Unit 28.2 - Capstone Two – Final Project Report
Gayla Rios

1. Problem Identification

With rising interest rates, how should buyers, sellers and agents evaluate the residential real estate market?

As a residential real estate agent in Sacramento, California, I know first hand what an anomoly our local market has been since the pandemic. Coming off a wildly competitive residential real estate market in 2021 where most properties were selling well above list price with numerous multiple offers, the housing market abruptly cooled in 2022. In fact by the middle of 2023, total sales volume decreased by 40%. During this same time period, the Federal Reserve steadily increased interest rates as a monetary tool to decrease inflation. For this project, we're going to build a model that incorporates interest rates along with other property variables such as square footage and number of bedrooms as predictors of home prices to see if changes in interest rates can be used to predict home prices.

2. Data Sourcing

As a member of the Sacramento Association of Realtors, I have access to Sacramento Metrolist data and was able to create my own dataset. There's a data limit within Metrolist. So the data was pulled as 22 separate .csv files – one for each year – and then combined into one dataframe. Two additional datasets were used – one with 20 years of PRIME interest rates and the other with Consumer Price Indexes from year 2000 until 2022.

Datasets and Corresponding Variables

1. Sacramento Metrolist – APN (Assesor Parcel Number), address, bedrooms, bathrooms, square footage, lot size (acres), year built, property condition, remodeled/updated, DOM (Days on Market), CDOM (Continuous Days on Market), original price, list price, close price (sold price), on market date, close date, multiple offers, number of offers
2. Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/series/DPRIME)
   Daily prime interest rate, Date
3. US Bureau of Labor Statistics
   ('US_Bureau_of_Labor_Statistics_CPI_for_All_Urban_Consumers_CPI_U_2000_to_2023.csv')
   Year, Month, Value

3. Data Cleaning - Metrolist

1. 'Property Condition', 'Remodeled/Updated', 'On Market Date', and 'Multiple Offers'
   Had hoped to include them in the model, but unfortunately they were newer fields and had far too many missing values to be included. Those columns were dropped from the dataframe. These features could be inlcuded in a possible future study with a shorter timespan.
2. 'APN', 'Lot Size', 'Square Footage' and 'Year Built'
   Dropped rows with missing values in columns
3. 'Close Date'
   Converted to datetime object
4. 'Bedrooms'
   Before converting to an integer type object, removed all characters after the hyphen leaving only the lower range of number of bedrooms. Then cleaned up outliers by limiting rows to only those with between 2 and 5 bedrooms.
5. 'Bathrooms'
   Dropped characters inside "()" which denotes full and half baths leaving only the total bathroom count. Assigned integer type and handled outliers by limiting total bathrooms to between 1 and 5.
6. 'Original Price', 'List Price', 'Close Price'
   Deleted extra random zero after decimal

7. ‘Address’, ‘Zip Code’
   Converted to strings
8. ‘Lot Size’
   Removed outliers by limiting to interquartile range (IQR .17 - .27) I have personal industry knowledge of the region and feel like we could safely have increased the upper limit to .50 acre, but decided to stick with the IQR and leave this as a possible variable to adjust in future studies.
9. ‘Original Price’
   After sorting dataframe, three obvious typos where there were several more zeros than corresponding list and sold prices. Removed extra zeros.

