STOR 565 Group 3 Report

The Creation of a Post-Crisis Housing Price Prediction Model

2024-12-03

I. Project Overview

In this project we attempted to predict future housing prices during and after time periods when the housing market has experienced a state of crisis. Our training data was recorded during the housing market crash of 2008 and our test data was recorded during the recent Covid-19 pandemic that caused volatility in the housing market. To achieve this goal we employed linear regression, forwards and backwards stepwise regression, Ridge and LASSO regression, PCR, PLS, and a variety of tree based methods. Throughout this report you will see a large variety of visuals to help demonstrate our findings. There will also be brief discussions on any numerical adjustments that needed to be made for the sake of comprehension and accuracy.

AI Usage Disclaimer

We **did not use** any AI features to construct any aspect of our project.

II. Data Cleaning

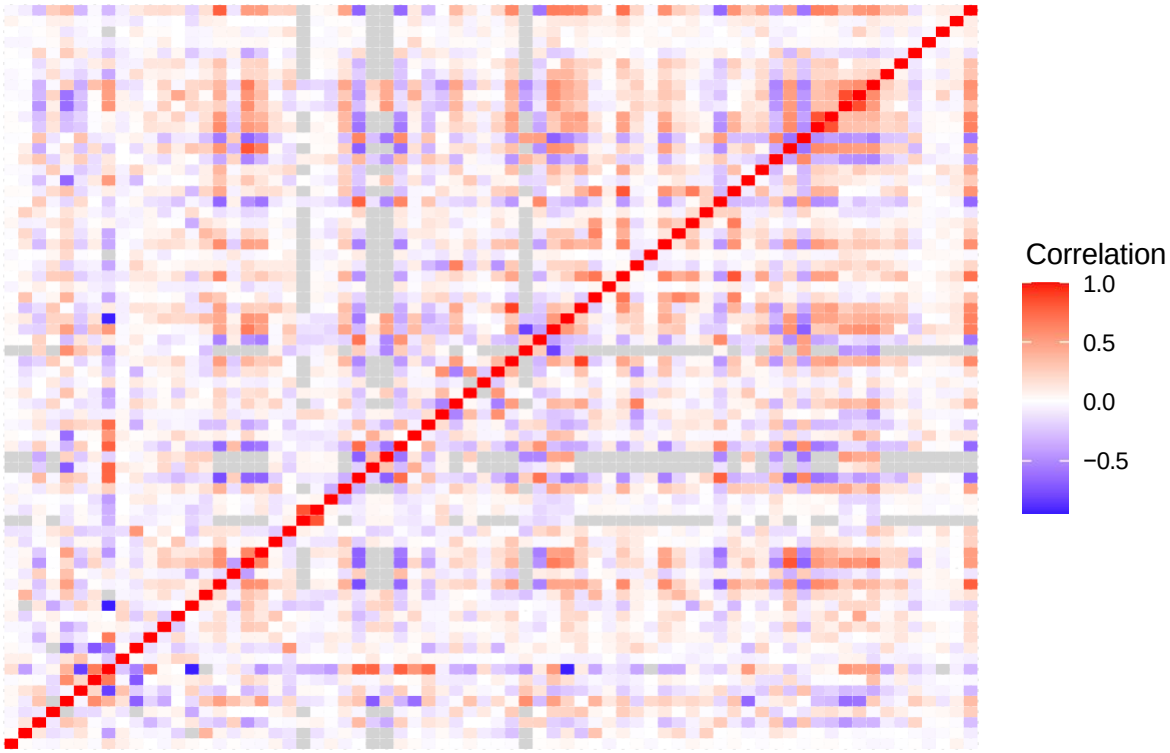
Looking into our original data set you will see there are a few major problems that needed to be dealt with in order to make our data set usable for this predictive analysis problem. The biggest of these issues was the lack of data collected for a specific few variables. For these variables we had 2 different options: either attempt to impugn the data or to get rid of the variables entirely. 19 of our 81 variables had missing values, so we looked at the total number of NA values for these variables and chose a cutoff point. Due to the severe lack in collected data for some of these variables we ended up having to get rid of the entire variable all together as impugning the data would cause more created data than recorded data. The 6 removed variables had between 259 and 1453 NA values. For the remaining variables which contained NA values, a vast majority of these values were from the same housing entries, so we filtered out these observations from the data set to allow for more accurate prediction probability. Lastly we split our data into training and testing data sets with a split of 20 percent test and 80 percent training.

