Predicting home prices in Solon, Ohio

Fouad Yared
Hunter College, Statistics 717: Multivariate Analysis, Fall 2020

Introduction

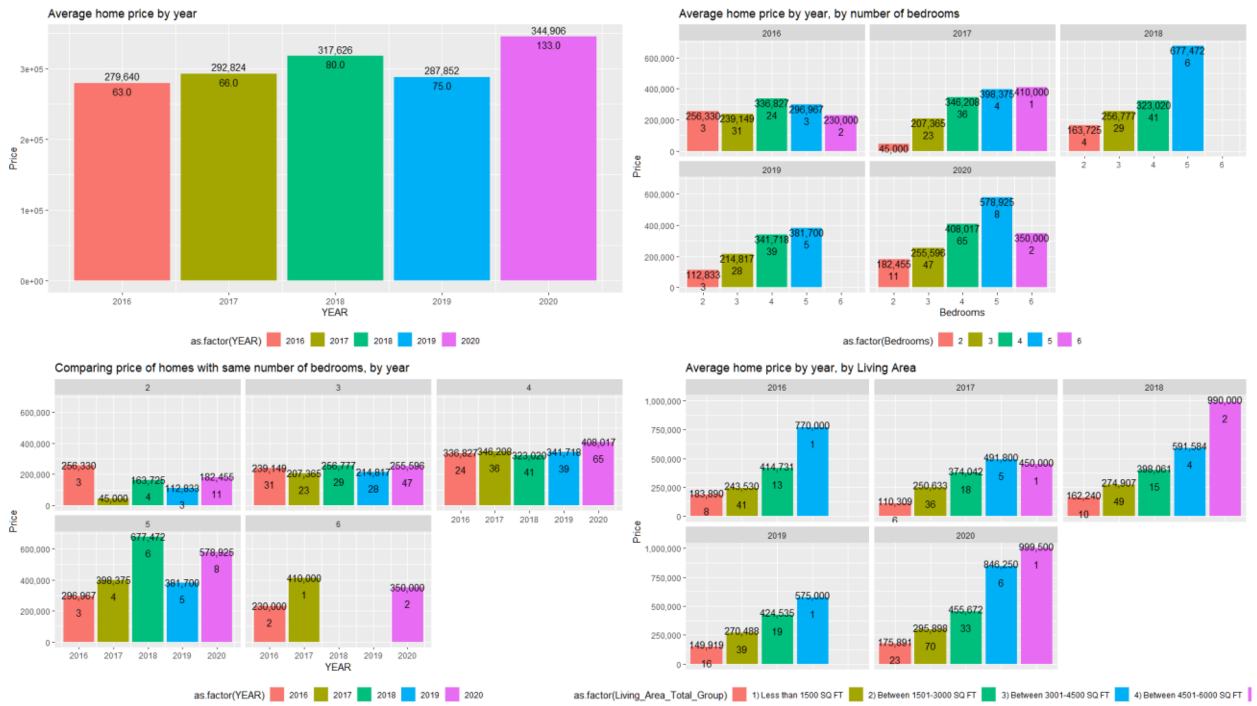
One of the consequences of COVID-19 is that the residential market has gotten more competitive. According to the Joint Center for Housing Studies of Harvard University, “August [2020] was the strongest month for home sales since 2006.” Since millions of Americans can work from home, they have reconsidered “their housing needs, especially in terms of space, location, and affordability.”

The purpose of this analysis will be to look at factors affecting sale prices of homes in Solon, OH between 2016 to 2020. Solon’s public school district is one of its main draws. The U.S. News and World Report ranks it as the fourth best school in the state and the top school in the Cleveland-Northeast Ohio region. According to the U.S. Census’s 2019 American Community Survey (5-year), Solon has a population of 22,947. Seventy-two percent of its residents are White, 13% are Black or African American, 11% are Asian (6% of the total population are Chinese), 3% are two or more races, and the remaining 1% is Other. Regarding ethnicity, 1.6% of the population is Hispanic or Latino (of any race). It is a middle to upper income area as the median income for families is over $133,000 and the mean income is over $171,000.

