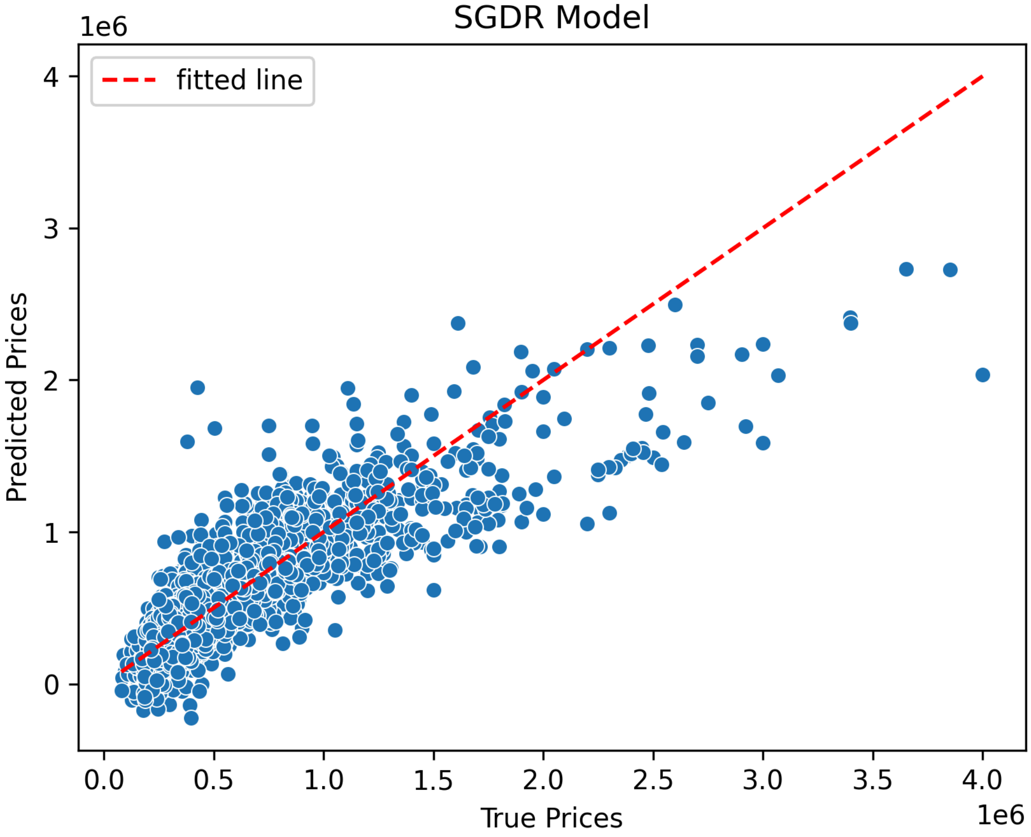
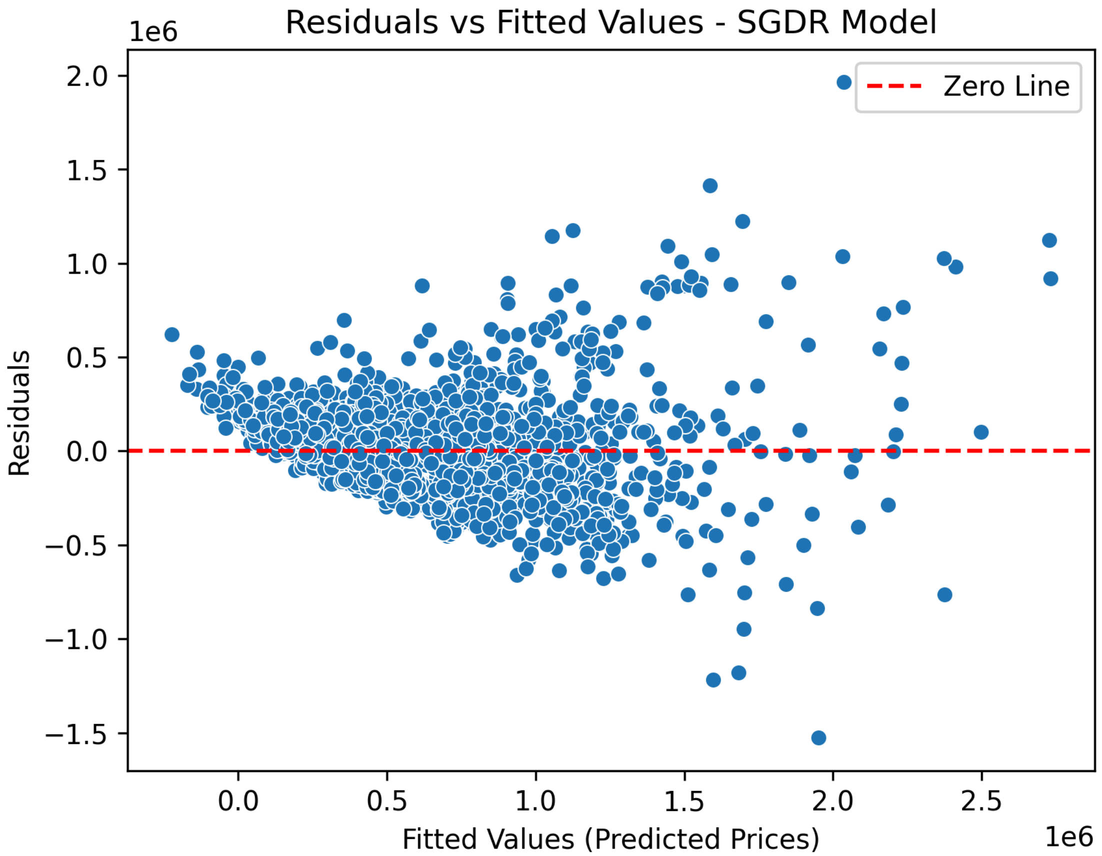
Chris Troop
2024-11-27
CSC 5610