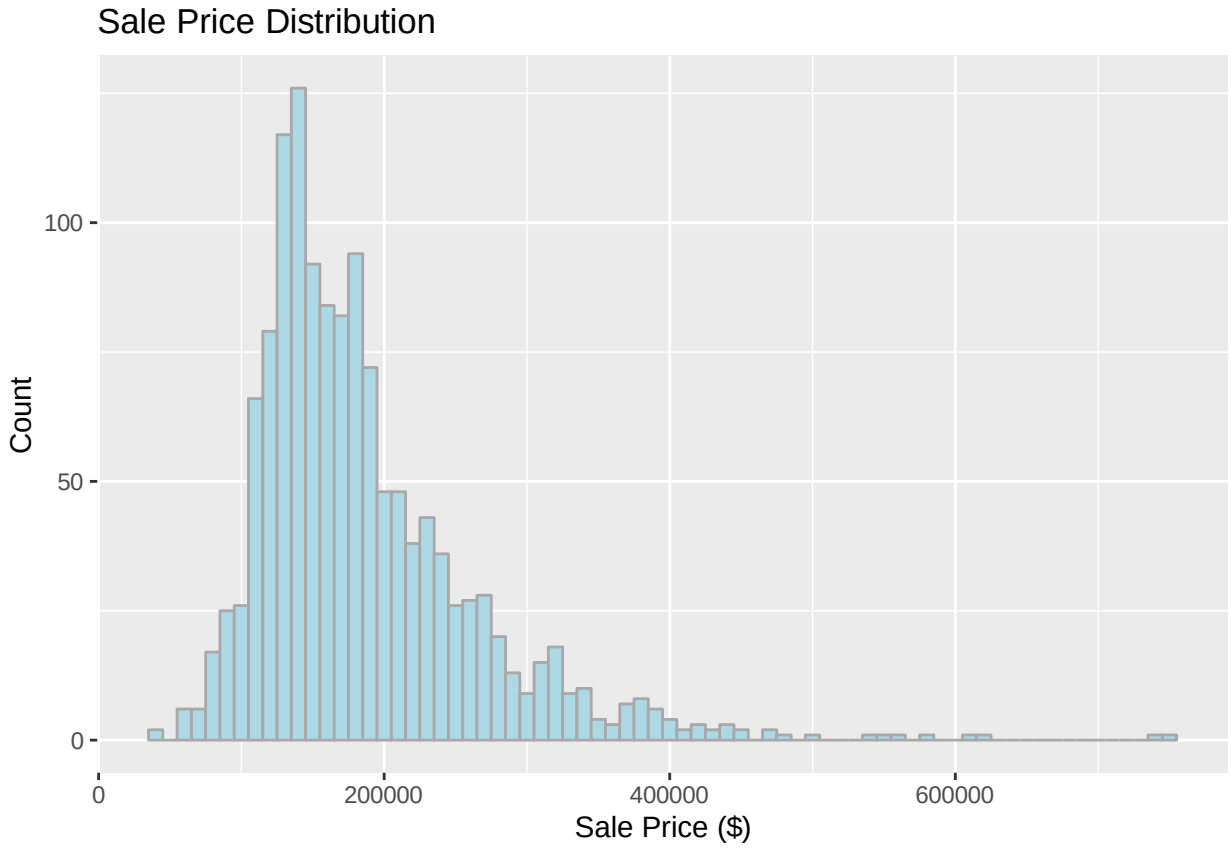
III. Exploratory Data Analysis

Before starting our regression modeling it was important to comb over the data with a few exploratory methods to point out any potential problems for making our predictions. The first of these methods we used was to create a correlation heat map. This analysis was performed to depict the relative correlations amongst our separate variables. Its importance was to make sure we understood which variables would typically group together or present similar inputs when it comes to our modeling. In other words it tells us which variables we need to keep an eye on when it comes time for our modeling. The second of these exploratory analysis techniques that we used was creating a histogram of the sale prices of the homes. We did this to better understand the skew of our data in terms of the sale pricing so that we knew where the modeling would most likely perform better and worse during testing.

Below you can see the correlation heat map along with the histogram of the housing prices after cleaning the data. You can observe that there are some potential outliers in the histogram, we decided to leave them in our analysis to start and if need be we'd remove them after some calculations.

Correlation Plot





IV. Inflation Rate and RMSE

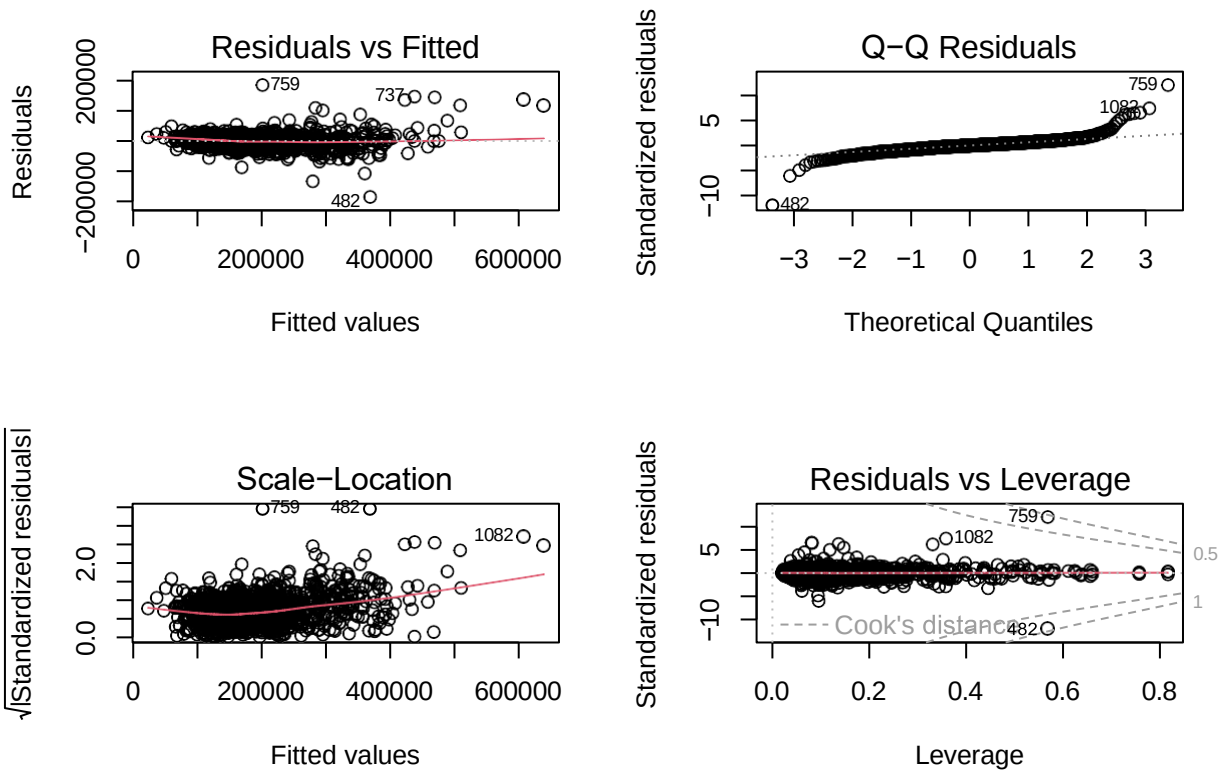
Since our data includes prices over a long range of years, it is important that we consider the value of money for each year. We decided to use a CPI inflation rate found from the U.S. Bureau of Labor Statistics. The rate was approximately 45%. This inflation rate represents the loss in value of the dollar from the year of 2010 to 2024. Using this inflation rate was critical in accurate predictions from our models. Secondly since regular Mean Square Error (MSE) created error values that were extremely large, causing confusion and reducing interpretability, we decided it best to use the Root Mean Square Error (RMSE).