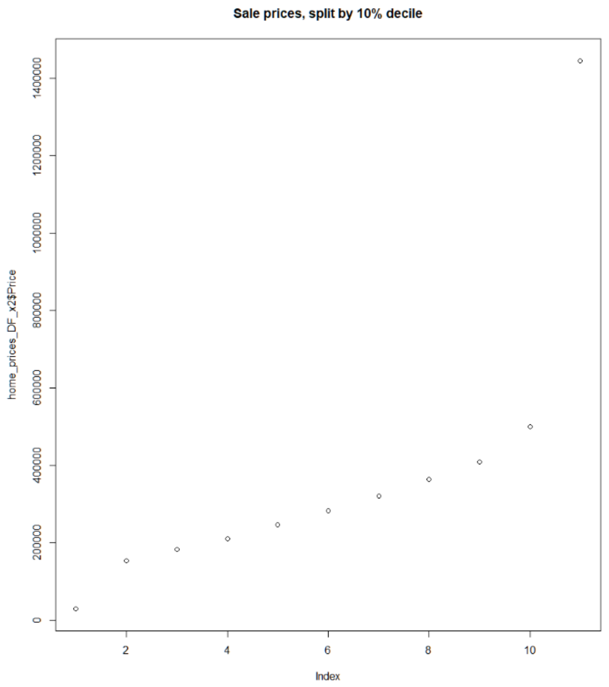
The data for this paper stems from two sources: Cleveland.com, a news website that has a searchable housing database (allowing users to see sale prices of homes and their addresses) and the county government, Cuyahoga County, publishes records on other housing attributes including the year a home was built, the total living area (in square feet), whether there is a basement, and more. To gather this data, addresses must be searched one at a time. For this paper, 417 records were collected. At least 15 records were collected for each quarter between January 2016 to June 2020. Twenty-six records were collected in the third quarter of 2020 and seventy records were collected in the fourth quarter of 2020.

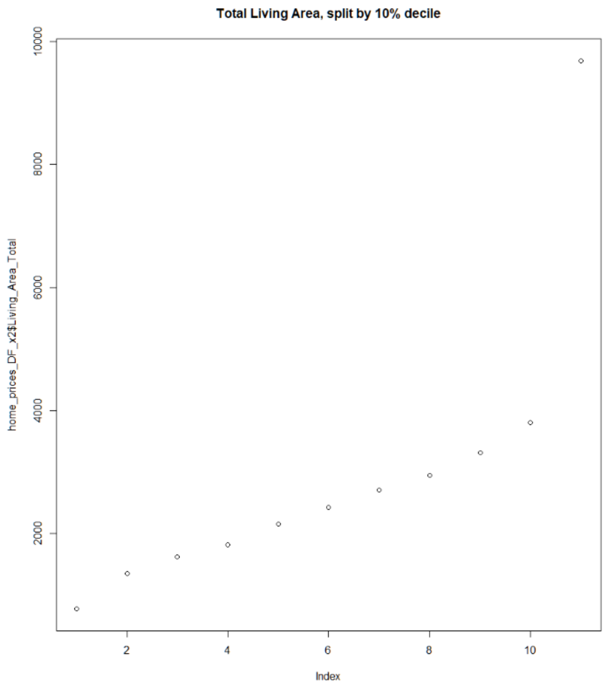
Data Exploration

Average values by year were observed to see how price changed over time. The top-left panel of the below plot shows that the average price tended to increase as prices decreased slightly in 2019 and rebounded in 2020. (The second row of figures show the count of homes in that group.) The top-right panel shows the average home price by number of bedrooms. Average prices each year tended to increase as the number of rooms increased. The lower averages for 5 and 6 bedroom homes in 2016 are likely attributed to the small number of homes meeting these criteria in this data. There could also be homes that have many bedrooms, but they may be small. In looking at the bottom-left chart, we see that when looking at a specific number of bedrooms over time, average prices tend to fluctuate somewhat. The average price for 2, 5, and 6 bedroom homes are not as reliable as those for 3 and 4 bedroom homes. With that said, there is an unusually high average home price for 5 bedroom homes in 2018 that is greater than any other year. This may be attributed to other factors, such as total square footage. The bottom right plot groups the total living area in homes in five classes: less than 1,500 square feet, 1501-3000 square feet, 3001-4500 square feet, 4501-6000 square feet, and more than 6000 square feet. Average home prices tend to rise as the living area rises. More robust figures are found in the 1501-3000 square feet and 3001-4500 square feet estimates.



