Technical Report

Introduction/EDA:

With a historical dataset of housing prices and detailed property information for houses in Pittsburgh, Pennsylvania and Richmond, Virginia, I sought to understand the most important factors in house value in each state as well as tested out models to predict the prices of homes. Through these analyses, I would help the team of realtors identify over or underpriced homes.

The data given had 1400 observations, with 687 missing values, all of which happened to be a part of the fireplaces predictor which we subsequently removed. zipcode was also omitted as it had 45 levels, which would make for difficult interpretability. Also, AvgIncome already records the average income for zipcode. yearbuilt was used as a categorical variable, which was adequate since the distribution of years is approximately normal. VA houses have a slightly higher median price than PA houses, but both states have their share of extremely high housing prices, causing the price distribution to be severely right-skewed. As expected, there is a moderately strong positive correlation between totalrooms, bedrooms, bathrooms, and sqft. These variables are also positively correlated with price. 90.8% of the houses are single-family houses.

Methods:

For the model building and fitting, train.csv was split into a training set (70% of observations) and a testing set (30% of observations). By best subset selection, the subset size that yielded the lowest test MSE (14.1 billion) was 21 predictors, which included all the levels of the categorical predictors exteriorfinish, rooftype, and desc. As a result, all predictors were considered in the following models, with a few exceptions. First, multiple regression was performed with all predictors. At a 5% level of significance, all predictors were significant except for totalrooms. Another multiple regression was then fit, this time omitting totalrooms and desc since there are too few observations of Mobile Homes within desc. Multiple regression yielded a 14.3 billion test MSE.

Next, models were built through ridge regression and lasso. For both of these models, a range of lambda values was considered for a tuning parameter, and the one yielding the lowest test MSE was chosen. The value of lambda used for ridge regression was 2848.04 and the value of lambda used for lasso was 1629.75. Both models produced a test MSE of 13.5 billion. Lasso zeroed out the predictors descMOBILE HOME, exteriorfinishConcrete, exteriorfinishLog, rooftypeROLL, and totalrooms.

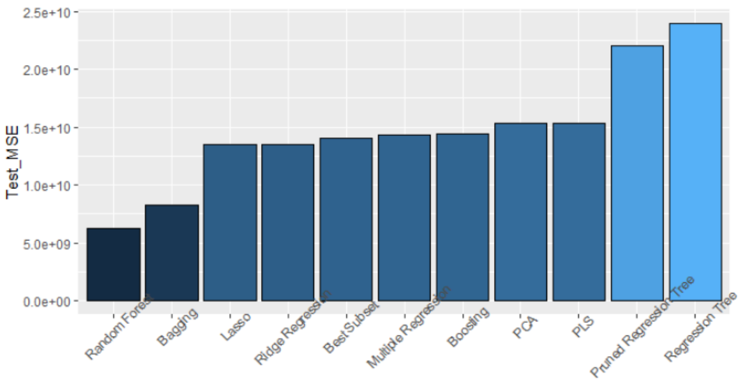
For principle components analysis and partial least squares, only state, basement, bedrooms, bathrooms, sqft, lotarea, and AvgIncome were considered, as many of the categorical predictors were difficult to interpret with this method. Also, these selected variables were significant in multiple regression. Both PCA and PLS yielded a 15.4 billion test MSE and 5 components were selected.

In the regression tree, all predictors were considered. However, the variables actually used in construction were sqft, bathrooms, rooftype, state, and AvgIncome, yielding a test MSE of 24.0 billion. The regression tree was then pruned, using cross-validation to select a tree of size 10 which used the variables sqft, bathrooms, rooftype, and state. Pruning the tree improved the test MSE to 22.0 billion.

