Ames Housing Data

Predicting Housing Price with Regression

Predicting Housing Price with Regression

This study applied regression techniques to predict housing sale price using the Ames Housing dataset, which contains information from individual residential properties sold in Ames, IA from 2006 to 2010.

The purpose was to discover both the most useful modeling techniques for predicting sale price, as well as to uncover actionable insights for homeowners and buyers.

1. Data Cleaning and EDA
2. Preprocessing & Feature Engineering
3. Regression Models
4. Conclusions & Recommendations

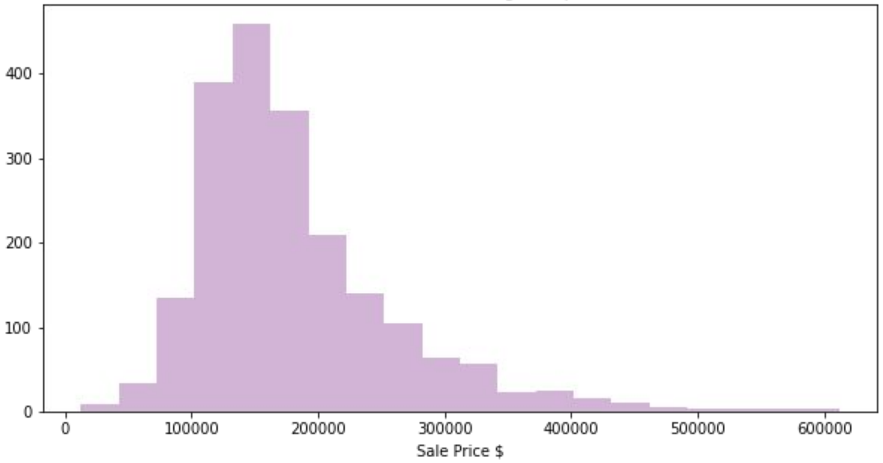
Data Cleaning & EDA

The Data

The Ames housing dataset contains:

- 2051 observations
- 82 individual variables
- Sale Price data from the years 2006-2010
- 878 observations with no Sale Price info

Distribution of housing sale price



The Process

Cleaning steps for:

- 20 continuous variables
- 14 discrete variables
- 23 nominal variables
- 23 ordinal variables

Dealing with Outliers

Missing Data

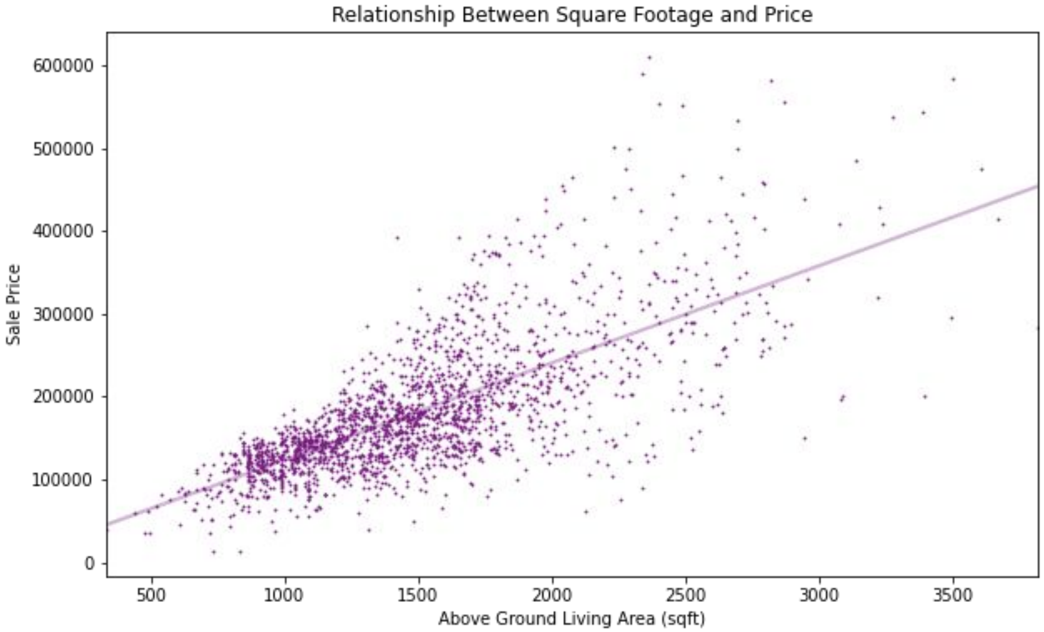
Continuous Variables

20 Continuous Variables

Assumed missing values meant the feature didn't exist for that property for 'Lot Frontage' and 'Mas Vnr Area' and set missing values to 0.