The below plots show the sale price at each decile (in the left panel) and the total living area (in the right panel). There is a dramatic increase in sale prices and total living area when going from the 90% decile to the 100% deciles.





Preparing the data

Variables were classified according to their type in R. Ratio features were treated as numeric, categorical features were treated as characters, and ordinal features were treated as ordered factors. Examples of numeric features are basement area (in square feet) and total living area. There was some ambiguity in how discrete and time variables should be treated. Specifically, the number of bedrooms and the year something was built can be ordinal or numeric. They were treated as ordered factors as the range of values was fairly limited for these features.

Removing unnecessary variables

While the original data set has 48 columns, many were removed for various reasons. Living Area Total is a composite of Living_Area_1, Living_Area_2, Living_Area_Basement, and Living_Area_Upper. Full Bathroom and Half Bathroom were separate variables and were grouped into one. Eleven categorical variables were removed because over 90% of the values were the same (since we want variation within variables). Two variables were removed because they have a lot of missing values. Several address-related features were also removed as they cannot help with fitting models and generating predictions. Twenty-one variables remained after removing unnecessary ones.

Regrouping levels of categorical variables

To produce potentially better and more robust estimates, levels of categorical variables were combined with other levels when their count was low, when a best guess could be made on how to reclassify these levels. For example, Bungalow homes were combined with Ranch homes since there were only three Bunalow homes in the data set. (Four-plex homes also appeared three times in the data. Since there was not a similar level, there were left as is. The observations with four-plex homes were eventually removed because of singularity errors.)

Regrouping yearly variables

The year the home was built and the year the garage was built were recoded as the decade either structure was built as there is likely not as much variation year over year as there is between decades. An ordered factor was used for each of these variables.

Replacing missing values

There were five records that did not have a garage. Missing data on garage size and capacity were imputed with a zero. The year the garage was built was imputed with an arbitrary value, which was used at the first level in the ordered factor of the decade the garage was built.

