

Capstone 2: Ames, Iowa – Housing Data

1. Introduction:

The US Census Department reports that 28.9% of US household wealth is in the form of home equity, making housing the second most significant store of wealth behind retirement accounts¹.

In fact, for most people in the United States, the purchase of a home is one of the largest financial transactions they will ever make ². Homes are not traded on a liquid market and unlike some other assets the market price of real estate is often difficult to determine. This can leave buyers without a reliable mechanism to determine if they are paying an appropriate sum for a given house.

The goal of this project is to create a model that will use the distinct features of homes to predict the Sales Price of a given house. This type of model could be used in a variety of business settings including by real estate agents who wanted a tool to advise clients or by institutional investors who were looking to purchase properties or by lenders who want to ensure the purchase price of the mortgages they are underwriting are reasonable. Although the dataset is unique to a specific location, this type of analysis could be performed for residential and commercial properties in any location where information from past sales is available.

2. Dataset

For this project, I utilized a publicly available dataset of recent home sales in the town of Ames, Iowa. The dataset consisted of 2,930 distinct sales within the town of Ames between 2006 and 2010. In total there were 82 columns in our data set, each representing a feature of the home. Features include quantifiable characteristics of the home such as:

- Total Square Footage
- Lot Size
- Number of Bedrooms
- Number of Fireplaces...etc.

Additionally, the features include categorical data such as:

- Zoning
- Neighborhood
- Condition of the home
- Whether the garage is attached or detached ...etc

¹ US Census Bureau:

<https://www.census.gov/content/dam/Census/library/publications/2020/demo/p70br-170.pdf>

² National Association of Realtors:

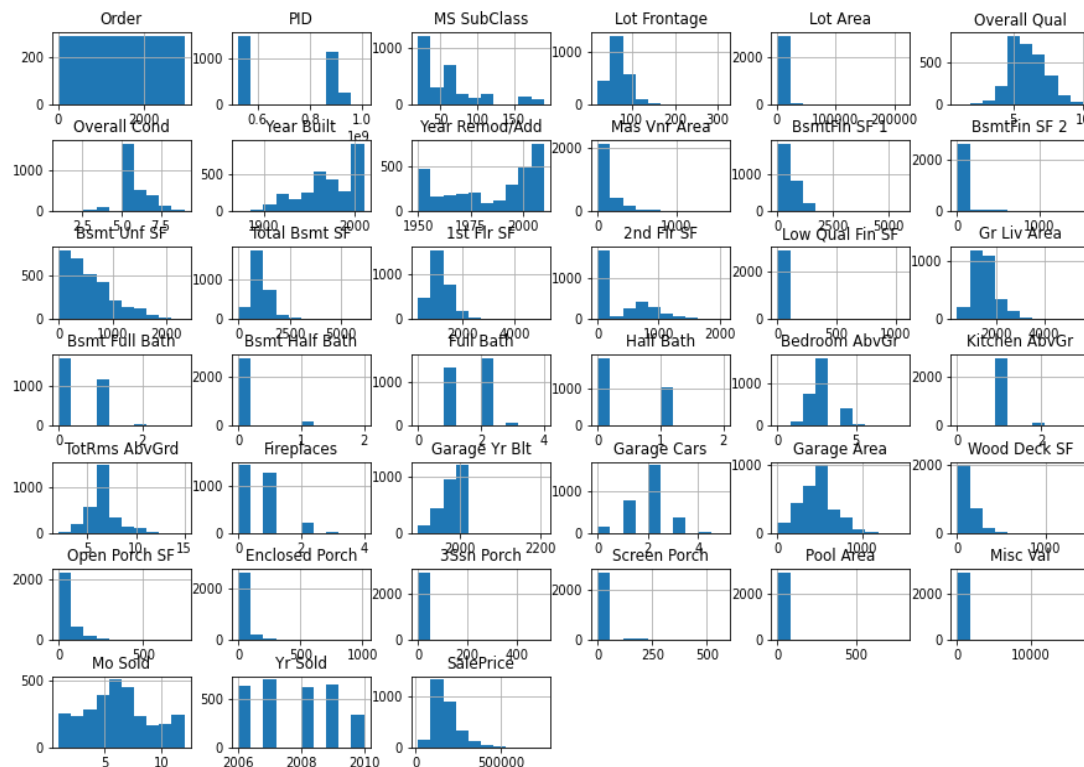
<https://www.nar.realtor/research-and-statistics/research-reports/highlights-from-the-profile-of-home-buyers-and-sellers>

The original data was provided in a .CSV file and was imported into a pandas data frame called housing_data.