$$RMSE = \sqrt{\frac{\Sigma(y_i \breve{} \hat{y}_i)\check{\mathrm{s}}}{n}}$$

V. Modeling Techniques

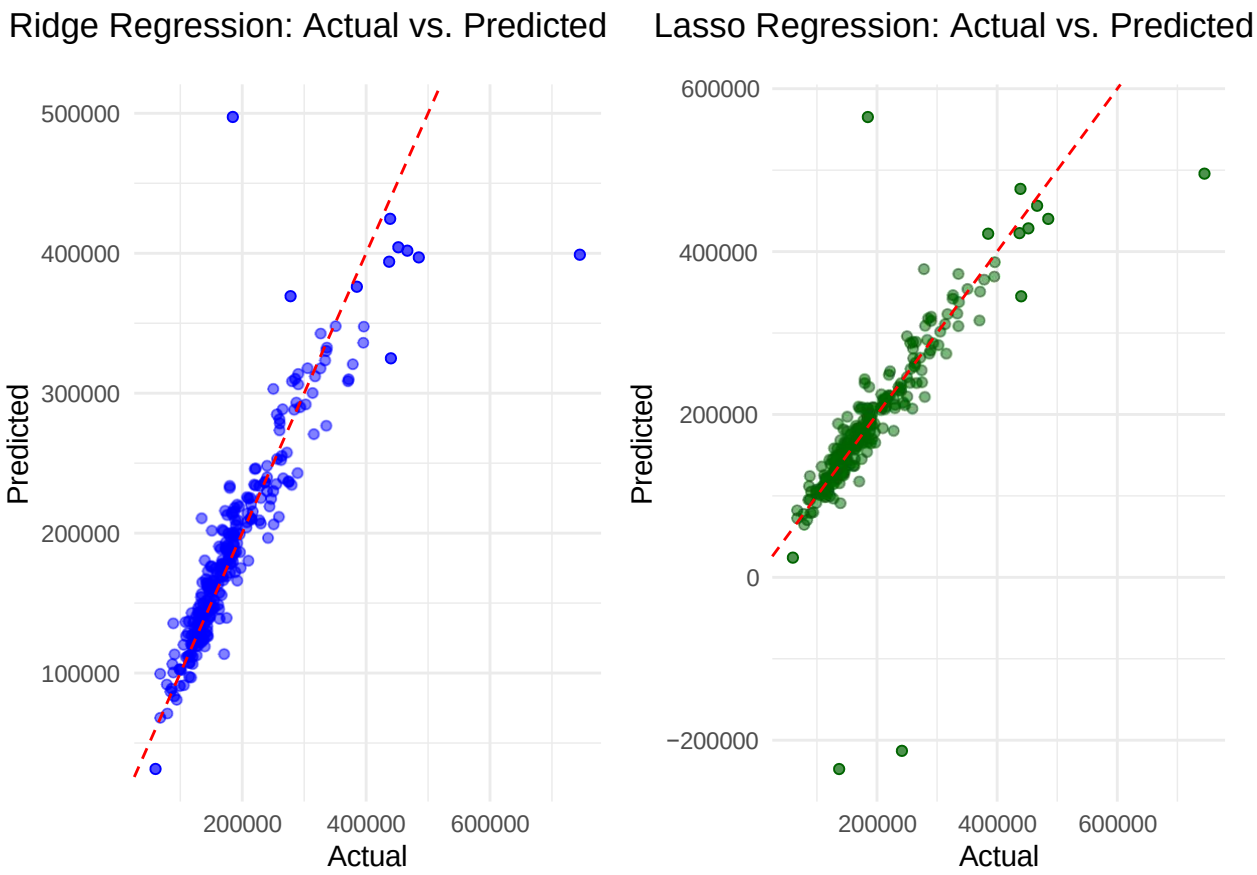
We first attempted to solve this problem with a linear regression model. We knew that our data was most likely not going to be most accurately predicted using this technique, and our main goal was to gain more insight into the data. Using a QQ plot we were able to determine that the data did relatively follow a Gaussian Distribution, which would be important for modeling choice. This model did perform the worst when it came to RMSE, with a value of 114,448.6. Even though it was the worst it still performed better than expected, suggesting that the data may have an underlying linear relationship. Below you can see the diagnostic plots for the linear model.

```
## Warning: not plotting observations with leverage one:
##   111, 167, 247, 298, 363, 375, 468, 534, 546, 608, 870, 919, 1087, 1127, 1142, 1163, 1168, 1189, 12
```