Outliers: Removed two rows with 'Gr Liv Area' greater than 4,000 square feet, following guidance from data dictionary.

Missing Data: Two rows were also removed for having significant missing data for several columns.



Discrete Variables

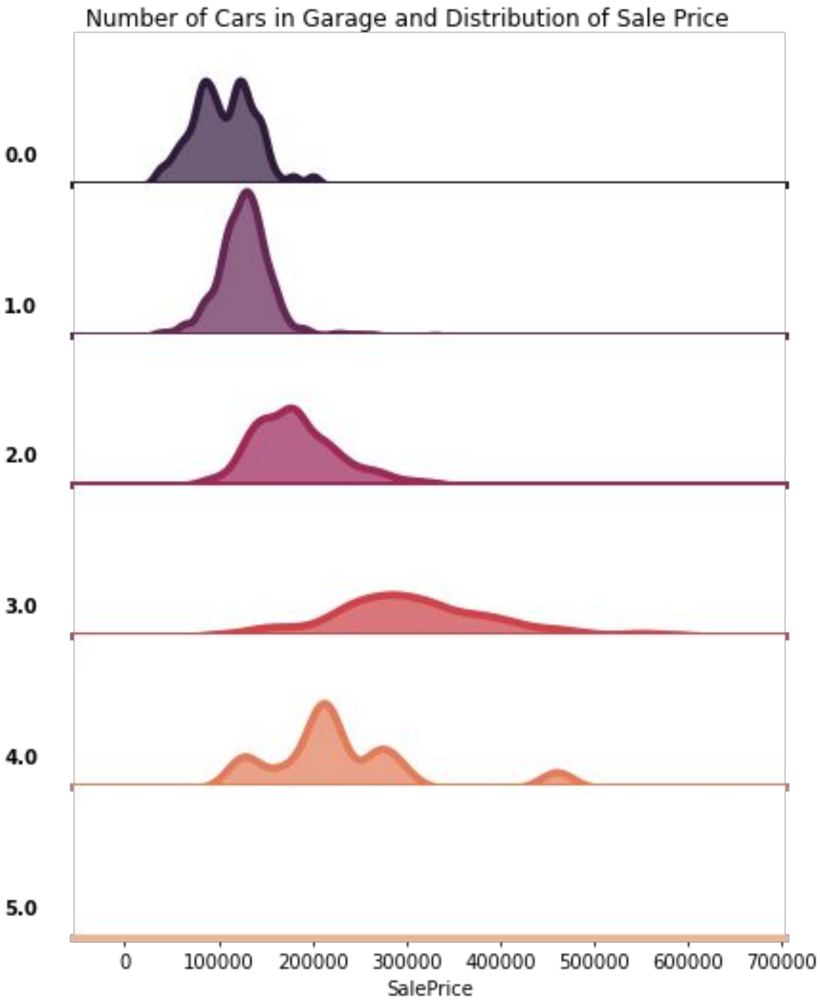
14 Discrete Variables

Filled missing values for ‘Bsmt Full Bath’, ‘Bsmt Half Bath’ to zero.

Imputation: Imputed ‘Garage Yr Blt’ based on the average value for houses built in the same year.

Feature engineering: Built new features for more interpretable results (more on this in next section).

‘Mo Sold’ and ‘Yr Sold’ were treated as Nominal variables.



