makes sense, for example, that houses with more bedrooms will cost more than houses with fewer bedrooms. The question to address, though, was whether this difference was significant. Using the built-in out-of-bag estimates in R proved to be sufficient. This is not a perfect method, however. Instead of considering changes in accuracy whenever one feature is not present in the bootstrap sample, one could instead analyze the accuracy whenever related pairs or trios of features are absent, like bedrooms, bathrooms, and total rooms. This may be able to better address collinearity between features.

Ultimately, we are quite satisfied with the final model because of its good predictive power and relative simplicity. It is just as dependable as models that contain all the predictors and is sufficiently interpretable.