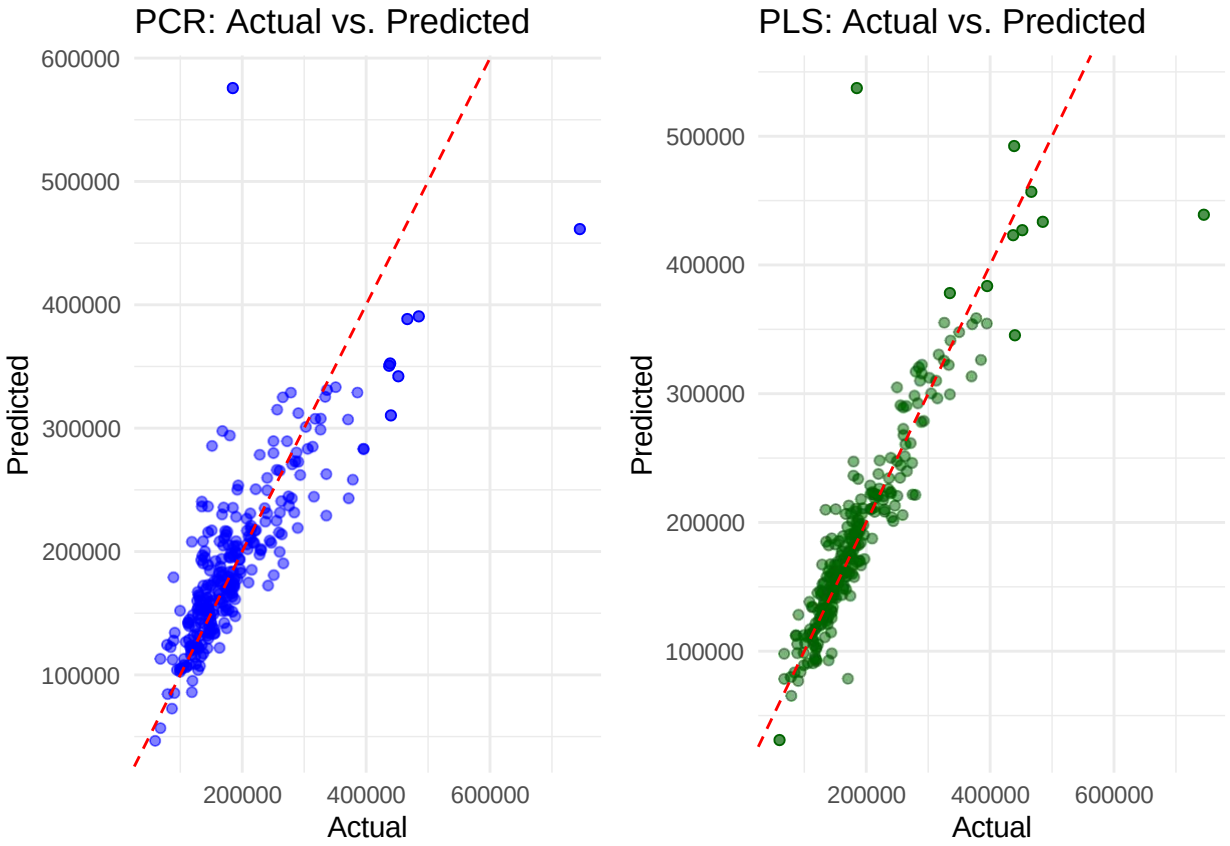


For our next attempts we used forward and backwards stepwise regression. Once again computing these models we were not expecting them to be our best performing models but we did gain insight using them. These models had similar RMSE values with forward regression having a RMSE of 22,058.8 and backwards having an RMSE of 22,800.8. These models ended up performing better than everything except Random Forest, and helped us identify variables that could have strong predictive qualities.

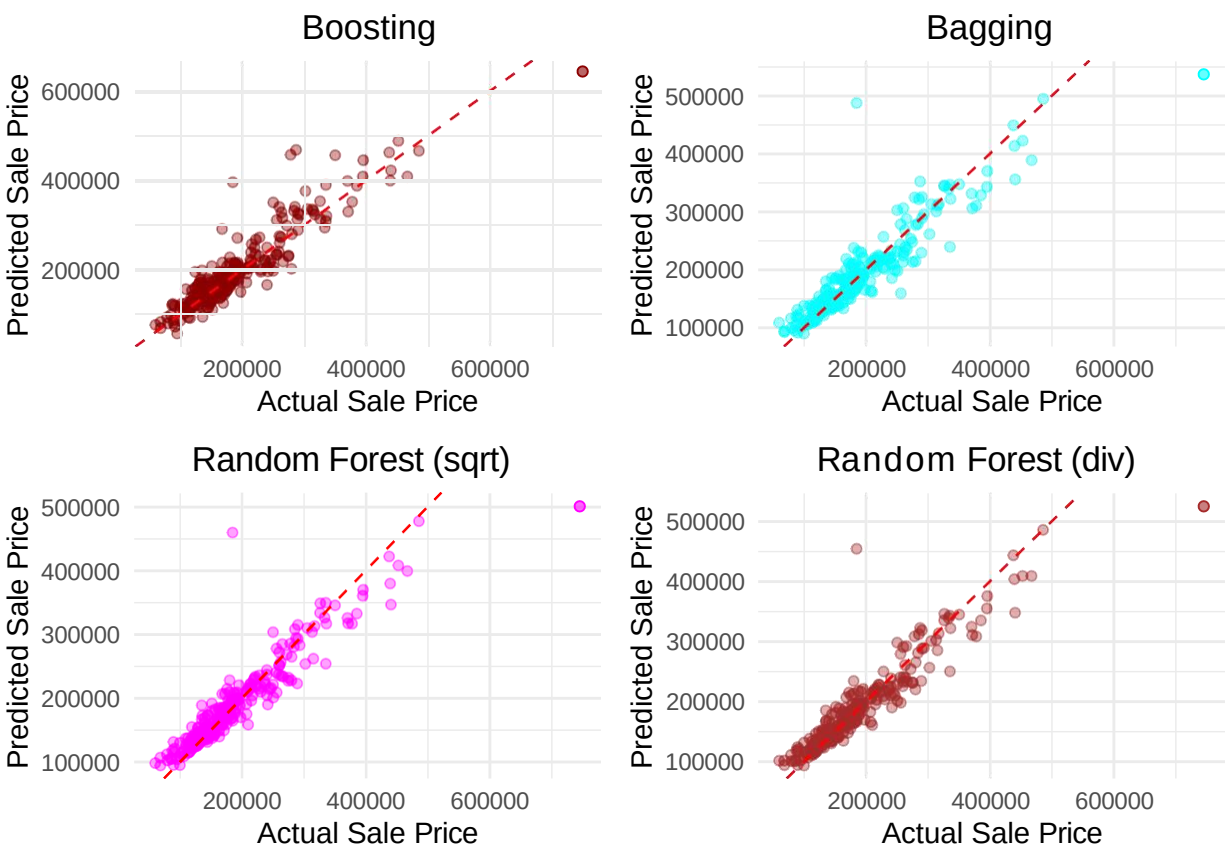
The next two models we created were Ridge and Lasso models. We decided to create these models due to the high complexity of our data (ie. the large amount of predictor variables). After the creation we tested these predictions against our test data and found RMSE values of 50,127.9 and 37,192.7 respectively. While these values were much lower than the value of the linear model, we were surprised to see that they did not perform nearly as well as the previous stepwise models attempted. You can see a visual comparison of these models' performance below.