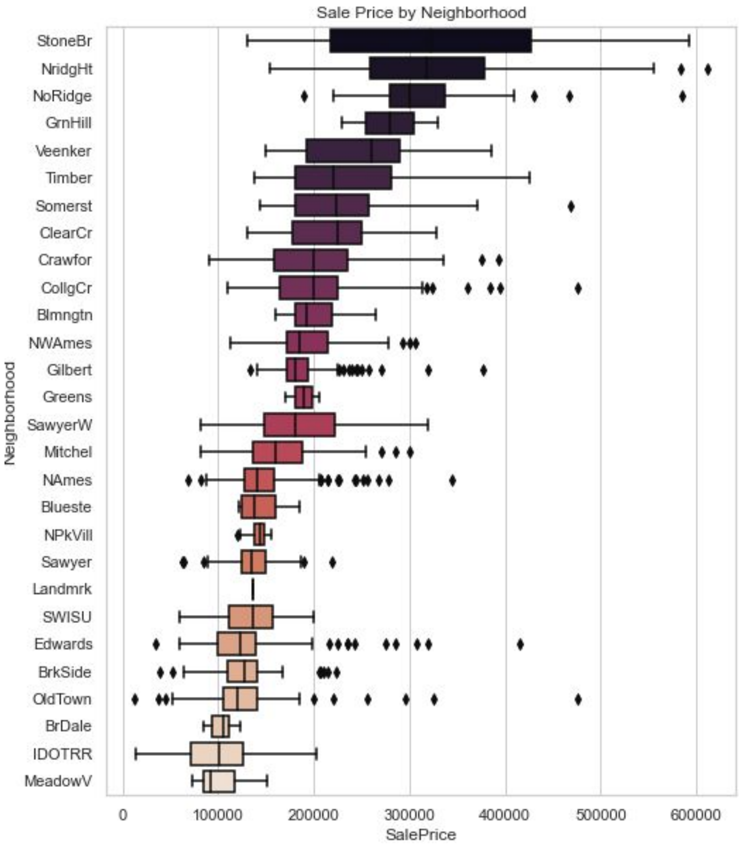
Nominal Variables

There were 23 Nominal variables.

Any missing values were given the label 'None'.

'MS SubClass' was converted from an integer to a string.

All nominal variables were one hot encoded into dummy columns.

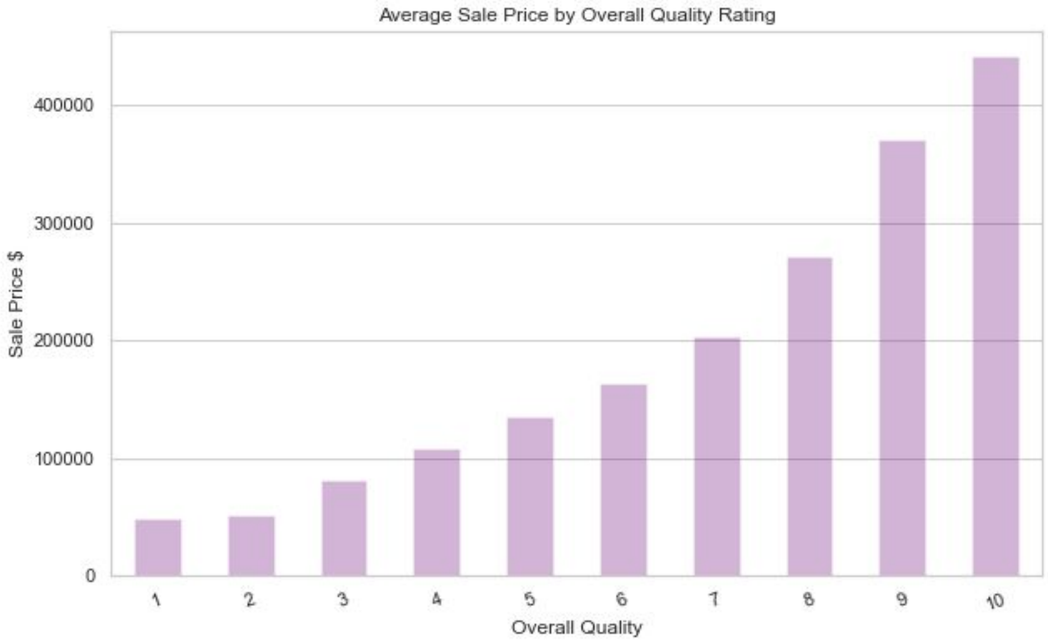


Ordinal Variables

There were 23 ordinal variables.

All ordinal variables were assigned positive numerical values based on hierarchical information provided in the data dictionary.

For example, the 'Bsmt Qual' variable was recoded so that 'Poor' was assigned a value of 1 and 'Excellent' was assigned a value of 5.