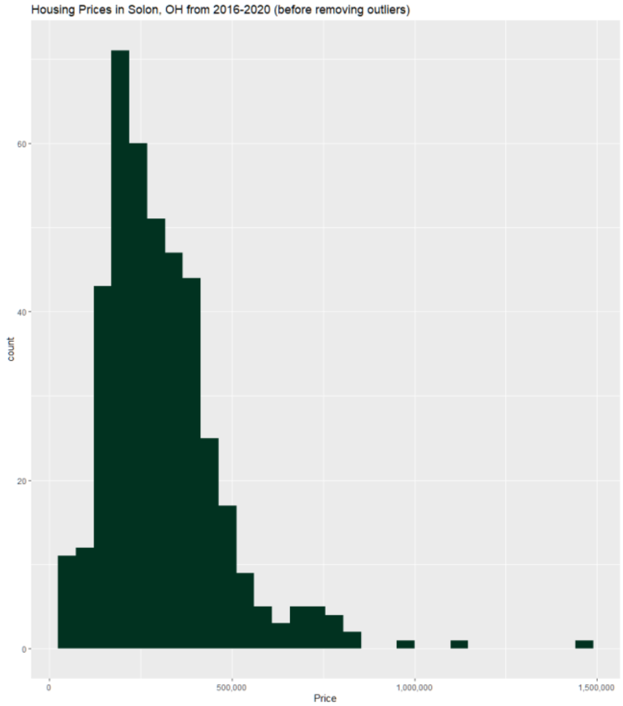
Evaluating outliers

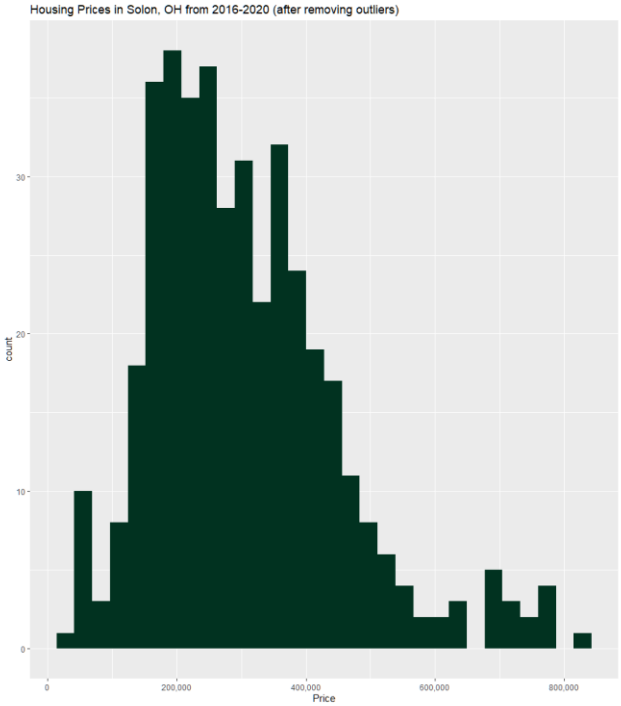
Univariate and multivariate outliers along with high leverage points were sought to potentially improve the model fit. To find univariate outliers and gather Z-scores, we centered and scaled each numeric feature. We looked at the Z-scores for eight features: Price, Rooms (overall), Total Living Room Area, Bedrooms, Garage Size, Garage Capacity, Basement in Square Feet, and Total Bathrooms. Ultimately three records were found to have zscores greater than 4.5 in the Price, Total Living Area, Basement in Square Feet, Garage Size, and/or Garage Capacity variables.

These three observations were removed to improve model performance. (The prices of the three removed items were 850,000; 999,500; and 1,445,000. It is likely these observations were not data errors, but they were very large when compared to the values of other observations.

To find multivariate outliers, we looked to see whether observations were close to the multivariate center of the distribution using the Mahalnobis D for the numeric columns. Z-scores were subsequently gathered for these values. Only one observation had a Z-score above

5.0 and it was previously removed (Price 850,000). The below plot shows a histogram of sale prices before and after outliers were removed. Please note the sales around $1 million have been removed.





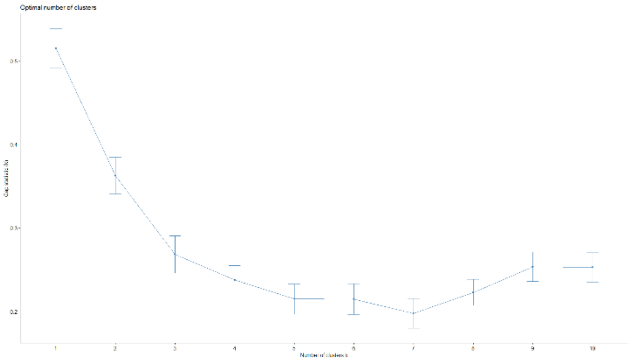
R provides many ways of finding potentially high leverage points. The standardized values of Cook's distance were used. Cook's distance measures the effect of one observation on all fitted values (Pennsylvania State University 2018, Boston University School of Public Health 2016). Observations were flagged when Cook's distance was greater than (4/n). Five high leverage points were observed that had a Cook's distance above the threshold. All five of these observations were homes sold in 2020. Since they each have fairly average attributes, they will remain in the model.

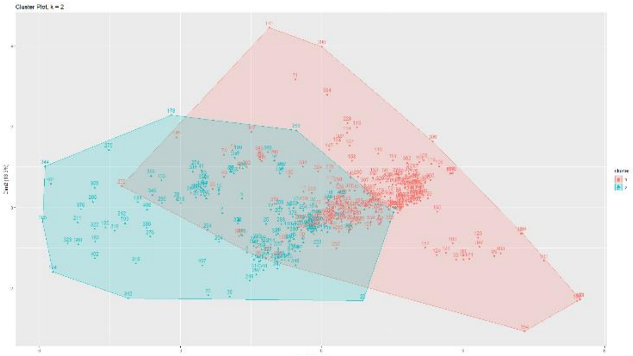
Fitting Models

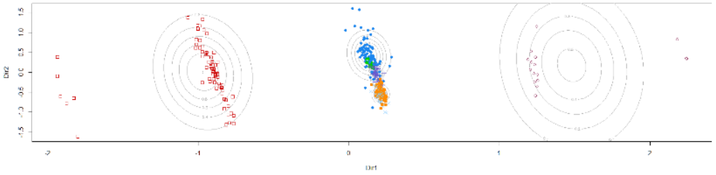
To reduce the twenty-one variables in the data even further, lasso regression was used. Both lasso regression and ridge regression are very similar to multiple linear regression. They rely on the least squares estimates and add a shrinkage penalty λ. The shrinkage penalty reduces the B terms to exactly zero (with lasso) or close to 0 (with ridge). While lasso regression helps with feature selection, ridge regression helps with multicollinearity. As λ increases, our coefficients further shrink towards zero. When λ=0, the least squares estimates are produced. When λ=∞, all B=0. (There is a lot of variation in values for λ, depending on the data set. For instance, values may be in the hundreds or thousands.) Cross-validation was then used to find the right level of λ.