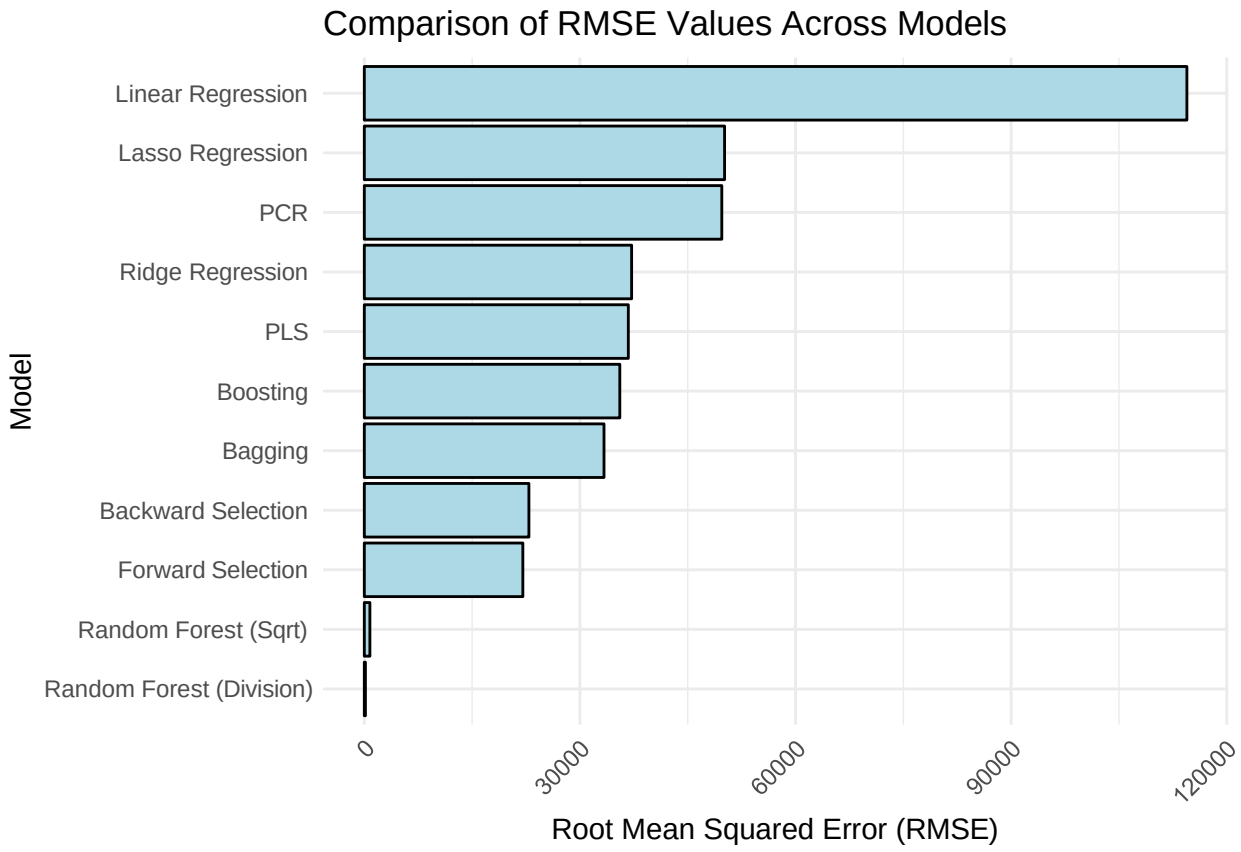
Our next attempts used the dimension reduction techniques of PCR and PLS. Observing the output of these two models we see PCR obtained a RMSE value of 49,742.5 and PLS obtained a RMSE of 36,732.5. You can see a visual comparison of these models' performance below.



Lastly we attempted a few different methods of tree based modeling. The first of these methods we used was Bagging. Using this had a RMSE value of 33,348.9. Next was Boosting, which produced a RMSE of 35,548.6. Lastly, and most importantly, we tried Random Forest models using both the square root and a divisor of three for the features. Through these we were able to generate extremely small test RMSE values of 776.8 and 152.8 respectively. These two methods were by far the best at predicting the test values. Below you can find a comparison of the performance of all four models.