Feature Engineering & Preprocessing

Feature Engineering

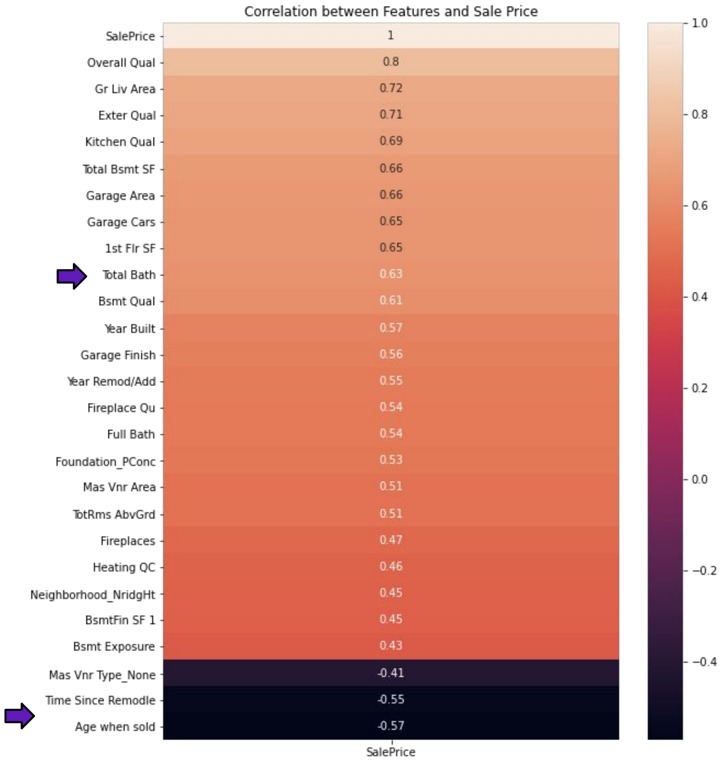
At the end of the data cleaning process there were 229 total predictors.

Additional variables were created to improve model interpretability:

'Total Bath': sum of all bathrooms in the house with a .5 weight on half baths

'Time Since Remodel': The difference between 'Year Sold' and 'Year Remodel/Add'

'Age when sold': The difference between 'Year Sold' and 'Year Built

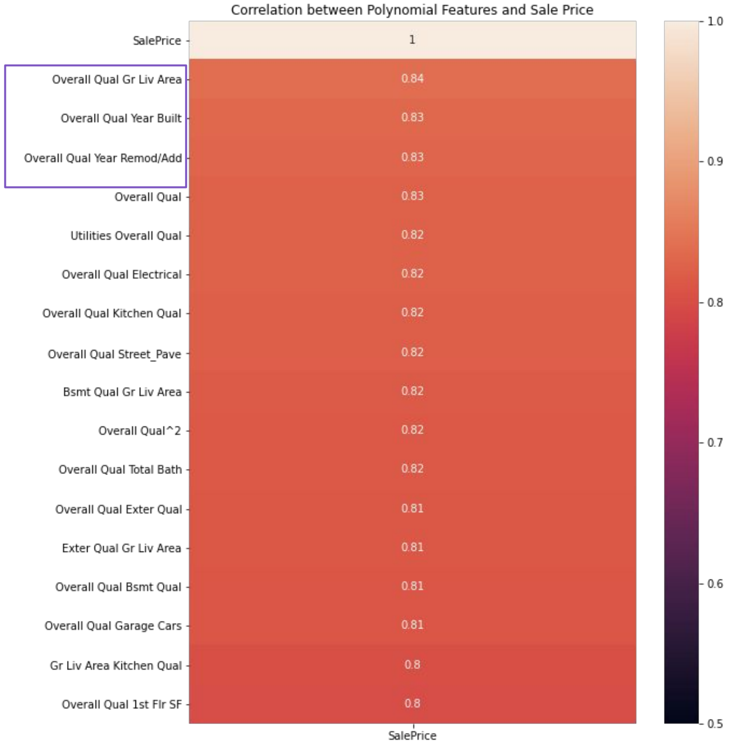


Polynomial Variables

In addition to features built by hand, PolynomialFeatures was used to automate the generation polynomial and interaction derivatives of all variables.