For lasso regression on the 21-variable model, a 75/25 train/test split is used. Cross validation is used on the train set to produce a λ of 1638.745 as this is the lambda value associated with the smallest MSE. The model was then fit and values were predicted. Since lasso regression brings some variables or levels of variables to exactly zero, all variables that had at least one non-zero coefficient for a specific level were retained. The Test R^2 value is 73.62% and the MSE is 5,181,633,471. The initial use of lasso regression reduced the number of features from 21 to 18 which had non-zero coefficients.

Clustering







Using the fviz_nbclust() function in R, we gather within cluster sums of squares to determine the most appropriate number of clusters. Within sum of squares methods show 2-3 clusters should be used while the silhouette method shows two clusters should be used. The gap statistic says only one cluster should be used. A kmeans cluster of size 2 was used, but because of the large amount of overlap it was not used for modeling purposes.

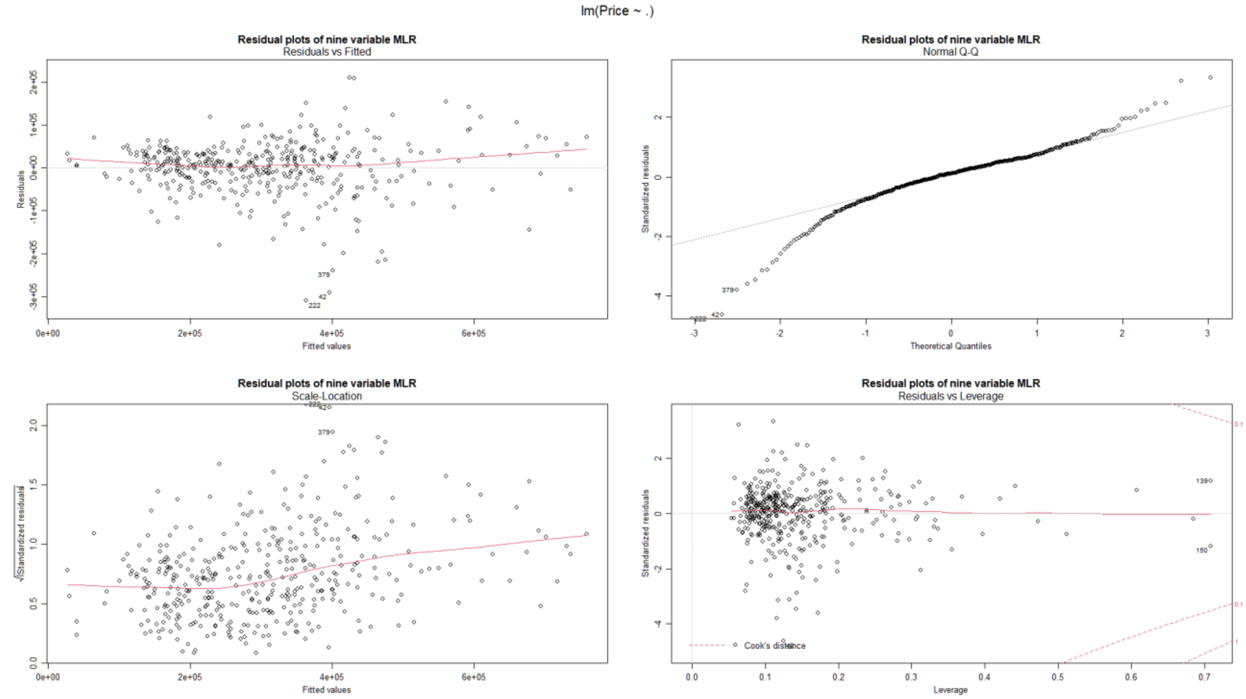
Multivariate clustering methods were also used. The clustering model that had the lowest BIC (Bayesian Information Criterion) and ICL (Integrated Complete-data Likelihood) was the 7 group VEI model, with diagonal distribution, variable volume and an equal shape. The multidimensional plot is shown on the left. The clusters chosen were used in model development, so there are now 19 variables in the models.