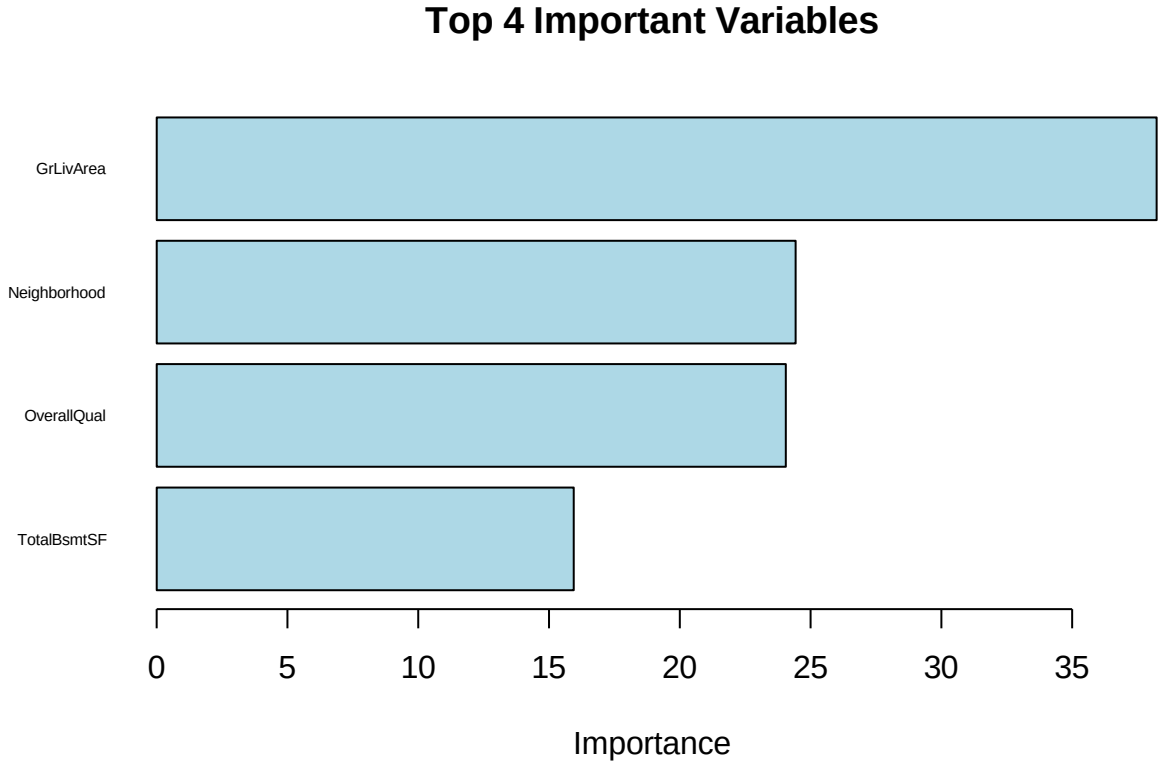
In summary, you can see a plot comparing the RMSE value for each model below.

