Understanding How to Predict the Value of Homes in Ames Iowa

Audience

The intended audience for this project would be someone looking at homes in the Ames, Iowa area while focusing on whether the home is priced correctly. While comparing all the different home factors and prices would be a tedious process for someone, they would be able to input all the information about their particular home into my linear regression model. This model would serve as a tool to help this potential homeowner. It would allow him/her to compare a home they are interested in purchasing to the rest of the data set and do a single home vs the entire data set as a train set.

Motivation

As I enter into the next stage of my life, I am starting to think of home ownership. I am feeling overwhelmed by the process simply because I want to make sure that I get a good deal in any home I look at seriously purchasing. I want to make sure that I understand what features make play a role in a fair priced and quality home that potential buyers would be attracted to. Initially, I wanted to use the built model to locate underpriced homes. I chose the Ames, IA housing data set to start my project. One way to achieve this goal was to compare the estimated price of the home to the actual price of the home and start looking at homes in the Ames, IA area.

Some of the variables that had to be taken into consideration when valuing the homes were the square footage of the home, age of the home, finished vs unfinished basement, size of the garage and the neighborhood. With these variables in mind, I realized the goal required a more holistic approach to best synthesize the data. Over the course of the project, I was able to arrive at a conclusion through a root mean squared error.

Methodology

The goal of the project was to build a model to predict the value of the house based upon the given factors which include the size of the home, age of home, number of bedrooms, and a list of other important factors. The method required the skills of linear regression. I wanted to take it a step further and visualize the outcome to show a case by case predicted vs. actual price of the home.

Data Set Description

The project used a data set provided by the city of Ames, Iowa. The set included roughly 3,300 separate instances over a 4-year time span. Each home included roughly 80 variables with it. Some extensive cleaning went into the data set during the organizing period. One column was dropped for missing 17% of the values. By the end of the cleaning process, every value was filled in and dummy variables were used in the case of categorical values. After running a correlation test, the variables with poor correlation were removed.

Data Wrangling

Like many other large public data sets, the Ames home data had a number of missing values. To ensure the data was relevant to the project early on, it was necessary to make sure that there was an actionable data set which would allow for a holistic exploration of the data. Beyond filling in the NA values and removing columns with large amounts of missing data, we had to use some additional libraries to help fill in the additional missing information.

The Ames housing data set is comprised of all the home sales from 2007 to 2010. There are 3,300 tickets and about 70 different columns of info on each home sold. We were provided with the sale price, roof style, house style, building type, garage size, basement size, living area, number of fireplaces, year built as well as some other information. Using some Pandas, I added columns for the age of home, years since remodel and size of the home. After this step, I removed the respected columns used to build these columns to ensure these features did not receive too much weight during the building of the model.

When looking at the data set initially, it is clear that a lot of attention was placed into just submitting as much of the information as possible. The columns with the most missing values still had 83% of the information present. Later on in this project, I removed this column and a few others because there was poor correlation between some categories such as the pool area, month sold, size of the porch, basement finished, year sold, order, basement half bath, and low quality finished square feet.

The other challenge that arose during the data wrangling portion included creating dummy variables for the categorical values. I found the categorical data by selecting data type. The categorical data was moved to dummy variables so that they could be used alongside the numerical data.

I finally ended the data wrangling portion of the project by fixing the skew of the data. When checking the skew of the data, I found that the values were not properly arranged. To fix the skew, I transformed the columns so they would be values between 0-1.

I was very pleased with the final outcome of the data. There were no missing values, the categorical data were transformed, and the skew was fixed which allowed me to confidently move onto the model creation portion of the project.