A “relaxed” lasso regression model was then fit. Since lasso brings all variables towards zero, we re-ran lasso with only the variables that have non-zero coefficients to see if there’s any effect. The reduced lasso regression removed one predictor from the model. The Test R^2 value is 73.84% and the MSE is 5,127,080,501. There are now eighteen variables in the model.

Ridge Regression on the 18 variable model performs similarly to lasso regression. We used a cross validation technique that produced a λ of 11,833.06 (the λ for a non-standardized model is much greater). We then fit the model and predicted values. The Test R^2 value is 72.86% and the MSE is 5,336,746,763 for the 18 variable Ridge Regression model. Multiple regression model

Multiple linear regression, one of the more traditional and interpretable models, assumes a linear relationship between the predictors and the response. Another assumption is to have constant error variance (homoscedasticity) and a lack of correlation among the error terms. In a previous section, we dealt with univariate and multivariate outliers and high leverage points.

Ten-fold cross validation was used on the train data set. Eighteen predictors were regressed against Home Price and the adjusted R^2 is 80.12% and the MSE is 6,307,926,843.

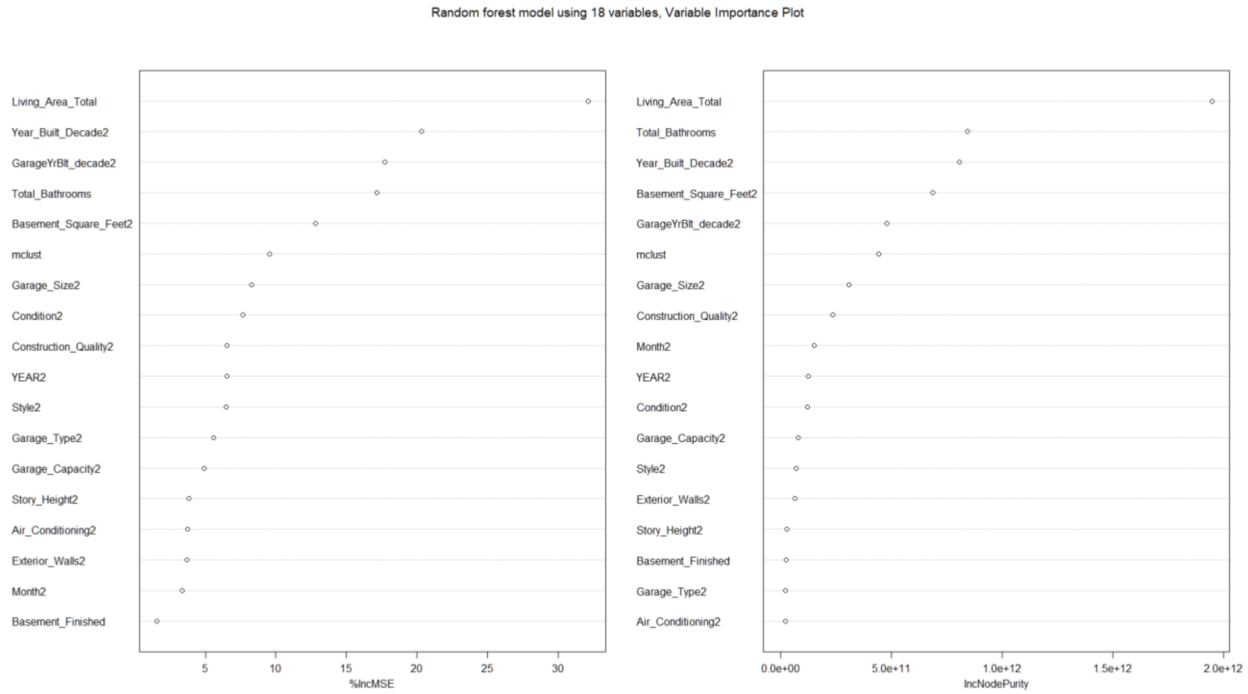


The residual plots above show that residuals are fairly linear. The qq-plot shows the data is fairly normal as it follows the theoretical quantiles. Transformations can be attempted to improve normality. The bottom left plot shows a constant variance among the residuals. The final plot shows leverage points using Cook's distance. There are a few major leverage points that may need to be looked at further.