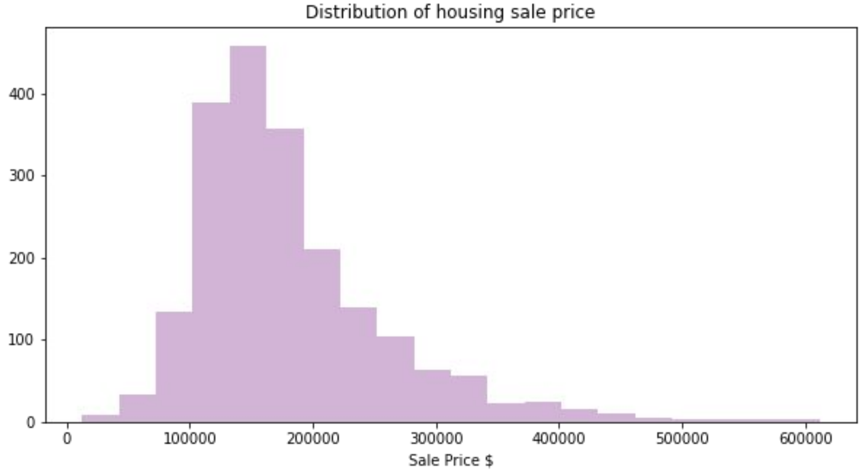
Over 27,000 additional features were created. Using subsets of the new features most correlated with Sale Price did not improve modeling results.

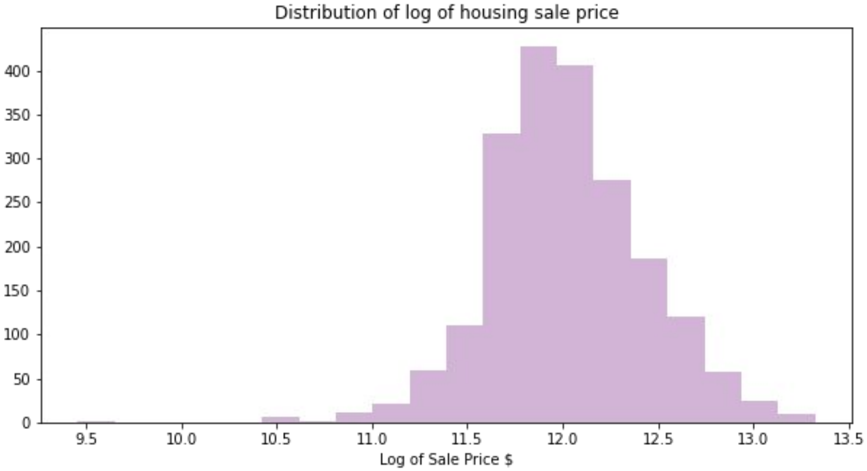
Only the top 3 interaction variables with correlation coefficients higher than the most correlated original feature were kept.



Preprocessing

1. Train/Test split using default test size of 25%
2. Scaled train and test data after fitting on training data
3. Log transformed Sale Price to have more normal distribution on target variable
4. After evaluating models on holdout data, retrained models on full data set