3. Data Cleaning and Data Wrangling

Each home in our dataset was assigned a 'Parcel ID' which is a unique identifier that I used to ensure no home sales showed up in our dataset more than one time.

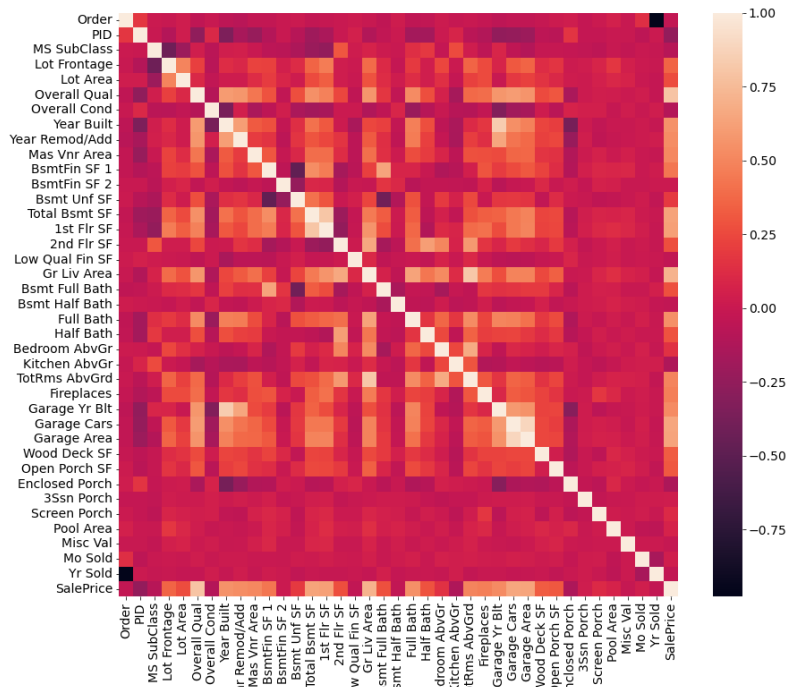
With no duplicated data in our dataset, I graphed all of the numerical data into histograms in order to visually inspect the data for outliers or other anomalies. Some of our data showed outliers, for example the lot frontage on a few parcels was around 300 feet. While wider than most, a lot width of 300 feet seemed reasonable for a smaller town in Iowa and unlikely to be an error.



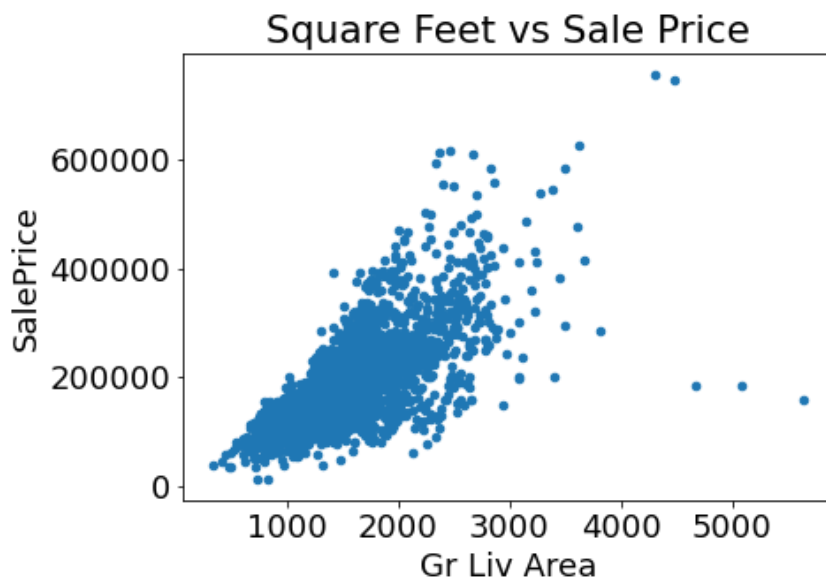
During this step I also looked into missing data and into data that was mostly homogenous. The dataset included four variables where more than 80% of the data was missing. These were examined individually and one variable "Misc Feature" was deleted as it was determined that the value of each miscellaneous feature identified in this column was reported in a separate column.