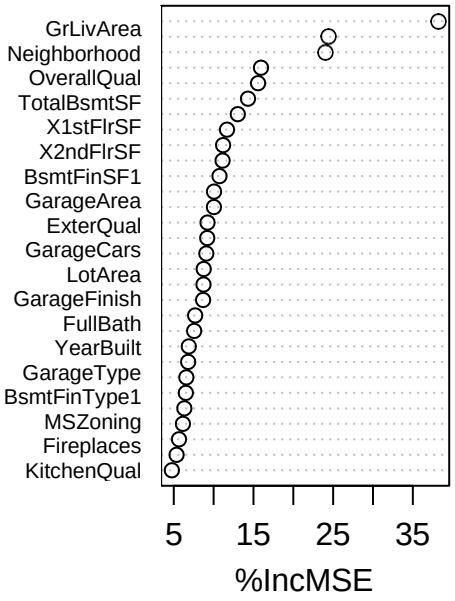


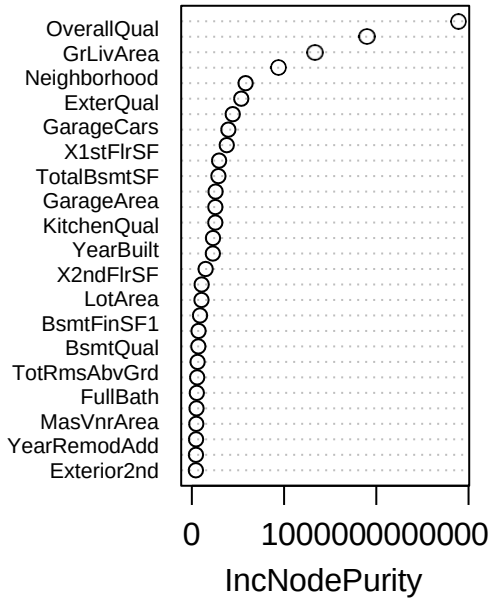
VI. Influential Features

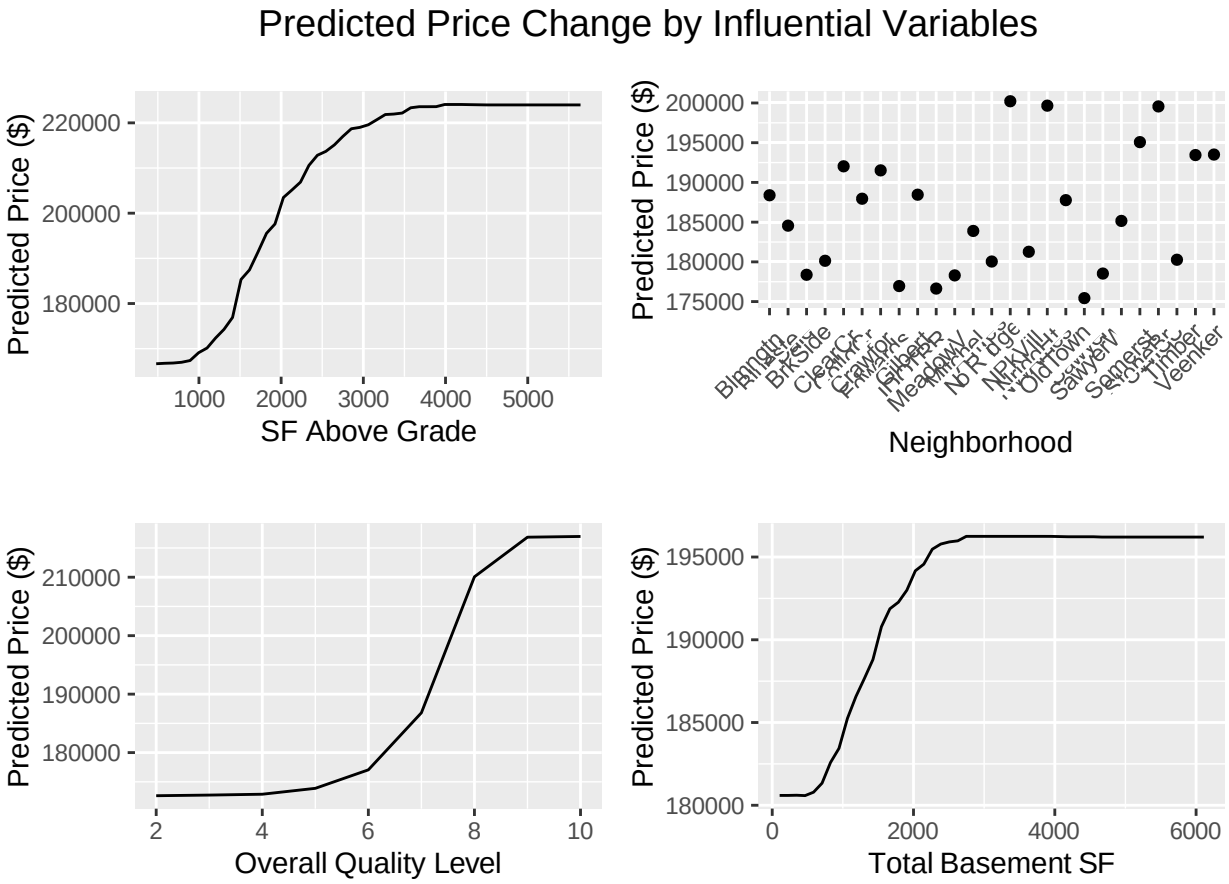
Seeing how the Random Forest tree based methods were far superior in terms of predictive power, we wanted to take a deeper dive into which features the Random Forest models found to be the most influential. Using graphical techniques we are able to see that there are four overarching features that provided the most influence to the models. These four being: Non-Basement Square Footage, Neighborhood, Overall Quality, and Total Basement Square Footage. These features are listed in order of importance with a brief explanation of each, followed by a ranked importance plot along with a plot of how each variable affects the sale price of a home.

- Non-Basement Square Footage (GrLivArea)- Square footage of the house from the ground level floor upwards.
- Neighborhood - The name dictated to the area that the house is located in.
- Overall Quality (OverallQual) - A numeric rating from 1-10 indicating the quality of the culmination of features in and outside of the house.
- Total Basement Square Footage (TotalBsmtSF) - Square footage of the levels of the house below the ground level.


Top 4 Important Variables
GrLivArea
Neighborhood
OverallQual
TotalBsmtSF
0
5
10
15
20
25
30
35
Importance