Project M4

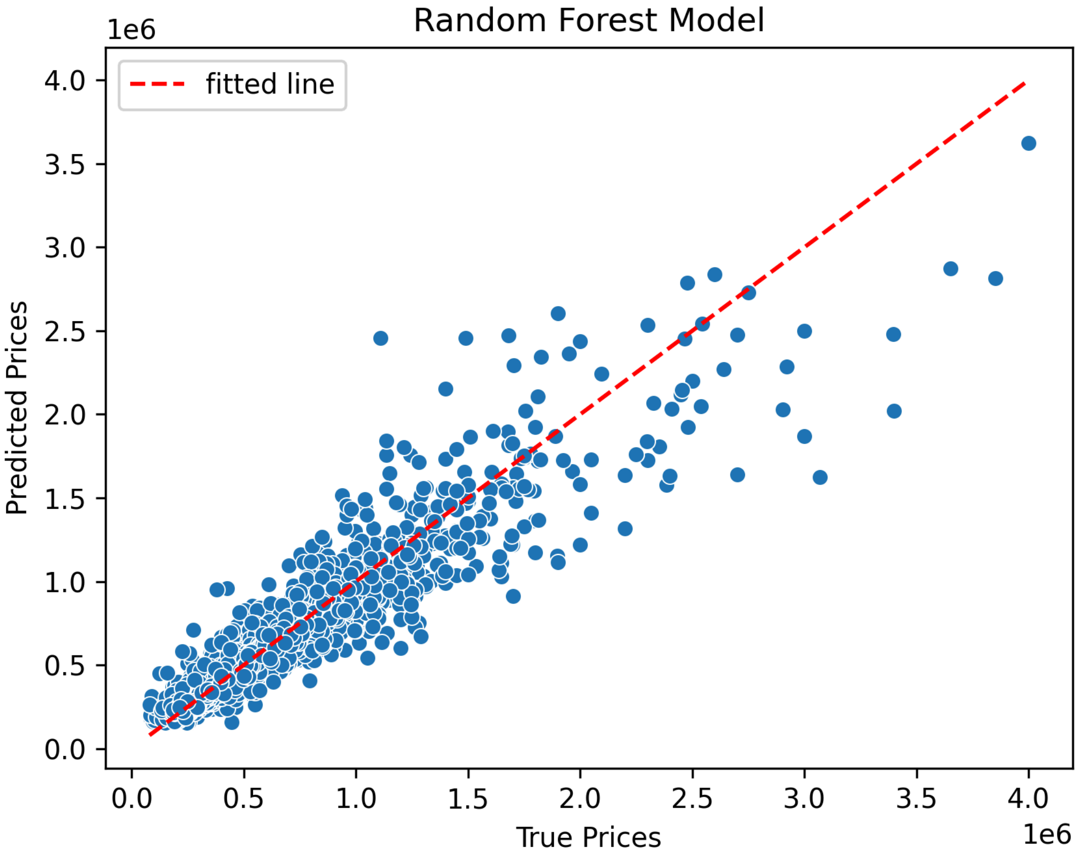
The data set provided was a list of sold houses in King County, Washington from May 2014 to May 2015. The initial data set required cleaning, I dropped any records with 0 bedrooms, 0.0 bathrooms, and there was a home with 33 bedrooms that was dropped. For the building of the predictive model of sale price, I only kept the most recent sale of the property. This trimmed the data set from 21,613 down to 21,419.