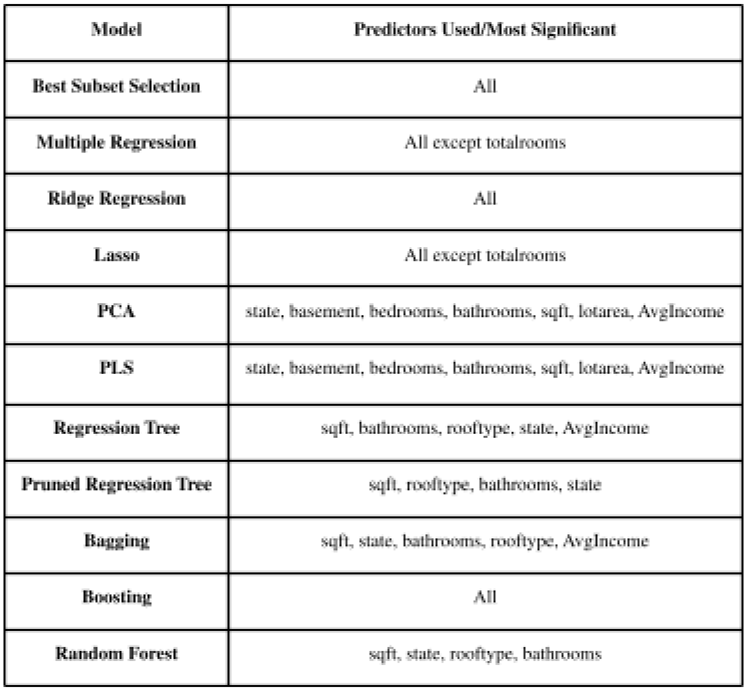
Bagging was performed next. Observing the variable importance plots, sqft, state, rooftype, and AvgIncome were found to have the highest %IncMSE. In other words, the exclusion of these variables produced the greatest mean decrease of accuracy in predictions on the out-of-bag samples. In decreasing order, sqft, bathrooms, rooftype, and state had the highest IncNodePurity, which is an average measure of the total decrease in node impurity that results from splits over that variable. sqft had a significantly higher IncNodePurity than the other variables, and variables past state did not show much difference in IncNodePurity. Bagging produced a test MSE of 8.2 billion. From boosting, the minimum test MSE of 14.4 billion was produced when lambda = 0.04.

To build a random forest, mtry = 4 was used, since it is close to the 13 (the number of predictors) divided by 3. The same four predictors as those from boosting have the highest %IncMSE, now ordered as sqft, state, rooftype, and bathrooms. The IncNodePurity is still lead by sqft, with bathrooms closer behind, followed by totalrooms, and lotarea. The random forest with mtry = 4 produced a test MSE of 6.2 billion. Multiple random forest models were looped over varying values of mtry (1 to 13), and it was found that random forest with mtry = 4 does yield the lowest test MSE.

Summary of Results:



The chart to the left shows the relative values of test MSE for each model used in the analysis. It is clear that random forest and bagging perform best. The test MSEs for lasso, ridge regression, best subset selection, multiple regression, boosting, PCA, and PLS are all similar, with the shrinkage methods performing the best of the bunch and dimension reduction methods performing slightly worse. The pruned and unpruned regression trees have the highest test MSEs out of the models analyzed. From the table below, it is shown that the most important variables are state, bathrooms, sqft, rooftype, and AvgIncome, as these variables are significant in at least 9 of the 13 models and also have the highest %IncMSE and IncNodeImpurity found in bagging, boosting, and random forest.



Model	Predictors Used/Most Significant
Best Subset Selection	All
Multiple Regression	All except totalrooms
Ridge Regression	All
Lasso	All except totalrooms
PCA	state, basement, bedrooms, bathrooms, sqft, lotarea, AvgIncome
PLS	state, basement, bedrooms, bathrooms, sqft, lotarea, AvgIncome
Regression Tree	sqft, bathrooms, rooftype, state, AvgIncome
Pruned Regression Tree	sqft, rooftype, bathrooms, state
Bagging	sqft, state, bathrooms, rooftype, AvgIncome
Boosting	All
Random Forest	sqft, state, rooftype, bathrooms

Conclusions:

It can be safely concluded that state, bathrooms, sqft, rooftype, and AvgIncome are the most important variables for predicting housing price and that a random forest model with mtry = 4 and ntree = 500 gives us the most accurate model, likely because random forests perform well on data with lots of noise, much like this real-life example.

The most challenging aspect of this particular dataset was determining which predictors to use in each model, especially since best subset selection did not narrow anything down. By the time we began building trees, however, it became more clear which predictors were important. I trust my “best” model in the sense that I trust it to consistently outperform the other models in the face of noisy data. However, I do know that test MSE values are volatile depending on how the training and testing set is split, so I would not recommend the test MSE of the best model to be something to be advertised or cited. This could have been mitigated with cross-validation.

In the future, we could focus more on separating the observations from PA and VA. Rather than lumping PA and VA together and simply using state as a categorical predictor, future analyses could subset the states and find more in-depth similarities and differences in significant predictors and relationships for each respective state.