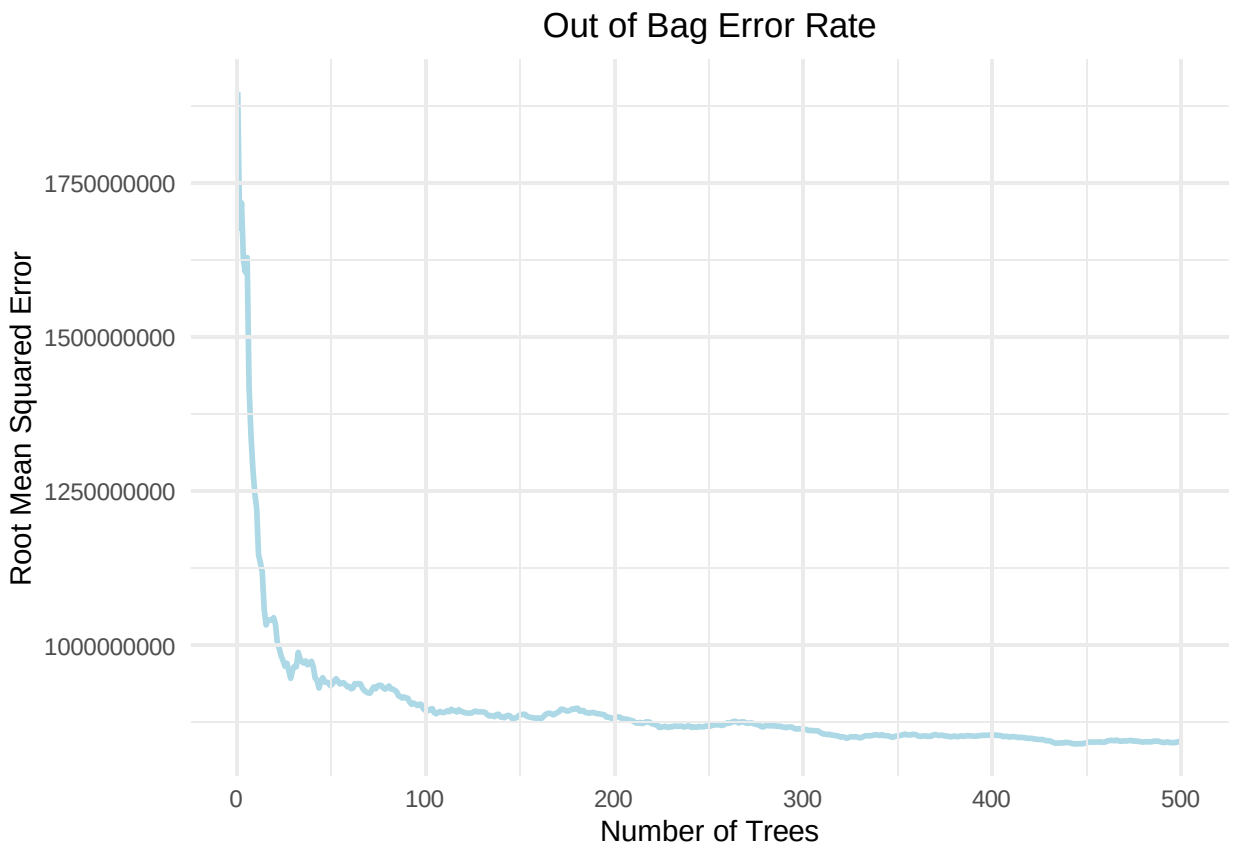

GrLivArea
Neighborhood
OverallQual
TotalBsmtSF
X1stFlrSF
X2ndFlrSF
BsmtFinSF1
GarageArea
ExterQual
GarageCars
LotArea
GarageFinish
FullBath
YearBuilt
GarageType
BsmtFinType1
MSZoning
Fireplaces
KitchenQual
5
15
25
35
%IncMSE


OverallQual
GrLivArea
Neighborhood
ExterQual
GarageCars
X1stFlrSF
TotalBsmtSF
GarageArea
KitchenQual
YearBuilt
X2ndFlrSF
LotArea
BsmtFinSF1
BsmtQual
TotRmsAbvGrd
FullBath
MasVnrArea
YearRemodAdd
Exterior2nd
0
1000000000000
IncNodePurity



VII. Model Diagnostics

There is extreme value and importance in looking into your model and seeing how it works and what the error rates look like. As stated earlier our best tree based model has a RMSE of 152.8. Located below is a graph indicating the out of bag rate. Observing this graph we can see the y-axis contains numbers that are extremely small, which is where we want the out of bag error rate to be.



VIII. Current Housing Value Testing

The first house we predicted for was located near the Northwest Ames neighborhood and had 672 square feet above ground, 300 square feet of basement, and an overall quality rating of 6. The price for this house listed on Zillow was $175,000, and the price our model predicted was $173,693.80.

The second house we predicted for was located near the Brookside neighborhood and had 1,774 square feet above ground, 450 square feet of basement, and an overall quality rating of 9. The price for this house listed on Zillow was $286,000, and the price our model predicted was $285,145.40.

The third house we predicted for was located near the Northridge neighborhood and had 2,625 square feet above ground, 1,823 square feet of basement, and an overall quality rating of 10. The price for this house listed on Zillow was $1,350,000, and the price our model predicted was $688,279.2.

Our predictions for the first two houses were less than $1,000 and $2,000 away from the actual price respectively. It makes sense that our model failed to predict the third price since the most expensive house in the training data was priced at $755,000. All of our predictions were less than the actual value of the house price, even if it was a minimal difference.

IX. Shortcomings and Limitations

Going into the limitations of our best model there are a few things that need to be spoken about. The first of these being the use of an incomplete data set. To work with the data set, as discussed previously, we had to make some alterations to the data set that could have had an impact on the final results of the model. Secondly, our model uses data from a town in Iowa and is skewed towards cheaper houses. A problem arises here because it is not guaranteed that data from this location is transferable to prediction in other areas. In fact one of our most influential features shows how deterministic location is in prediction. The skew towards cheaper housing prices also makes the model less likely to make accurate predictions for higher value homes.

Other limitations include houses not located in neighborhoods and subjective factor variables. We found that during the process of testing the model on current listings, the neighborhoods of the listings weren't included in our levels in the training data. One of the homes also was in a more agricultural area which was not listed as a neighborhood option. To elaborate on the subjectivity of factor variables we found that many variables, such as quality or condition, are not rated on a scale in a home listing, it is up to the viewer to determine for themselves. This posed the possibility of the prediction being wrong due to bias from the individual conducting the prediction.

X. Future Projects

While our model only has strong prediction power for the town of Ames there are plenty of future options that could be implemented to help create a more generalized model. A few ideas for this would be to include data from urban and rural areas, to use multiple locations, and to add other housing variables. Some other project ideas would be to pull data from other crisis event times to use for prediction and testing or predicting the future values of other living accommodations, such as apartments, condos, etc.

XI. Conclusion

The goal of our project was to make accurate predictions of future housing prices based on the housing markets during/after crisis events, such as Covid-19 and the housing market crash of 2008. During our exploratory analysis we came across a few issues with the data, these including incomplete data and skewness. While we did control for incomplete data, we felt it best to not remove more observations than absolutely necessary. When testing our models we found a Random Forest with a divisor of three created the most accurate model. Our conclusion on why this is the case was because housing prices usually culminated around whole values in thousands. For example, $150,000 opposed to $134,922. We also had lots of factor variables that could be better handled by a tree based method. We ran diagnostics, found our influential variables, and came up with our models' limits and ideas for future endeavors. Finally, in the end we created a model and found it to be accurate in two out of three test cases by a $2,000 margin.

References

Cityofames. (n.d.). https://www.cityofames.org/home/showpublisheddocument/52755/637050278030900000

U.S. Bureau of Labor Statistics. (n.d.). CPI inflation calculator. U.S. Bureau of Labor Statistics. https://www.bls.gov/data/inflation_calculator.htm

Zillow, Inc. (n.d.-a). 1217 Scott Ave, Ames, IA 50014. https://www.zillow.com/homedetails/1217-Scott-Ave-Ames-IA-50014/93960195_zpid/

Zillow, Inc. (n.d.-b). 1507 Grand Ave, Ames, IA 50010: MLS #65409. https://www.zillow.com/homedetails/1507-Grand-Ave-Ames-IA-50010/93955629_zpid/

Zillow, Inc. (n.d.-c). 5373 cobblestone CT, Ames, IA 50014: MLS #65161. https://www.zillow.com/homedetails/5373-Cobblestone-Ct-Ames-IA-50014/375239427_zpid/