Similarly, some variables were dropped because the data was exceedingly homogenous. For example, a column named 'Utilities' was dropped since all but three of the 2,930 sales had the same value indicating they were connected to public utilities.

4. Exploratory Data Analysis

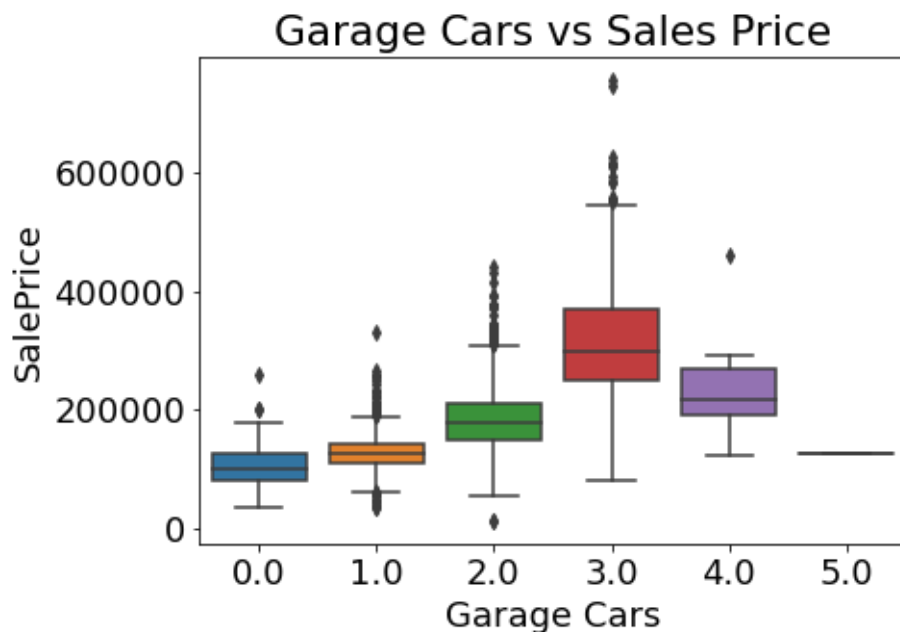


In the exploratory data analysis step, the relationship between variables is assessed. Of particular importance is the relationship between Sales Price and other variables. In the above heatmap we can see a positive correlation between Sales Price and a few variables including Overall Qual, Gr Liv Area, Full Bath, Garage Cars and others.

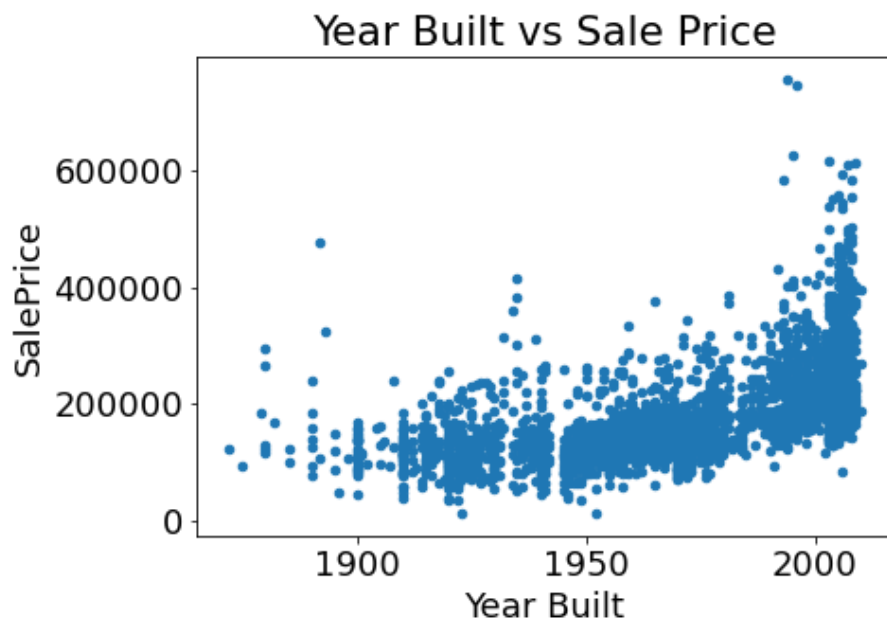


A strong positive relationship between Gross Livable Area (square feet) and sales price is apparent in the scatter plot above.