4. Data Cleaning - DPRIME

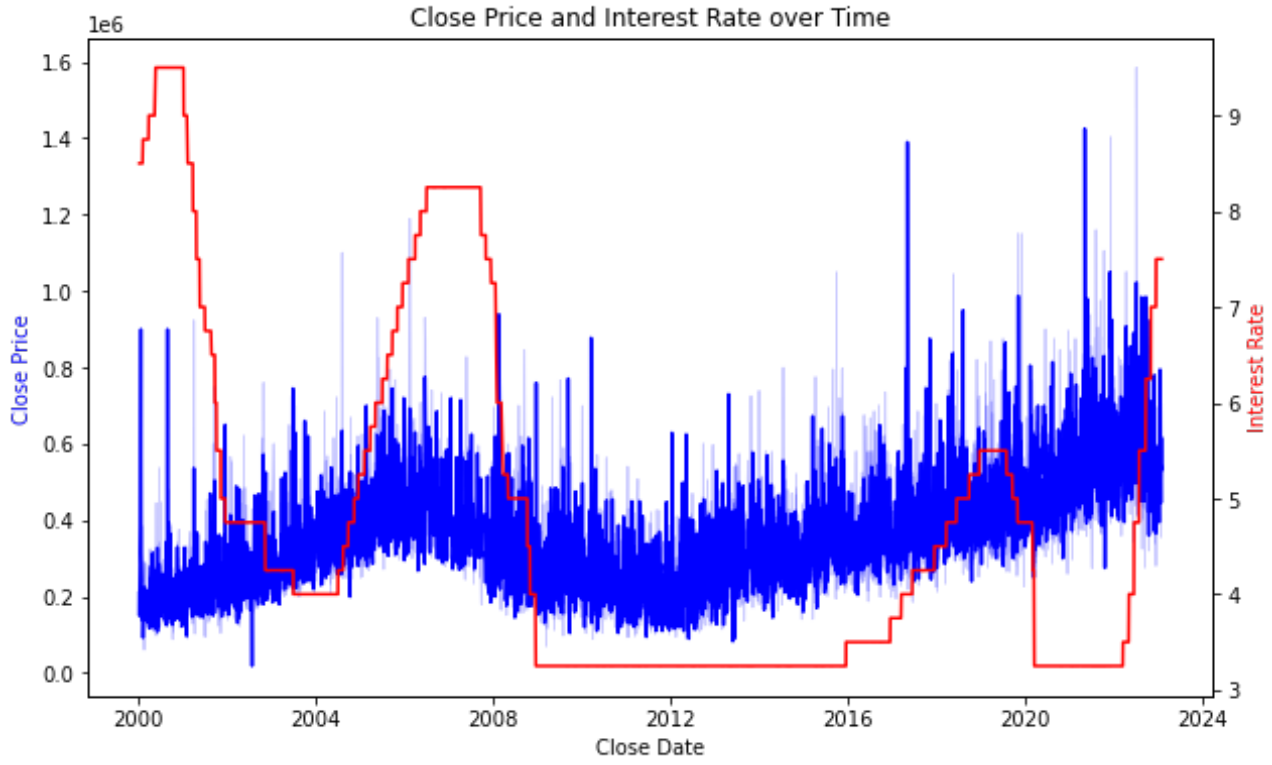
1. ‘DATE’
   Converted to datetime object so that it could be used to merge with Metrolist dataframe.
2. ‘DPRIME’
   After merging with Metrolist, converted to float and removed 11 rows with “.” values.

5. Data Cleaning - CPI

1. Calculated the inflation adjustment factor for each year based on CPI dataset and then added ‘Inflation Factor’ to Metrolist dataframe. We then used the ‘Inflation Factor’ to add ‘Adjusted Original Price’, ‘Adjusted List Price’, and ‘Adjusted Close Price’ to dataframe.