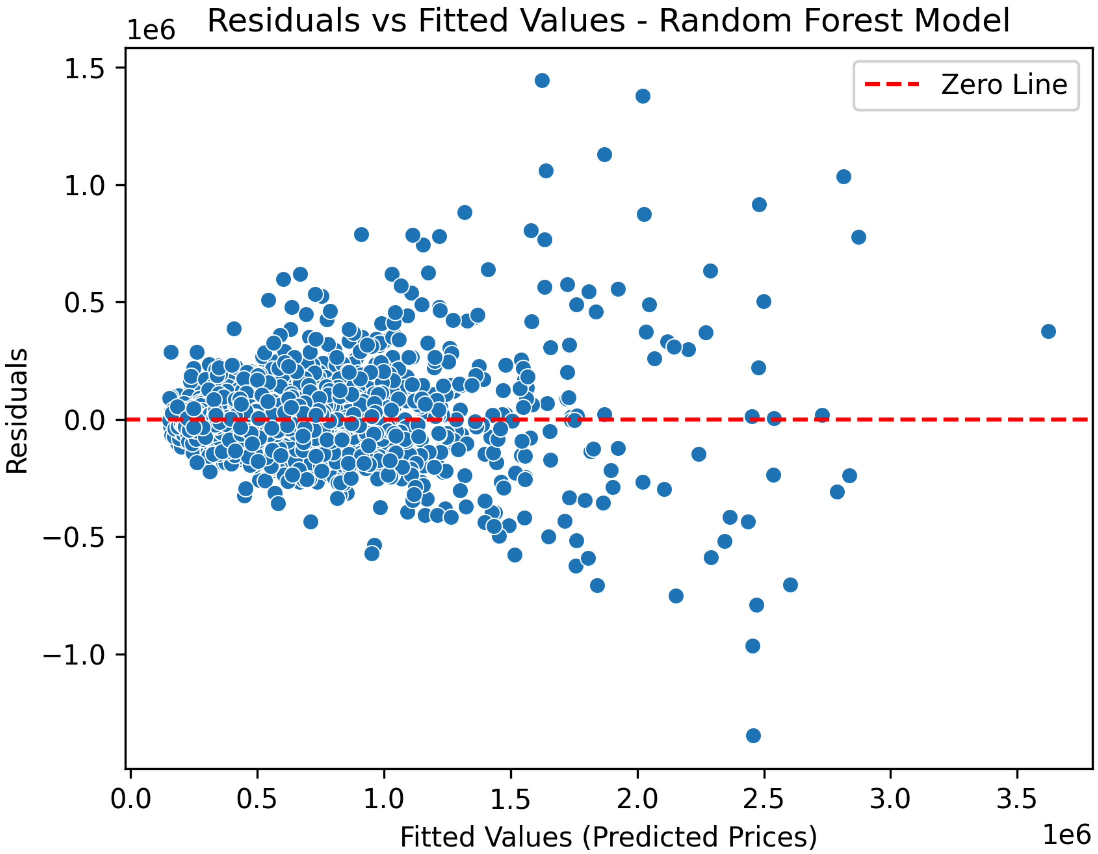
I built linear regression, ridge, SGDR, and random forest models to see which performed best. Linear regression, ridge, and SGDR all performed similarly with MAPEs all over 25% and RMSE over $188,000. Below is the SGDR model plots of predicted price vs real price and the residuals vs the predicted values as an example of what these three models produced.