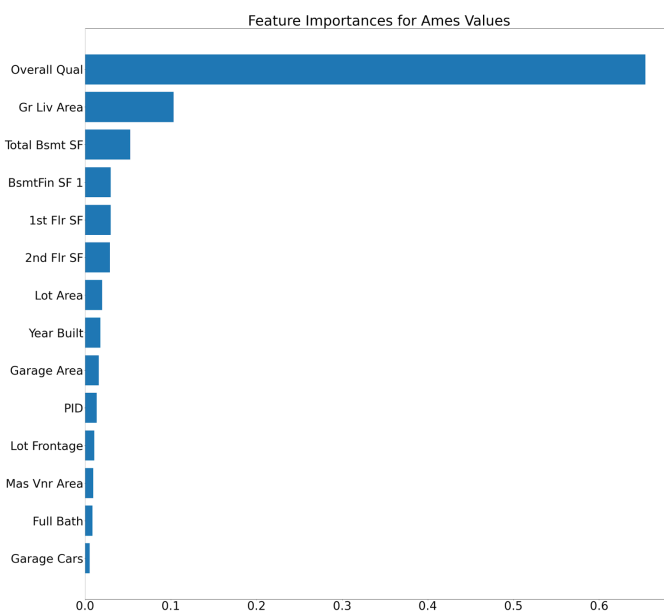
Likewise, a logical relationship is found between sales price and the number of cars the garage will hold where homes without garages sell for a discount while those with three-car garage command a higher sales price.



One additional variable of note is the year built. As shown in the chart below, recently constructed homes appear to command a high sales price.



Finally, during this step I ran a Random Forest Feature Importance analysis.



The results of the feature analysis, shown above, indicate that by far the most important variable in the analysis was the 'Overall Quality' Variable. Unfortunately, there was limited information available regarding how this quality was assessed or measured and by whom. Unlike square footage and lot size I

had concerns about whether this variable would be consistent among various datasets including using this model on new and unseen information. Ultimately, I determined that it would be best to remove this variable altogether.

5. Pre-Processing and Training

Imputing Values and creating Dummy Variables

During this stage some missing values were imputed. A total of 490 home sales (16.7% of the total) were missing the 'Lot Frontage' variable. Instead of removing this column, the missing data for this one variable was imputed with the mean.

Twenty three properties (0.78% of the total) were missing information about their masonry veneers. It was assumed that the lack of information most likely indicated that there was no masonry on the property. The masonry veneer type and area were imputed with their respective modes of 'None' and 'zero square feet' of masonry which is consistent with what I would assume the missing data implied.

One property was missing information about its electrical system. It was assumed that all properties had access to electricity so the value for this property was imputed with the mode.

At this stage, I was careful not to impute data for garages, pools, air conditioning or other features that may not be present in every property.

Since many of our variables were categorical, at this stage I utilized the `get_dummies` function to create dummy variables for these categorical features. One dummy variable from each categorical feature was dropped to avoid unnecessary collinearity.

After imputing data and creating dummy variables, I drop any missing data and the resulting data frame had 2,747 rows (unique sales) and 244 variables (columns).