Distribution of housing sale price


Distribution of log of housing sale price

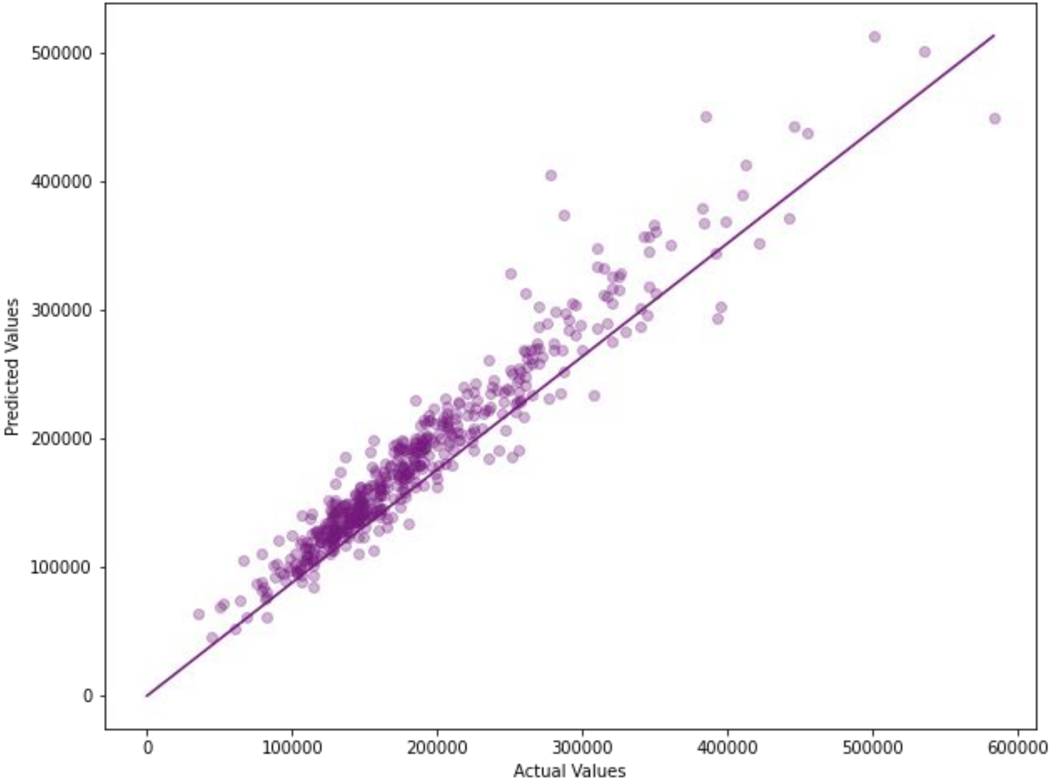
Results & Recommendations

Ridge Regression

The model with the lowest Root Mean Squared Error (RMSE) for the Kaggle submission was a Ridge Regression which iterated over 100 values to optimize alpha.

- The optimal alpha: 422.9
- R2 for Training set: 0.93
- R2 for Test set: 0.92
- The RMSE for the Test Set: $21,075

This model explains 93% of the variance of the Sales Price for our test data. The high and similar R2 for Train and Test indicate this model should perform well on unseen data.

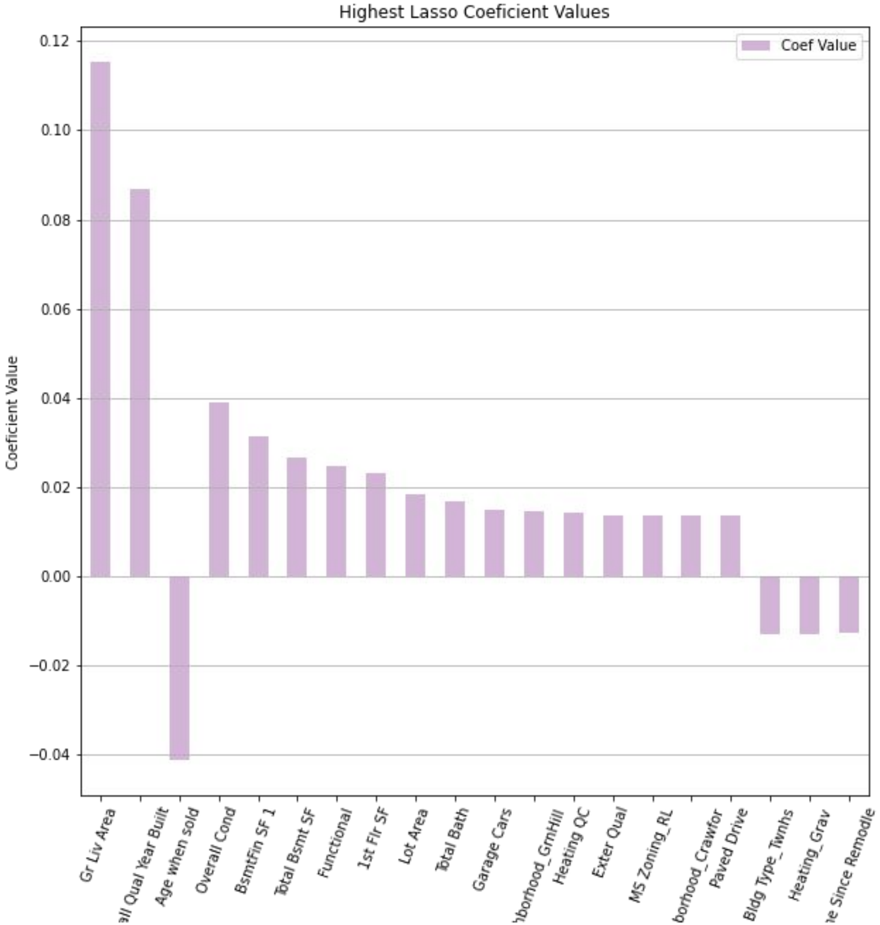


Lasso Regression

Despite not performing as well on the Kaggle set, Lasso was used to help determine features most likely to be useful for a more interpretable linear model.

- The optimal alpha: 0.0028
- R2 for Training set: 0.93
- R2 for Test set: 0.93
- The RMSE for the Test Set: $20,909

This plot shows the highest Lasso coefficient values after regularization. These features were used to develop the final (and most interpretable) linear model.

