Ames Housing Data

Using Regression Models to Predict the Sale Price of Homes

Veronica Giannotta
12/07/2018

Ames, Iowa

The city of Ames, Iowa has a population of 65,000, and is considered a desirable place to live. The Ames housing data set is a compilation of individual residential properties sold in the city from 2006 to 2010. Based on this data, it is possible to build a regression model that can predict the sale price of a home, as well as identify the leading factors in the market value of a home.

Modeling Sale Prices of Homes

Modeling Sale Prices of Homes

- Systematically clean and transform data features
- Conduct exploratory analysis, in order to understand the relationships between variables, and their wider implications on housing market value
- Build and fit a predictive model, based on a curated group of features from this data set, that can be used on unseen housing data to accurately predict sale prices

Modeling Sale Prices of Homes

What was in the data set?

- 82 columns and 2930 rows of housing data
- Compiled between 2010 and 2016
- Nominal, ordinal, discrete, and continuous values
- Descriptive features of homes and their most recent sale prices

Cleaning the Data

Cleaning the Data

Considerations

- Large number of columns
- Varying orders of magnitude
- Sizeable quantity of categorical data
- High probability of multicollinearity between variables

Cleaning the Data

Steps Taken

- Visualized the most important correlations
- Appropriately dealt with null values
- Converted categorical data into numeric values wherever possible
- Created interaction columns for strongly correlated features

Exploratory Analysis

Exploratory Analysis

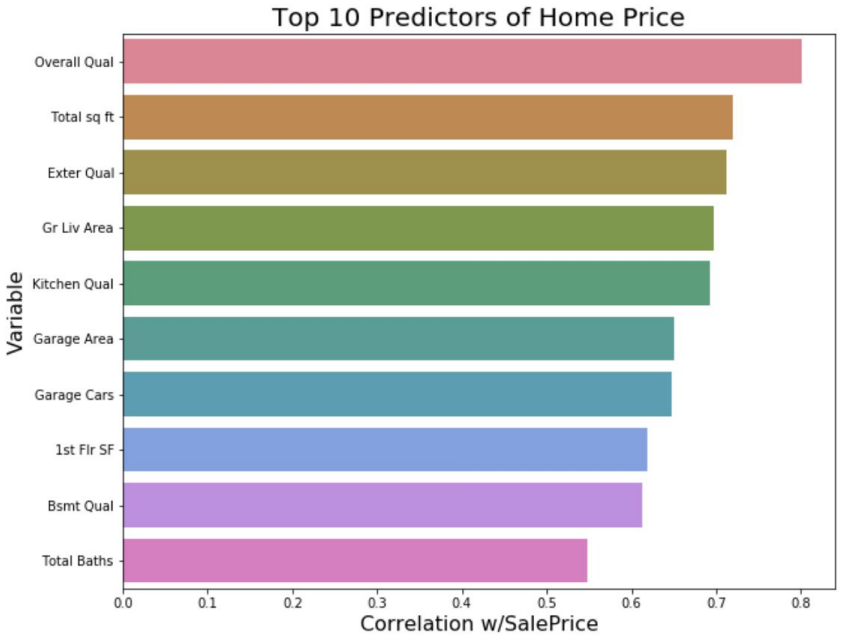
Visualizing the Sale Price Distribution

- Right-skewed distribution
- Outliers in the higher price ranges

Exploratory Analysis

Correlations with Sale Price

- Overall Quality
- Total Square Feet
- Exterior Quality
- Above Ground Area
- Kitchen Quality
- Garage Area
- Garage Car Capacity
- First Floor Square Footage
- Basement Quality
- Total Baths



Preprocessing & Modeling

Preprocessing & Modeling

Preparing the Data for Modeling

- Selected a final features list based on correlation with the Sale Price variable
- Scaled both variables
- Ran cross-validations
- Fit Linear Regression, LassoCV, and RidgeCV models to the data

Preprocessing & Modeling

Preparing the Data for Modeling

- Selected a final features list based on correlation with the Sale Price variable
- Scaled both variables
- Ran cross-validations
- Fit Linear Regression, LassoCV, and RidgeCV models to the data

LassoCV r2 Scores:

Training Score: 0.8165
Test Score: 0.8687

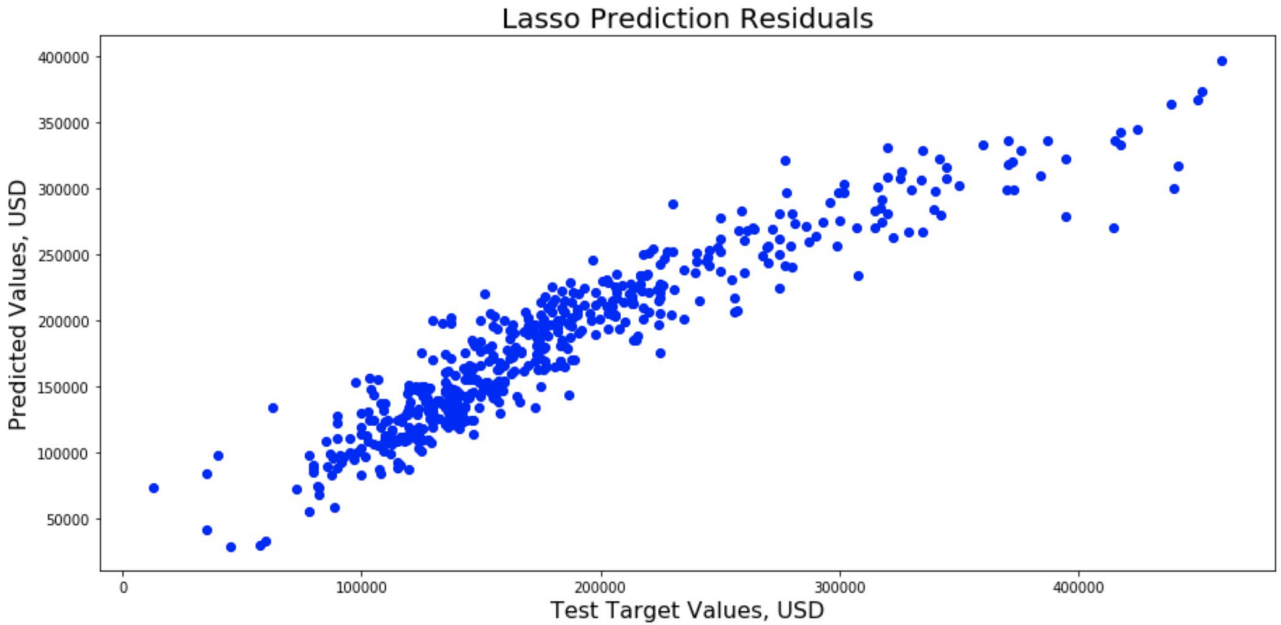
Modeling

Results

Modeling

Results

	Coefficient
Overall Qual	19693.580132
Exter Qual	10786.159755
Total sq ft	9651.743522
Kitchen Qual	8250.417916
Gr Liv Area	5832.191019
Garage Cars	5407.029956
Bsmt Qual	5308.482250
BsmtFin SF 1	5022.147672
Fireplace Qu	4673.339035
1st Flr SF	4464.944497
Mas Vnr Area	4040.707778



Insights

Best Predictors of Sale Price:

Insights

Best Predictors of Sale Price:

- Overall Quality
- Exterior Quality
- Total Square Footage
- Kitchen Quality
- Living Area Square Footage
- Garage Car Capacity
- Basement Quality
- Basement Condition
- Fireplace Quality
- First Floor Square Footage
- Area of Exterior Masonry

Thank You.