Addressing multicollinearity

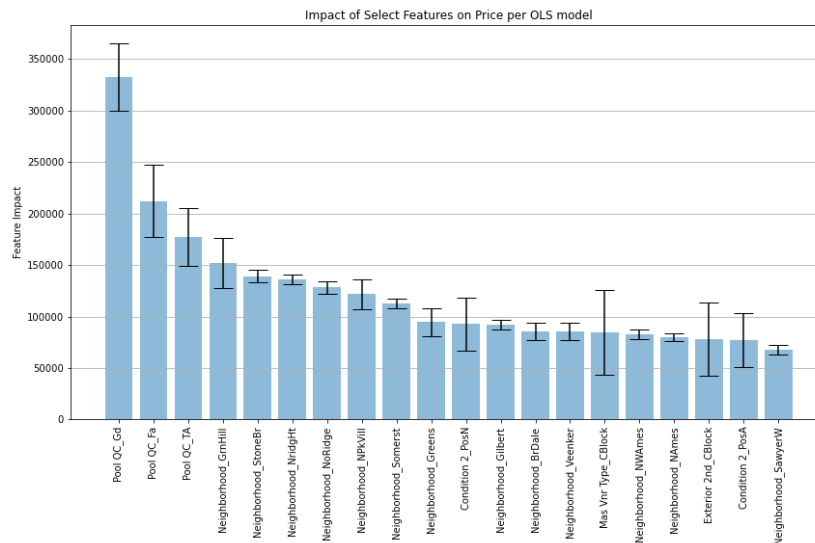
The first step was addressing simple collinearity among my variables. To do so, I examined all variables with a correlation coefficient of 0.95 or greater with any other variable and then simply eliminated one of the variables. In this step I eliminated four variables.

Next, I examined a measure of multicollinearity called the Variance Inflation Factor (VIF). In this step I used some value judgements to eliminate variables with an emphasis on retaining variables who were shown to be important in the Random Forest Feature Importance exercise. The first variables to be eliminated were dummy variables who said basically the same information as another variable for example, condition good and condition fair convey similar information. At this point I eliminated many of the zoning variables as I suspected they may be correlated with other location variables such as 'neighborhood' and because they were not shown to be important in our Feature Importance analysis. This process was repeated using my best judgment until the remaining variables all had a VIF score below 10.

6. Two more measures of Feature Importance

Linear Regression Feature Coefficients

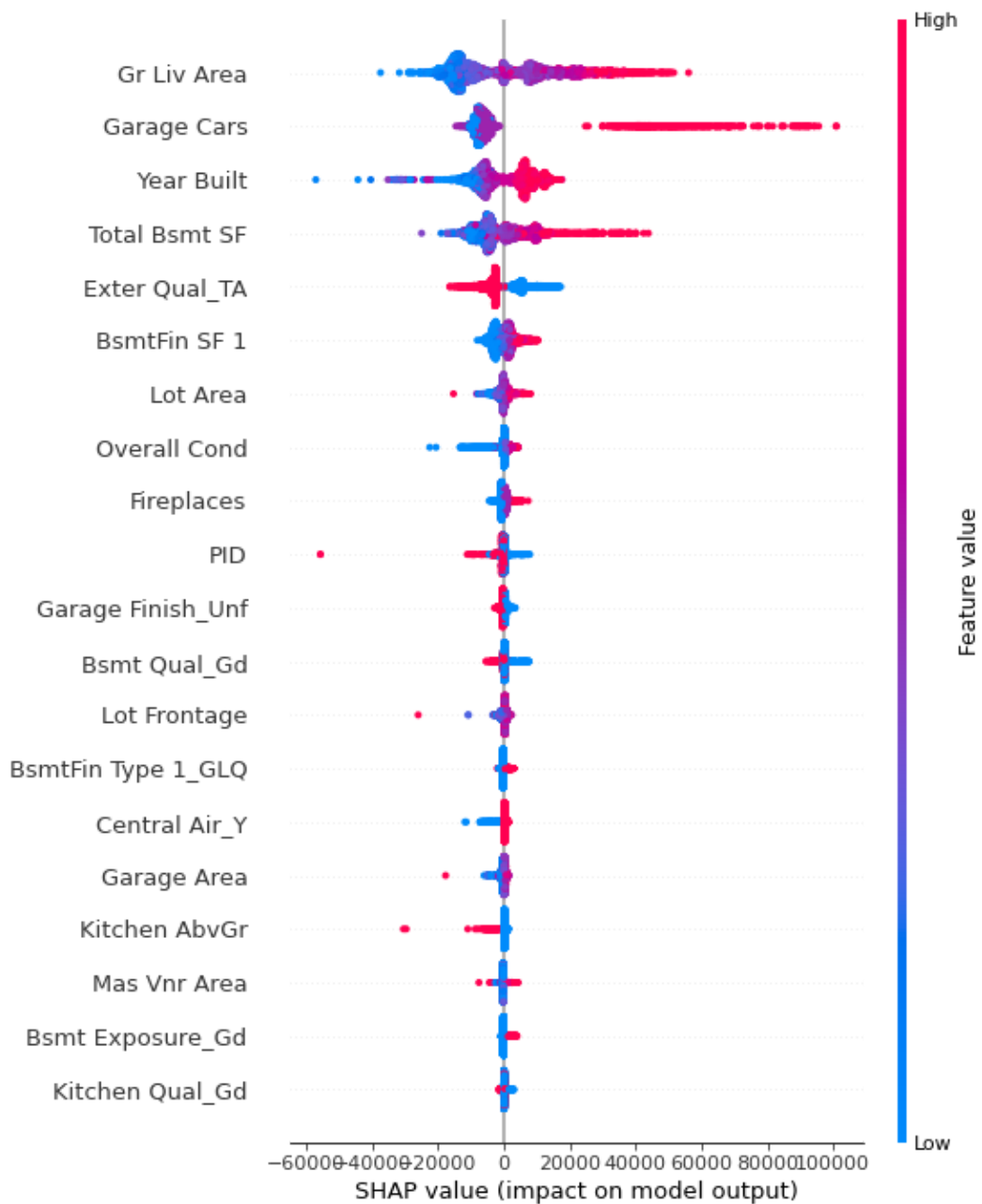
At this stage a significant amount of multicollinearity had been removed from our dataset so another feature importance analysis was conducted, this time using OLS regression. The first step was to standardize the data. Using the Robust Scaler feature, all of the independent variables were standardized. The standardized data was fit into an ordinary least squares regression and the coefficients were extracted from the regression. Given the data was standardized, the coefficients should represent the relative impact of the features on the Sales Price.