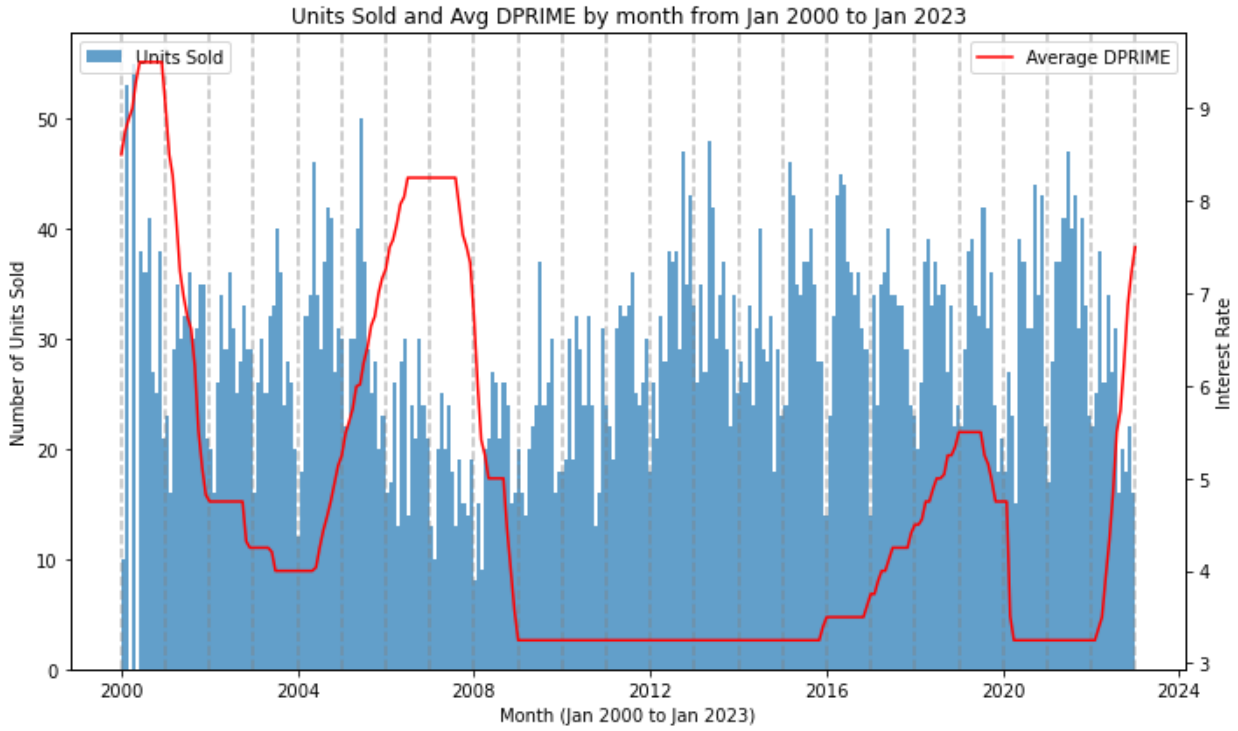
6. Exploratory Data Analysis

Let’s first explore the relationship between ‘Close Price’ (sold price) and ‘DPRIME’ (daily prime interest rate). This is of particular importance today as the Federal Reserve steadily raises interest rates to combat inflation citing established ecomonic theory that as interest rates rise (the cost of borrowing money), demand for goods and services decreases which causes a decrease in prices overall.