Exploration of the Data

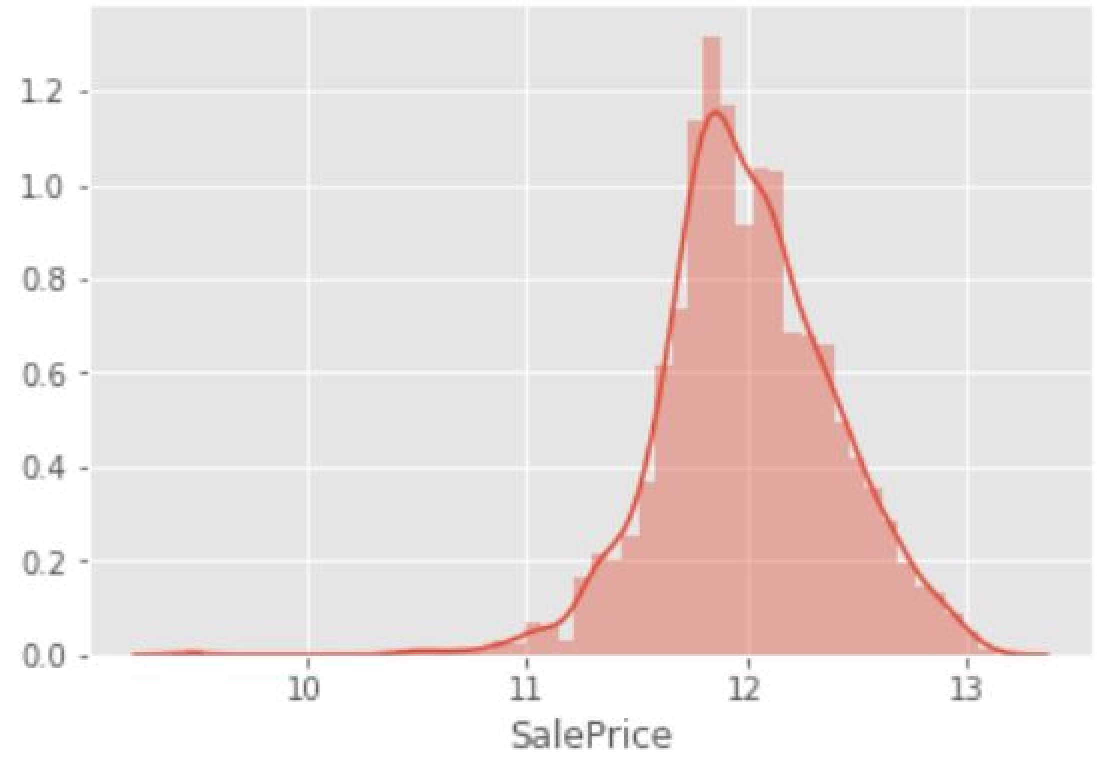
In my spare time, I work on renovating foreclosed homes with my family. I specialize in custom brickwork and building custom staircases. My goal when taking on this project was to understand when a home is priced well and what features need to be taken into consideration when purchasing a home. I found this Ames, IA data set valuable because of how many features it included for me to work with.

The first hurdle that I had to overcome was cleaning the data. After countless hours of reviewing tutorials and source code, I finally produced a clean data set. My largest challenge with this was the creation of the dummy variables for the categorical data. After finally getting the data set cleaned, it allowed me to move closer to my goal of figuring out the value of the home using the homes sold between the years of 2007 to 2010.