What was interesting is that the features identified as most important on the OLS model did not match what we had previously identified as important in the Random Forest Feature Importance. Additionally, many of the variables identified as important in the OLS model were relatively rare features like a swimming pool and dummy variables for a specific neighborhood. It may not make sense to build a model around these fairly rare features so a second Feature Importance analysis was conducted.

SHAP Feature Importance

The second Feature Importance Analysis that I utilized is called the Shapley Additive Explanations (SHAP) which is a measure of the average marginal contribution of a feature across all the possible combinations of features. This analysis, presented below, was more consistent with both logic and the Random Forest analysis previously conducted.



7. Train Test Split

In advance of modeling, I split the data into a training set consisting of 75% of the recent sales and a training set consisting of the remaining 25% of the data. Each of the models was fitted on the training set and then tested on the test data in order to evaluate the performance of the model.

8. Modeling

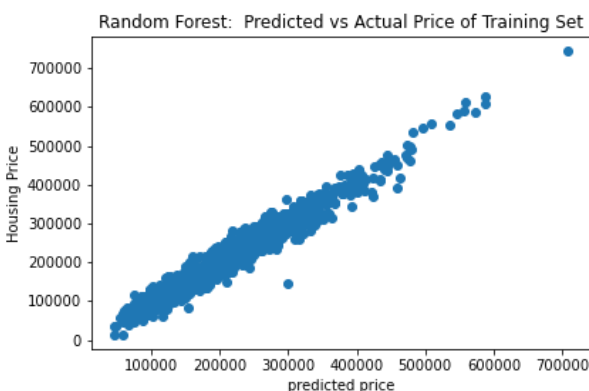
The housing data was evaluated using three different models, Random Forest , Ridge Regression and Gradient Boosting. Each of the three models was run first using all of the features in our dataset and then utilizing only the top twenty features as identified in our SHAP Feature Importance analysis. For both feature sets the relevant hyperparameters were tuned and the results of each model were compared in order to select the best model.

Random Forest

The first model employed was Random Forest. For this algorithm I elected to optimize both the `n_estimators` and the `max_depth` hyperparameters. Instead of a grid search I elected to use a random search to optimize the two hyperparameters since grid search was computationally very expensive. When modeling the entire feature set, it was determined the optimal hyperparameters were `n_estimator` of 5,308 and `max_depth` of 20. When modeling only the top 20 features the optimal parameters were `n_estimator` of 10,000 and `max_depth` of 110.

