In this assignment, we are tasked with analyzing historical data concerning property information in order to predict current house prices in both Pittsburgh and Richmond. The overarching purpose of these predictions is to identify over/under-priced homes based on a comparison between list price and our predicted price. Throughout this process, we partake in both inference as well as prediction in order to create models that have high predictive accuracy, while also keeping interpretability in mind.

Upon parsing through the data, we notice that there are multiple oddities present in both the training and test sets. The first is that all of the values for fireplace in the "VA" observations are missing. We confirm this by selecting all rows in which the fireplace takes the value "NA" and comparing to all the rows in which state takes the value "VA". The selected rows in each case are identical, thus proving that if and only if the state is "VA", the fireplace value is "NA". To resolve this issue, we cannot simply fill the null fireplace values with the overall mean number of fireplaces because this would create problems with interpretability. Instead, we remove the fireplace predictor entirely. Although this proves to be a problem with the dataset, we resolve it without great difficulty.

Another data oddity we encounter is the sole observation belonging to the class "MOBILE HOME". This observation is discovered when it causes an error in our Leave One Out Cross Validation (LOOCV) multiple regression model. Because there is only one mobile home in both the train and test data, we elect to remove it. We believe it is favorable to remove this point because it is far more likely to be a high leverage observation or outlier than it is to provide value in modeling the vast majority of homes, which are single-family.

Below we can see relative frequencies of each value of "desc", i.e., the different types of homes being analyzed in train and test sets.

Figure 1

Train Set House Type

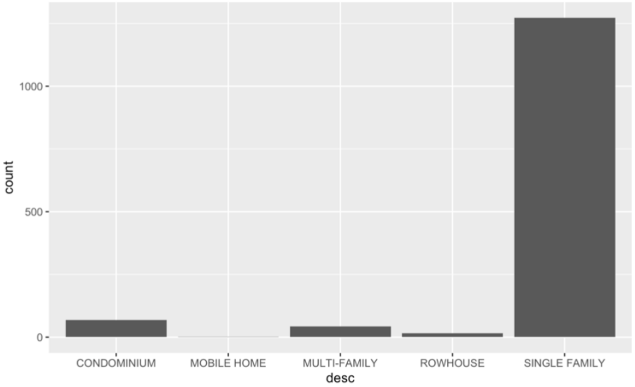
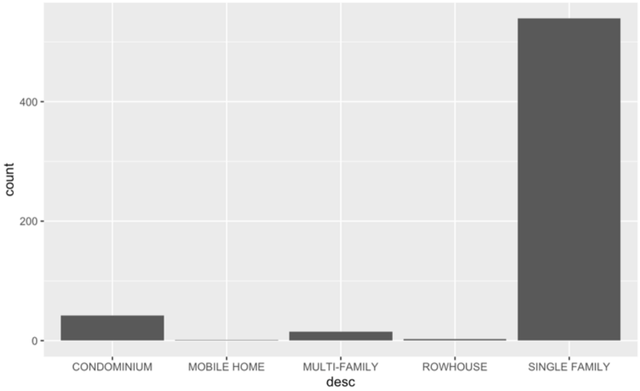


Figure 2

Test Set House Type



In can clearly be seen that mobile homes are not of great importance, and single-family homes comprise the lion's share of observations.

Besides the oddities noted above, the data are as expected. We discover via simple histograms of the continuous variables that most of the variables are right-skewed. This makes sense because very large and expensive houses tend to have well above average numbers for most metrics, however smaller houses cannot have comparably small or negative values. This results in long right tails created by the few multimillion dollar or otherwise unusual homes. For this reason, we do not remove any other observations.

We construct many different models using the training data. Some models utilize a validation set approach in order to approximate the generalizability error. Others rely on cross validation. The two metrics used for model evaluation are MSE and R^2. Because the MSEs are very large, in part, due to the fact that houses are relatively expensive, R^2 gives us a more interpretable measure of model fit.