Random Forest

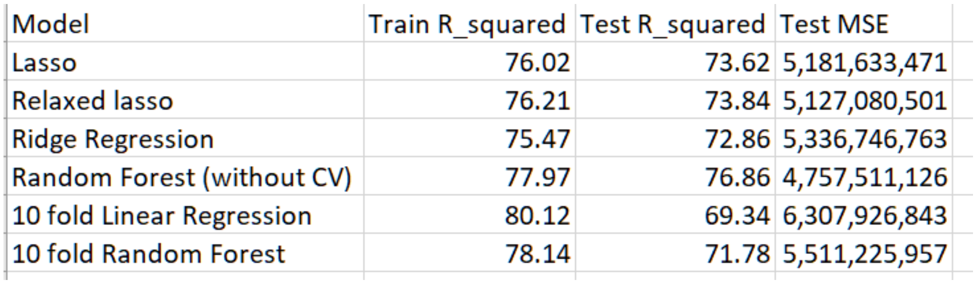
Random forest models are constructed with decision trees which split various features on some value (e.g., observations may be split on Total Living Room Area where values greater than or equal to 1500 are part of one branch and values less than 1500 are part of another branch). The Home Price in the terminal node is then averaged. Random forest works by selecting a random predictor for the next level (usually in the set of √p) and running many trees to determine which features are most important in predicting home prices. This also decorrelates the features. It is worth noting that random forest models do not have an explicit shape like linear models; they are non-parametric. Ten-fold cross validation was used on the train data set for random forest. The Test R^2 value is 76.86% and the MSE is 4,757,511,126 for the 18 variable Random Forest model.

The variable importance plot below shows that Living Area Total has the biggest impact on MSE while the decade the home and garage were built, and the number of bathrooms, and basement area.



Model Validation

The purpose of model validation is to see how good the various models performed on new (test) data. Cross-validation methods were used for multiple linear regression, lasso regression, and ridge regression. The table lists the evaluation criteria when comparing the different models.



Model	Train R_squared	Test R_squared	Test MSE
Lasso	76.02	73.62	5,181,633,471
Relaxed lasso	76.21	73.84	5,127,080,501
Ridge Regression	75.47	72.86	5,336,746,763
Random Forest (without CV)	77.97	76.86	4,757,511,126
10 fold Linear Regression	80.12	69.34	6,307,926,843
10 fold Random Forest	78.14	71.78	5,511,225,957

Evaluating the models

All models performed fairly evenly. The Random Forest (without Cross Validation) performed notably better than others in its test R^2 and test MSE values. The relaxed lasso model did pretty well too. For models with more rigorous cross validation, either 10 fold random forest or linear regression would work well.

Conclusion

This project shed light on which models are more accurate and which features play a large role in determining the sale price of a home. The most important features were related to total square footage, the decade the home or garage was built, the number of bathrooms, and the basement area. Further analysis can further reduce the feature space. It would also be worth talking to people who work in real estate to better understand the market and what features are seen as important.

References

Boston University School of Public Health. (2016). Regression Diagnostics. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression7.html

Hermann, A. Joint Center for Housing Studies of Harvard University (2020). *Homebuying Innovations have enabled the strong rebound in home sales.* https://www.jchs.harvard.edu/blog/homebuying-innovations-have-enabled-strong-rebound-home-sales

James, G., Witten, Daniela, W., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning: with applications in R. New York: Springer http://faculty.marshall.usc.edu/gareth-james/ISL/

Pennsylvania State University, The. (2018). Identifying Influential Data Points.

https://online.stat.psu.edu/stat462/node/173/

Rodríguez, G. (2020). Regression Diagnostics

https://data.princeton.edu/wws509/r/c2s9

U.S. Census Bureau Quickfacts: Solon City, Ohio (2019). American Community Survey, 2019 (Five year estimates). https://www.census.gov/quickfacts/soloncityohio

U.S. News. Top Ranked Ohio Schools (2020). https://www.usnews.com/education/best-high-schools/ohio