Random Forest Evaluation Metrics	Model: All Features	Model: Top Twenty Features
r2 Score Train	0.983	0.983
r2 Score Test	0.862	0.856
Mean Absolute Error (MAE) Train	6189.75	6308.99
Mean Absolute Error(MAE) Test	17905.1	18336.8
Root Mean Square Error (RMSE) Train	10268.8	10250.1
Root Mean Square Error (RMSE) Test	29797.4	30411.4
Optimal <code>n_estimator</code>	5308	10000
Optimal <code>max_depth</code>	20	110

For the Random Forest model, using all of the features resulted in a slightly more robust model with both a higher r-squared and a lower root mean square error. This model shows a strong relationship between actual and predicted values.



Ridge Regression

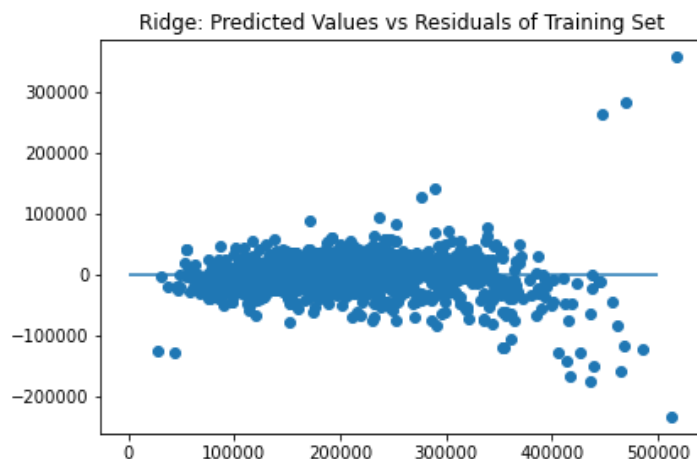
Ridge Regression is useful when multicollinearity still exists between the variables. Given the large number of features in our model, a ridge regression was used to model the data set.

The hyperparameter of alpha was optimized using Grid Search cross validation and a ridge regression was run using the optimized alpha.

Ridge Regression Evaluation Metrics	Model: All Features	Model: Top Twenty Features
r2 Score Train	0.88	0.789
r2 Score Test	0.883	0.829
Mean Absolute Error (MAE) Train	17309.2	22532
Mean Absolute Error(MAE) Test	18166.6	22350.1
Root Mean Square Error (RMSE) Train	27466.3	36415.1
Root Mean Square Error (RMSE) Test	27442.7	33171.5
Optimal Alpha	6	35

Once again, if we compare r-squared and RMSE, then using the full feature set resulted in a better model than using the top twenty features from our SHAP feature importance.

For the ridge regression I plotted the residuals to visually ensure linearity, and a constant variance. As shown below, nothing unusual was detected in the plot of the residuals.



Again, this model produced a satisfactory R-squared and a slightly lower RMSE than the Random Forest model.

Gradient Boosting

The final algorithm was a gradient boosting model. Utilizing a randomized search cross validation method, I optimized the learning rate and the max depth hyperparameters. As with our other models, using the full feature set resulted in the most robust model. This model performed slightly better than our other model when measuring both the R-squared and the RMSE. Some evaluation metrics of the Gradient Boosting model are shown below.

Gradient Boosting Evaluation Metrics	Model: All Features	Model: Top Twenty Features
r2 Score Train	0.974	0.97
r2 Score Test	0.89	0.881
Mean Absolute Error (MAE) Train	9603.46	10340.5
Mean Absolute Error(MAE) Test	16981	17789.9
Root Mean Square Error (RMSE) Train	12869.2	13781.8
Root Mean Square Error (RMSE) Test	26604.8	27666.8
Optimal Learning Rate	0.1	0.1
Optimal Max Depth	4	4

Selecting a Model

For this project I used three different regression algorithms to model the Ames, Iowa housing data with the objective of finding the best model to predict a dependent variable (Sales Price) based on the features of the property. Each of these three algorithms was evaluated using both the full feature set (all of the variables) and a reduced feature set with the top 20 variables as identified in our SHAP feature importance.

For each model I used a grid (or randomized) search cross-validation to optimize the hyperparameters.