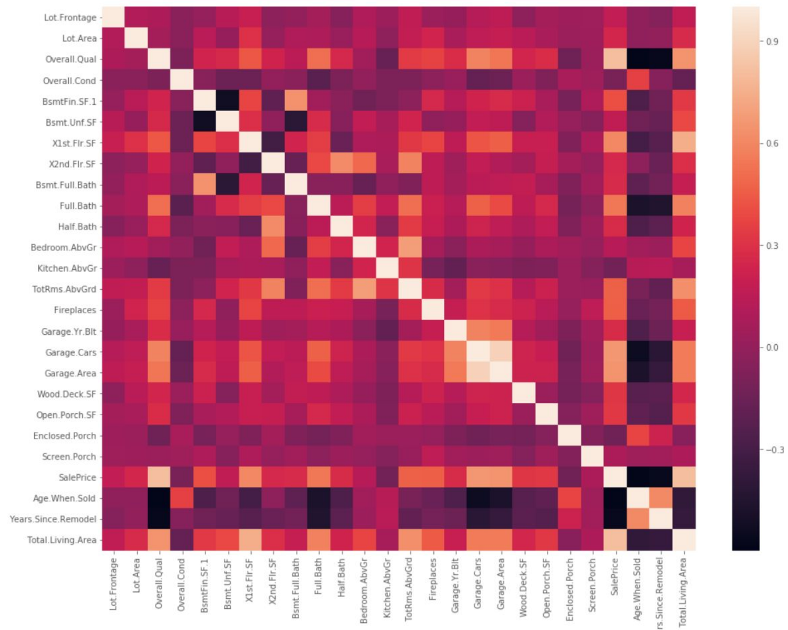
With my final goal of seeing how close I could get the predicted value of the home to the actual value of the home, I started off by checking the correlation of independent variables to the sale price of the home. I then removed any variables that had less than an absolute value of +- .1. At this point, I removed 8 columns from the data set because of the lack of correlation. I then removed the extreme outliers in regards to the size of the home.

After removing the poorly correlated values, I turned my attention to converting the categorical values to binary values. I created numerical values for the categorical values then connected them back to the numerical values. I decided to call the new table “results” because this made it clear that the data in this column presented my findings and reminded my audience of my purpose in analyzing this data set.

I then turned my attention to the skew of the data. At first, I scored a 1.16 skew on the numerical data. I was not pleased with such a high skew number and I was able to discover that the reason this skew was so high was due to how large my values were in the data frame. For example, when comparing a value of 1 for half baths to a value of $700,000 for the sale price of a home, the $700,000 would bring a lot more weight and skew the results. To improve the skew of the data, I transformed the columns to fall between the values of -0.5 to 0.5. Once this step was completed, I reevaluated my skew and found that I had a more desirable skew of -0.21.



Correlation Between Visuals



The image above is a heat map that shows how much of a correlation can be found between different variables in the data. In this case, I compared each numerical value inside of the project to itself. As you can see, the downward trending white line shows a perfect correlation when comparing a value to itself. The column that was transformative to helping me analyze this data set was the “sale price” column. This column, when compared with the other variables considered in the heat map help explain what features are key players in determining a fair priced and quality home.