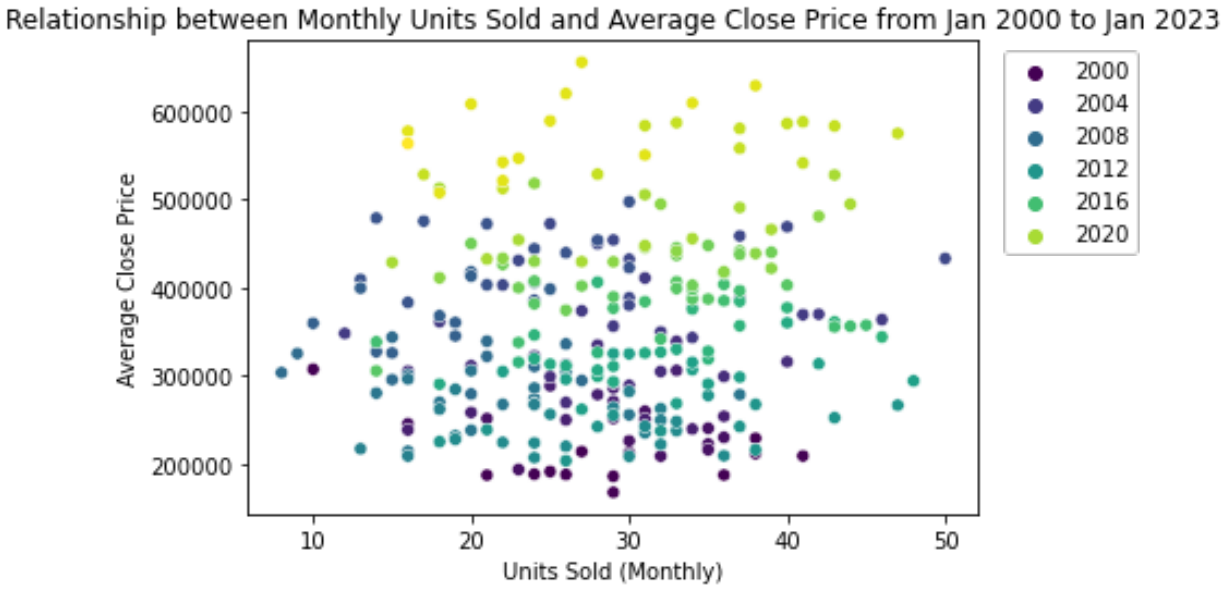


Unit 28.2 - Capstone Two – Final Project Report
Gayla Rios

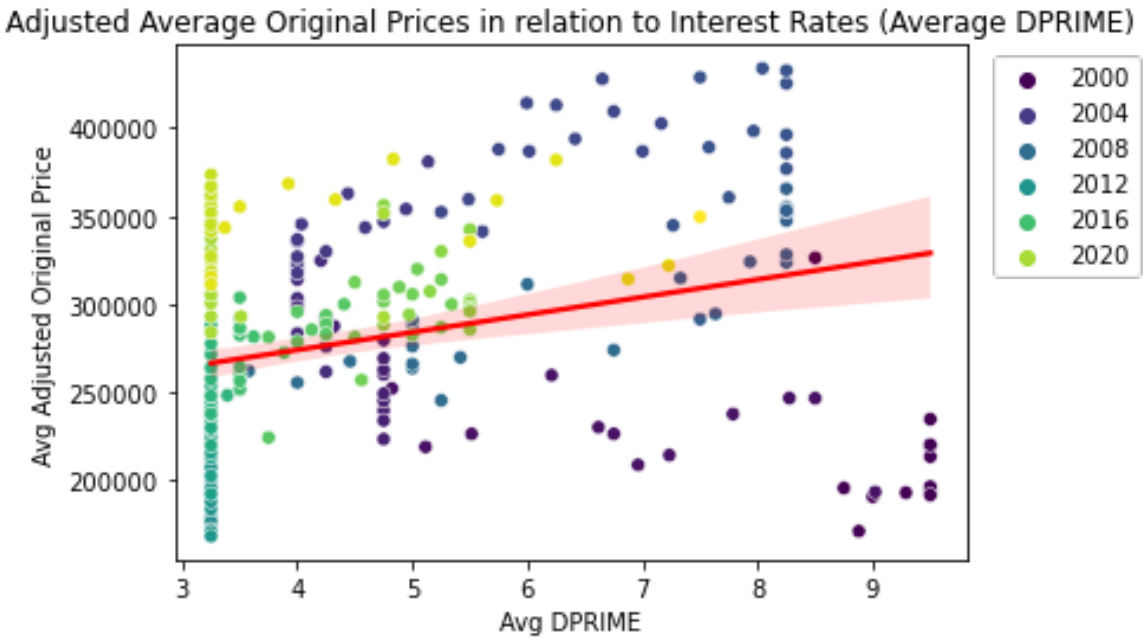
Let's also examine monthly sales volume (units sold) which is an indication of buyer demand. We see that as interest rates are higher, sales volume decreases and as interest rates are lower, sales volume increases.



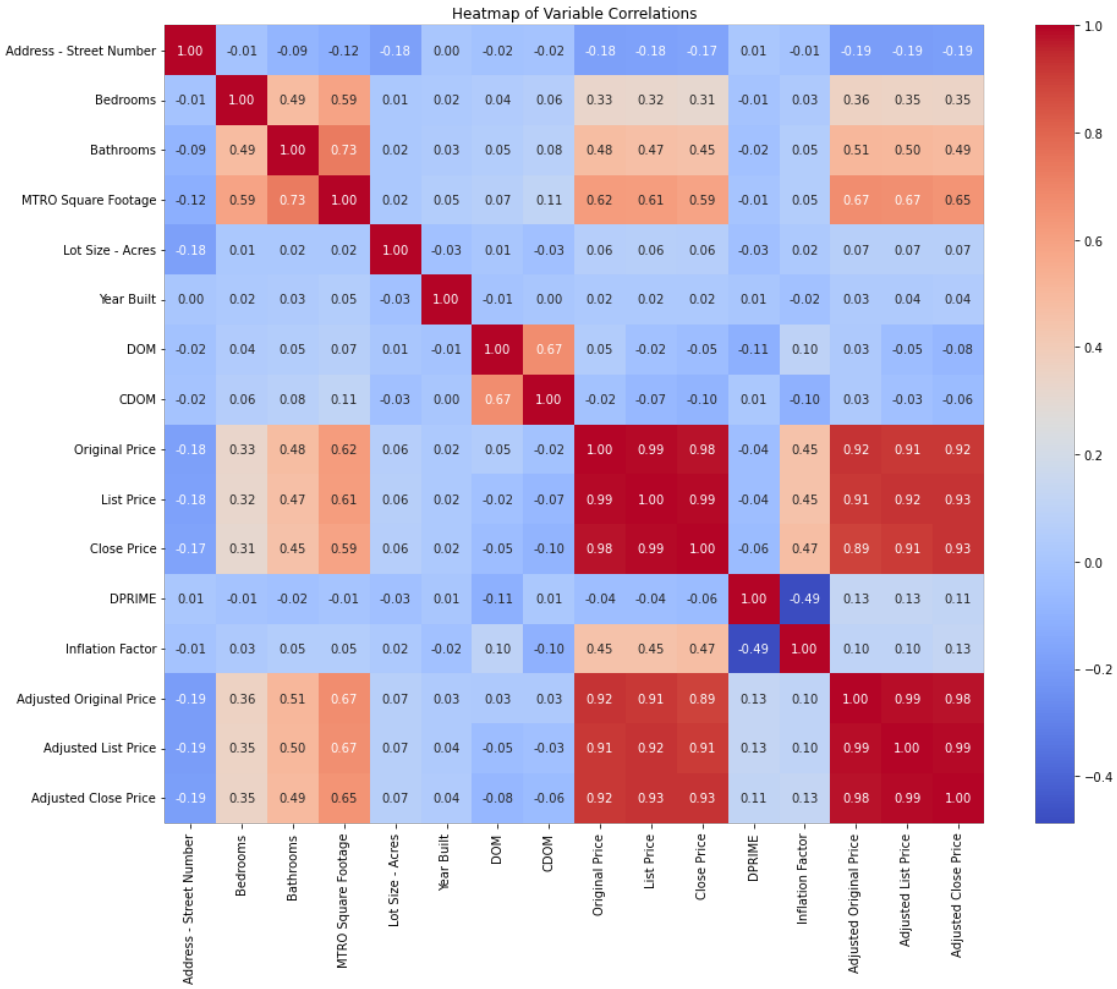
In the chart below, we see sales volume (units sold) and prices of sold homes over time. It looks like part of the increase in prices is due to inflation. With this in mind, I pulled in Consumer Price Index data to see if adjusting prices for inflation clarifies the relationship between prices and interest rates.