A key metric used to compare these three models was their R-squared, which is the extent to which the variance in the independent variables influences the variance in the dependent variable. This is often referred to as the 'Goodness of Fit' and is an indication of how well the data fits the regression model. A higher R-squared will indicate a model that creates a better fit of the data.

Of our models, the Gradient Boosting (All Features) yielded the best R-squared on the testing data.

The other metric I focused on was the Root Mean Square Error (RMSE) which is a measure of how far our predictions fall from our actual Sales Price. The RMSE is in the same units as our dependent variable (US Dollars) and is therefore easy to conceptualize its magnitude. A lower RMSE is preferred.

Model	r2 Score (Test-Data)	Root Mean Square Error (Test Data)	Hyperparameter 1	Hyperparameter 2
Random Forest (All Features)	0.862	29797.42	n_estimators = 5,308	max_depth = 20
Random Forest (Top 20 Features)	0.856	30411.44	n_estimators = 10,000	max_depth = 110
Ridge Regression (All Features)	0.883	27442.69	Alpha = 6	
Ridge Regression (Top 20 Features)	0.829	33171.55	Alpha = 35	
Gradient Boosting (All Features)	0.89	26604.81	Learning Rate = 0.1	max_depth = 4
Gradient Boosting (Top 20 Features)	0.881	27666.77	Learning Rate = 0.1	max_depth = 4

Since the objective of the model is to determine the price of a real estate offering where the sales price is unknown, therefore the r2 Score and the RMSE score of the Test Data (not the Training Data) was used for comparison. Given the superior performance of the Gradient Boosting (All Features) model both in terms of its R-Squared and RMSE on the test data, I would select the Gradient Boosting Model (All Features) and expect it to most accurately predict the Sales Price.

Selected Model	Gradient Boosting (All Features)
-----	-----
Model	Gradient Boosting (All Features)
Hyperparameter 1	Max Depth = 4
Hyperparameter 2	Learning Rate = 0.1
r2 Score Train	0.974
r2 Score Test	0.89
Root Mean Square Error (RMSE) Train	12869.0
Root Mean Square Error (RMSE) Test	26605.0

9. Running a Test Case

To illustrate the functionality of the model constructed I randomly selected one property from our test data and used the gradient boosting model to predict the sales price of the home.

On this particular home, the model predicted a house price of \$247,968. When I looked up the actual sales price of this home, it in fact sold for \$226,000. While a single sale is not a good indicator of the performance of the model, it is nonetheless fun to put the model to use in a manner that simulates its real-world application!

10. Conclusion

The goal of this project was to build a model that would use the features of recently sold homes to predict the sales price of a home in that same market.

This type of model has a number of business uses including:

- Real Estate agents advising their clients
- Investors valuing potential purchases
- Mortgage Lenders conducting due diligence prior to lending etc

Although our data was limited to single family residential real estate transactions in Ames, Iowa, this type of analysis could be conducted for multifamily, condominium, or commercial real estate in any number of geographical locations.

The data set consisted of 2,930 real estate transactions in the city of Ames, Iowa between 2006 and 2010. There were 82 columns in the data set each representing a feature of the home. Features were both categorical and numerical.

Once the data was rigorously cleaned, a SHAP feature importance analysis was conducted which revealed our top 20 variables. The data was split into a testing and a training set.

Three distinct models were employed to analyze our dataset. Each model analyzed both our full feature set as well as the top 20 feature set as identified in our SHAP analysis. Hyperparameters were tuned and the models were compared based on their r-squared and their root mean square error. One model

stood out as having both the highest r-squared and the lowest RMSE and was therefore selected as the optimal model for this dataset.

The model selected was a Gradient Boosting Model that incorporated all of our features (not just the top 20). This model had an r-squared of 0.89 and a root mean square error of \$26,605 compared to an average sales price of \$180,796.

One test case was then run to demonstrate the utility of the model.

A significant amount of time was spent optimizing the hyperparameters and each of our models included either one or two hyperparameters to tune. However, it is likely that including additional parameters in a future analysis could result in a more robust model.

Furthermore, while we assessed each model using two feature sets, additional feature sets i.e. top 50 variables may provide a more robust model.

Finally, the data available is more than a decade old, so future analysis could include revisiting with more updated data or evaluating different types of real estate including condominiums, or commercial buildings.