Model Summarized

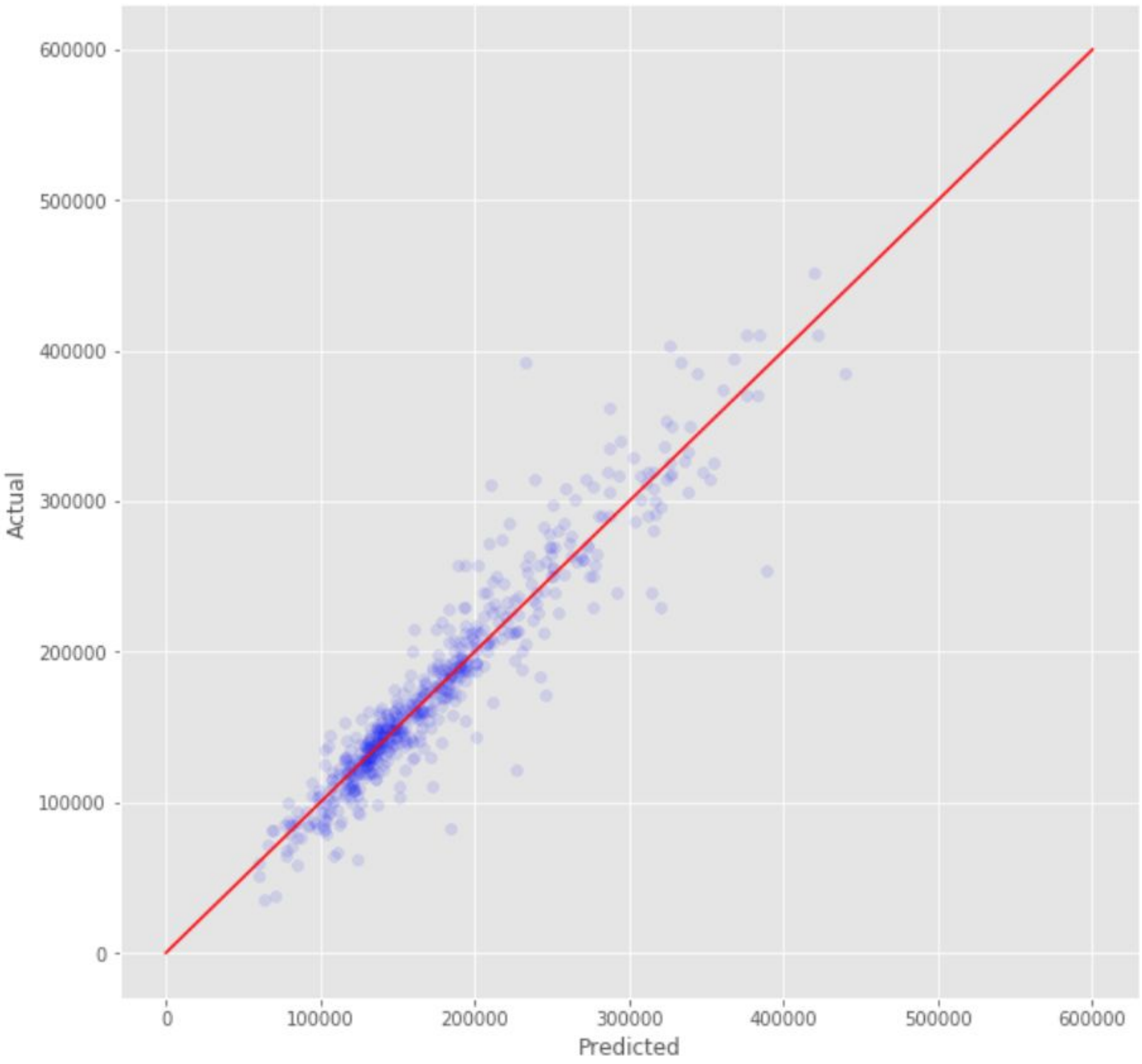
At this next stage in my project, the data is transformed into a set with no missing values and all the inputted information is properly skewed. Next, I took my new data set and split it into a train and test set. I did a 70/30 split and then for the second step of this split, I did 80/20. I then dropped the sale price variable from the test set.

I wanted to see how the root mean squared error scores compared on a 70/30 split vs an 80/20 split. The 70/30 split scored a 20,942 on the test set. The 80/20 RSME split scored a 21,527 on the test set. The score was better on the 70/30 split because the mean averaged out more with the larger test set. I am happy with this RSME score as it is relatively low for 2,900 values.

My first RSME score before I started changing the parameters of the equation produced a 0.84 RSME score. This was decent but I wanted to try to get a better score out of the equation. After some tuning on the model, I was able to get the score to improve to a 0.902 on the 70/30 split and a 0.894 on the 80/20 split. I was pleased with these results but wanted to also visualize what these models looked like and so I created an actual price vs. predicted price graph.

Plotted RSME MODEL

The below image shows a plotted RSME model. If the model was perfect, we would see a 45* line. I am very pleased with this line as there are just a few outliers on it. The x-axis of the graph is the predicted price of the home and the y-axis of the graph is the actual price of the home. It appears that up to around the $200,000 dollar mark, the model is able to do a great job of valuing the home.