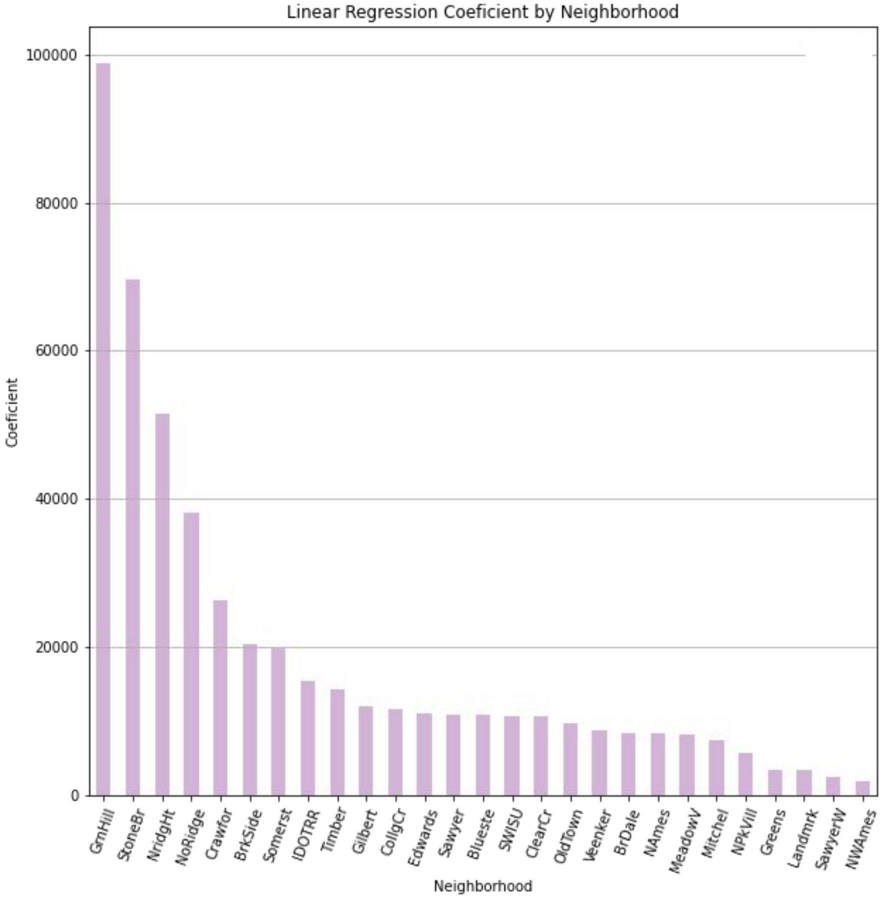


Linear Regression

Linear Regression was used on features determined by Lasso Regularization to produce interpretable coefficients.

- R2 for Training set: 0.91
- R2 for Test set: 0.91
- The RMSE for the Test Set: $23,423

This model still explains 91% of the variance for Sale Price on the testing data, and does not show signs of overfitting. Furthermore, the coefficients provide meaningful information about how much additional value each feature provides when controlling for all other variables.



Linear Regression

Linear Regression was used on features determined by Lasso Regularization to produce interpretable coefficients.

- R2 for Training set: 0.91
- R2 for Test set: 0.91
- The RMSE for the Test Set: $23,423

This model still explains 91% of the variance for Sale Price on the testing data, and does not show signs of overfitting. Furthermore, the coefficients provide meaningful information about how much additional value each feature provides when controlling for all other variables.

Interpretation Examples:

Holding all else constant, for every…

- Square foot increase in living area we can expect a $54 increase
- One unit increase in overall quality we can expect a $10,344 increase
- Additional car garage space available, we can expect a $7,062 increase

...in Sale price

Conclusions

With appropriate data cleaning techniques, regularized regression performs well fitting the Ames housing data set to Sale Price with minimum feature engineering.

Linear Regression without regularization provides quantifiable insights for homeowners and potential buyers:

- Neighborhoods matter - with some neighborhoods commanding a $100,000 price premium, all else held constant
- Recent Remodels, Add Ons, Basement finishes can improve the value of the home
- Space is at a premium - over $50 per square feet
- Projects improving the heating or exterior quality of the home are probably worth it, but specific home improvement costs can be checked against model coefficients to evaluate return on investment

Thank you!