Unit 28.2 - Capstone Two – Final Project Report
Gayla Rios



There's definitely something going on here. Let's see if a heatmap could help us determine which variables are most related to one another.



Unit 28.2 - Capstone Two – Final Project Report
Gayla Rios

While I see in the heatmap that 'DPRIME' and the 'Inflation Factor' are in fact inversely related on the heatmap, I'm seeing a weaker relationship between 'DPRIME' and actual prices. This surpises me.

7. Preprocessing

1. Dropped Columns: In preparing the dataframe for modeling, I first simplified the dataframe by removing 'APN', 'Address', 'Original Price', and 'List Price'. I selected 'Close Price' as the target variable.
2. Dummy Features: I then converted 'Close Date' to dummy variables (Month and Year) since it was a date-time feature (not numeric). Future studies might drop 'Close Date' entirely.
3. Scale Standardization: I standardized all the remaining features except the target 'Close Price' and dummy variables for 'Close Date'. I then dropped the 'Close Date' column because it's not numeric and had already been replaced with dummy variables.
4. Split Data into Test and Training Sets: Next I moved 'Close Price' to the last column before splitting the data into X and y subsets where X is all features except the target and y is the target, 'Close Price'. The data was split with 80% for training and 20% reserved for testing.

8. Modeling

I ran 4 models on the training set and selected **Random Forest Regression** because it had the lowest Mean Squared Error, Root Mean Squared Error, Absolute Error and also the highest R-squared.

1. **Linear Regression**
   Mean Squared Error: 5112855197.033736
   Root Mean Squared Error (RMSE): 71504.23202184425
   Mean Absolute Error (MAE): 48132.192234245704
   R-squared: 0.7791768307657686