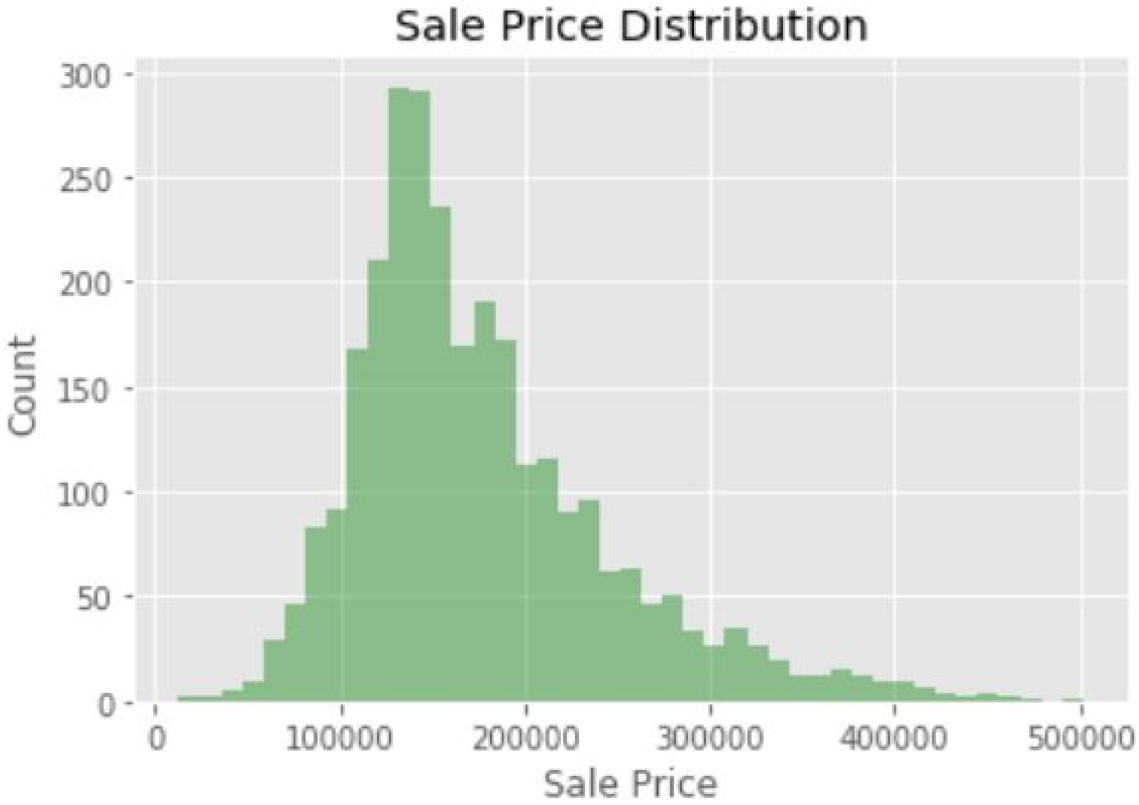


Trends in the Data

After completing the task at hand, I wanted to find some of the fun facts held inside of the deep dive. First, I looked at the most expensive and the cheapest homes in the data set. I found that the most expensive home sold for $501,837 and the cheapest home sold for $12,789 dollars.

Sale Price Distribution

The histogram below is the sale price distribution of the homes. I changed the histogram to show the value count by the number of homes per value. The mode of the data frame falls at $150,000 with roughly 575 homes near $150,000.



Walk Away

For my first RSME model, I am very happy with the results and potential for further interpretation of project conclusions. I learned a lot about data cleaning as well as modeling the data at the end. The score of 0.902 was very competitive. The take away from the data is that there are many variables that affect the value of a home, as such I think it would be extremely difficult, if not nearly impossible for anyone to build a perfect model. Despite the struggle to build a perfect model, it was encouraging to be able to draw conclusions from the problem statement, code, models, and clean data set.