These provided low confidence in predictive capability compared to the random forest which had a MAPE of 12.89% and a RMSE of $122,907.

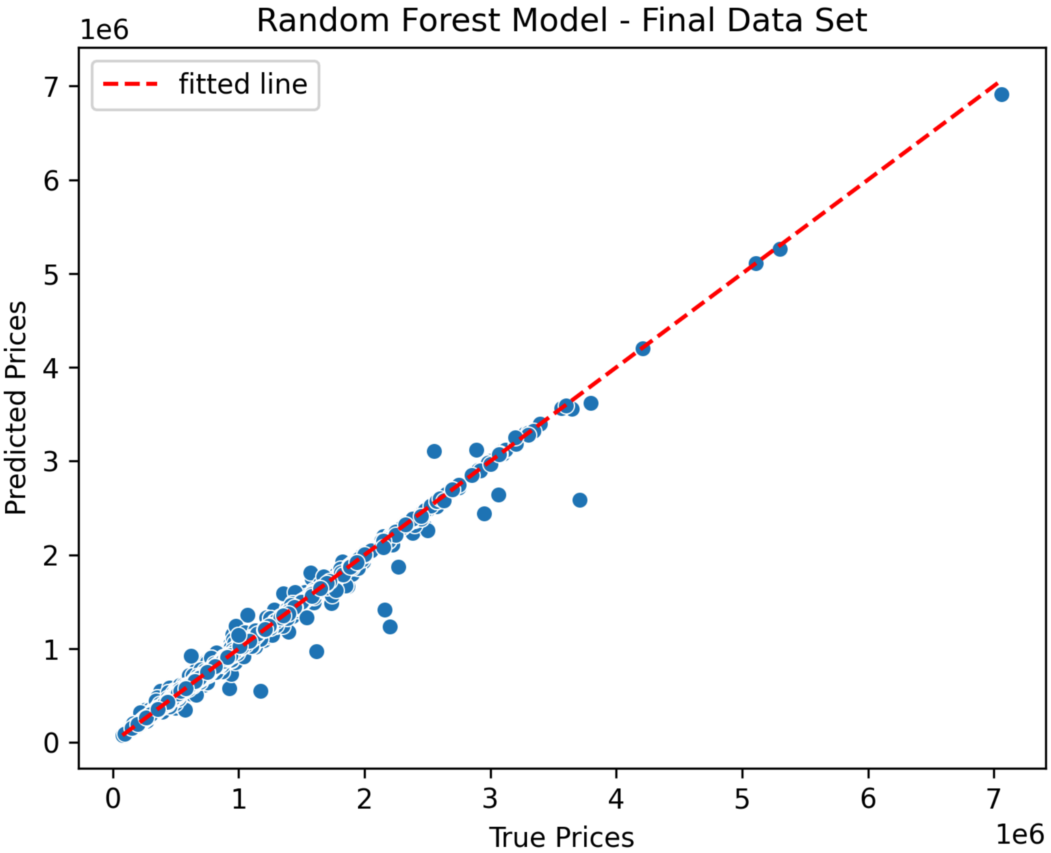

Random Forest Model
1e6
fitted line
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Predicted Prices
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
True Prices
1e6