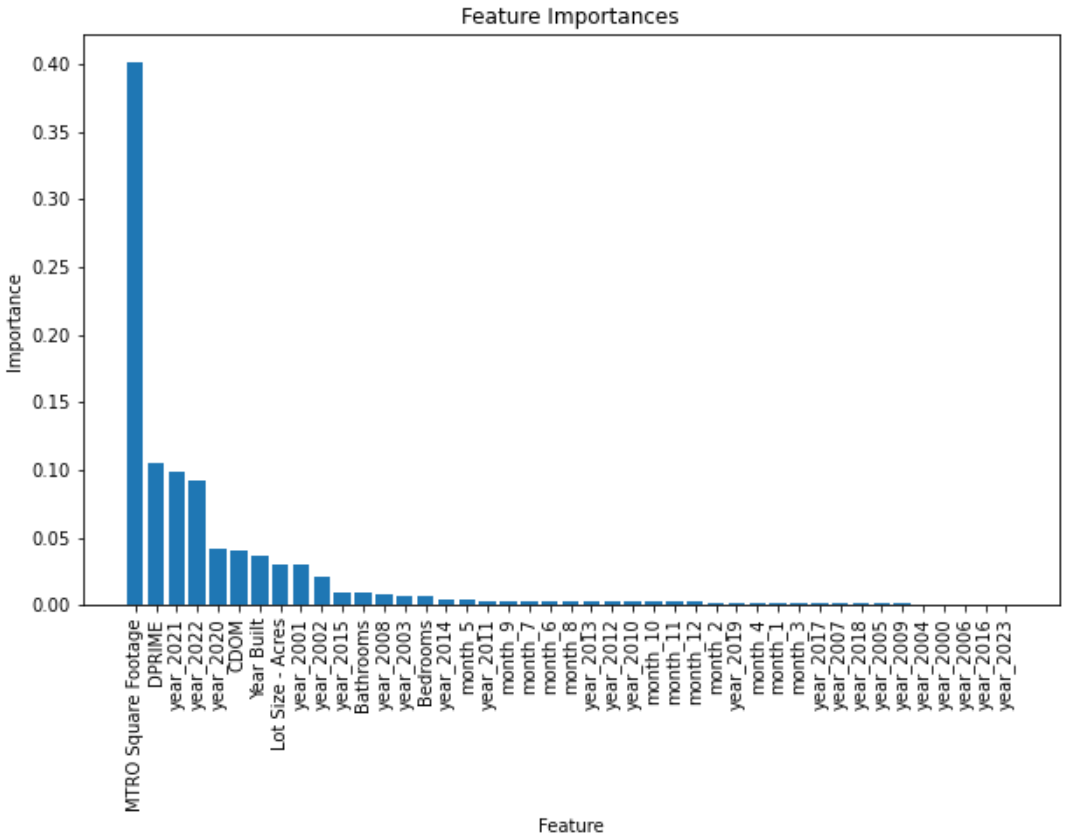
2. **KNN Regression**
   Mean Squared Error: 7493581342.712974
   Root Mean Squared Error (RMSE): 86565.47431114194
   Mean Absolute Error (MAE): 60574.90929344366
   R-squared: 0.6763537559263548

3. **Random Forest Regression**
   Mean Squared Error: 4880777997.125726
   Root Mean Squared Error (RMSE): 69862.5650625979
   Mean Absolute Error (MAE): 46996.30903670698
   R-squared: 0.7892001975179553

4. **Gradient Boosting Regression**
   Mean Squared Error: 5168093763.004319
   Root Mean Squared Error (RMSE): 71889.45515862753
   Mean Absolute Error (MAE): 48276.863356937865
   R-squared: 0.7767910884101754

Having selected Random Forest Regression, let's examine the feature importances. Looking at the chart below, we see that square footage and interest rates (DPRIME) are the two most important features. For future study, I might limit features to the just the top 10 or eliminate all date features. I wonder if reducing the features would increase the importance of interest rates and also increase model accuracy.

Unit 28.2 - Capstone Two – Final Project Report
Gayla Rios



9. Hyperparameter Tuning

Using GridSearchCV, I found the best parameters for the Random Forest Regression Model. Using those parameters yielded the best evaluation metrics. Each of the error metrics – MSE, RMSE, MAE – decreased while R-squared increased slightly as compared to all the other models.

Mean Squared Error (MSE): 4857022005.909885
Root Mean Squared Error (RMSE): 69692.33821525781
Mean Absolute Error (MAE): 46414.38505015631
R-squared: 0.790226213915139

10. Model Visualization

Unit 28.2 - Capstone Two – Final Project Report
Gayla Rios

11. Conclusion

The Random Forest model performs well especially for lower to mid-range of actual home prices. As actual home values increased, the model didn't perform quite as well. We see this clearly in the scatter plot with points corresponding to higher actual prices being farther away from the regression line. Looking at the Feature Importances, square footage and interest rates were the most influential features for predicting home prices. Its of particular importance to note also that the years 2020, 2021, and 2022 were nearly as important as interest rates in terms of prediction value.

12. Further Study

For this model, I focused on seeing how interest rates impact the residential real estate market, but there are many other unique economic factors at play since 2020. Looking at the Feature Importances of our model, we see that the years 2020, 2021, and 2022 were nearly as important as interest rates in making accurate predictions. I'm curious how pandemic lockdowns and subsequent PPP loans as well as student loan payment pauses and student loan forgiveness also impacted the market.