The first model we fit is multiple regression using the validation set approach. We chose a 75/25 split (75% of data to training set, 25% to validation set) for building and testing this model. We utilize the same split for all of our models involving validation set approach. This model performs mediocre relative to our other models. We improve our results by utilizing LOOCV on the entire data set instead of using the aforementioned test/train split. The R^2 jumps from 0.84 to almost 0.87 when making this switch and the MSE drops from nearly 22 billion to 18 billion. This is likely due to the fact that there is less bias when using LOOCV. We then run both five- and ten-fold cross validation, producing nearly identical results to those of LOOCV.

After running these multiple regression models, we use forward selection to build the best model of each size. Because all of the categorical predictors have been coded as factors, each of their respective levels shows up as its own variable. This causes there to be a 66 variable maximum model. Forward selection performs worse than the multiple regression models. This is odd because theoretically, at the very least, forward selection should just choose the maximum model and perform identically to multiple regression. This discrepancy is likely due to the fact that there are three "linear dependencies" in the data, and we are only able to run up to a maximum 63/66 predictor model.

While looking through the order of predictors added, we notice that the zipcodes, as well as avgincome are the last predictors added to the model. This indicates that in forward selection they are less important than the other variables. Zipcode and avgincome, perform very similar purposes: both quantify and categorize a house by the area it is located in. However, zipcode is likely only important because it associates a house to the average wealth of an area. This is exactly what avgincome does; it measures the average income within a given zipcode. Considering both of these metrics are incredibly similar, as well as relatively unimportant in forward selection, we elect to remove zipcode from some of the future models we run.

We next build lasso and ridge regression models using the validation set approach. These models perform comparably to the multiple regression models.

Next, we utilize PCR and PLS. Both perform poorly with R^2 values of 0.82 and MSEs of nearly 25 billion. This is probably because desc had to be removed from the model due to another error relating to factor variables. Desc had been the most important predictor in forward selection, so it is likely the culprit for PCR and PLS performing poorly.

After PCR and PLS, we run a GAM with smoothing splines on all of the quantitative predictors. This model performs relatively well with an R^2 of 0.85 and an MSE of 20 billion. Next, we try a regression tree. This tree performs the worst out of any of the fitted models with an R^2 of 0.72. Individual trees have very high variance, and this likely hampers performance. Pruning does not improve the test MSE either.

Bagging is the first model that produces results better than those of LOOCV regression. We attain an R^2 of 0.87 and an MSE of 17 billion. We then improve this even further by implementing a random forest model (reduce mtry to 14/3) and achieve an R^2 of 0.88 and MSE of 15.9 billion. This change in mtry is advantageous because it reduces correlation between trees which reduces the overall variance of the model.

Last, we build a boosting model. Like random forest and bagging, we utilize the test/train split to measure the generalizability error of this model. Boosting achieves an R^2 of 0.92 and an MSE of 11 billion, the most predictively accurate model by far.

When building bagging, random forest, and boosting models, our best three models, we first build them with zipcode included and then again without it. To our surprise, the models not only perform better with zipcode, but zipcode is actually one of the most important predictors in these models. This contrasts with the forward selection results. In both bagging and random forest, zipcode is the most important predictor when measuring via %incMSE (a measure of the MSE increase when said variable is removed). It is the third most important predictor when using IncNodePurity, behind sqft and bathrooms. In boosting, zipcode is also the most important variable when measuring via relative influence. Surprisingly, desc appears unimportant using these metrics. This is likely because it is correlated with other predictors.

Based on R^2 and MSE, boosting, random forest, and bagging perform first, second, and third best, respectively. This indicates to us that certain nonlinear models perform better than the standard multiple regression models. Although these were the best models that we were able to produce, I would not trust them to be able to isolate over/under priced homes. This is because, even though we were able to get the R^2 value up to 0.92, the MSE is still over 11 billion. This corresponds to an average difference between the predicted value and the actual value of over $105,000. This value is found via rooting the MSE. This is a huge amount of variability and error considering the average price of a home. Any over/under-pricing that occurs will likely only be of a few thousand dollars, so a model that misprices by 100k on average is of little use for generating marginal edge in housing markets.