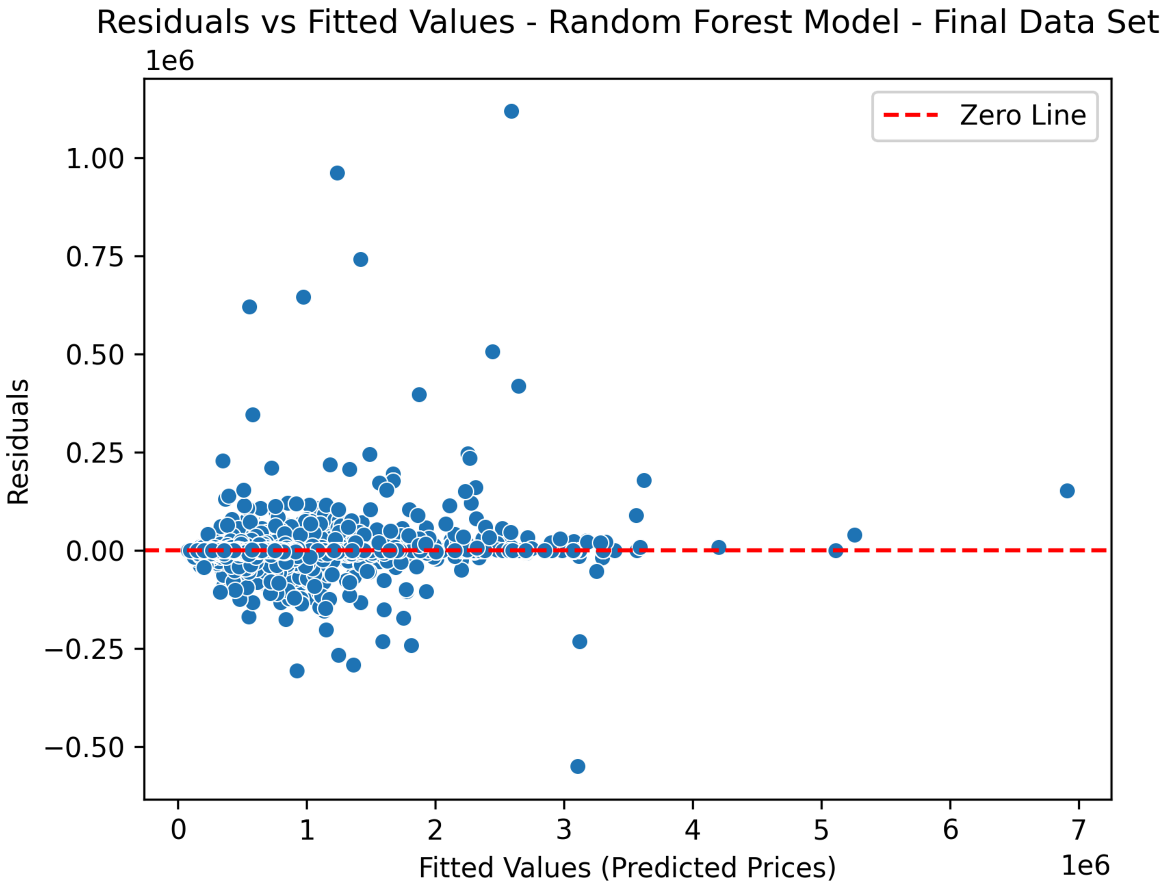


The RMSE is still very high but I will address this later through additional data and feature engineering

To improve the quality of the model I incorporated the population and schools data sets. Areas that are densely populated typically have higher property values due to their desirability and as every realtor will tell you there are three important L's in real estate; Location, Location, Location. Population size typically drives demand for services and businesses. The larger the population in the zip code the more desirable the area; is the thought process I took when selecting this. The other data set I added was schools, education is one of the more sought after services parents look for in a new house location. I engineered two variables to help quantify the schools in the zip codes that were being investigated. Number of schools in the zip code would quantify if the area is most residential and family friendly. The other variable was students per teacher, this was created to provide some sort of qualitative aspect of the school. Schools that can keep the student to teacher ratio low will give the students better experience and higher quality of service. This variable encases a few assumptions as well, less students per teacher means the school would be more expensive leading to higher home prices nearby.

After adding these new engineered variables and the population in the zip code. A new random forest model was generated and the results improved greatly. The MAPE was 0.35% and the RMSE was $16,601. The plots below provide and visual indication of the improvement.


Random Forest Model - Final Data Set
1e6
7
6
5
4
3
2
1
0
Predicted Prices
fitted line
0
1
2
3
4
5
6
7
1e6
True Prices



The provided data set did not have enough information about what might be driving home prices in each zip code. Adding population and school information to the data set provided more context and enhanced the predictive value. I would consider using the model to help with making an offer on a home but I would also need to add some more information into the model such as homes listed currently that are still for sale, how many days they have been listed, garage size, crime rate in the area. I think the predictive value is there with this model but I would still need to use some human judgment to decide if the value was appropriate.