README

Ames Housing

Problem Statement:

Is there a correlation between certain features or traits of a house that can clearly affect its price?

EDA & Data Cleaning:

Looking at the data provided, we can see there are a few columns of data that do not contribute much as they contain null values. These columns being 'Alley', 'Misc Feature', 'Fence', and 'Pool QC', all of which contain above 80% of their values being null. The loss of these columns will not affect our model as they provided little data already.

So are there values we want to keep? **YES!** Looking more at our data there are columns we need to give numeric values to build our model to be more accurate. This process led me to converting qualitative data into quantitative data that our model can read. Taking the columns that were marked with labels that could be scaled quantitatively as '5, 4, 3, 2, 1'.

For the remaining variables I created dummy columns to use in our machine learning model to give accurate predictions on which columns have the strongest correlation to the sale price of a home.

Exploratory Analysis:

Upon analyzing the data we can see a strong correlation between some categories and the final sale price of the house. In real estate it is often said that there are three things that affect price - location, location, location! In our analysis we can

see this to be true, as examining a visual of price per neighborhood tells the story of our data. Other areas that impact our overall price are square footage, number of bathrooms, and the amount of cars a garage can hold.

Pre-processing:

Using train test split and standard scaler we were able to scale our model. While using Numpy log to convert our y variable 'Sale Price' to be more normally distributed, giving us a better model to build up predictions. Since we use log, we also must call the use of Numpy's exponentiate when we call our predictions to scale our model back to its previous scale.

Modeling:

Having tested all three models: Ridge, Lasso, and Linear Regression. I have found that Lasso has performed the best, especially due to its ability to reduce variables that have a lower coefficient to the target. I have found that building our features on those areas that have the strongest correlation make not only the best sense to our data, but also logically.

Given how our model works with the features that are used, we could train the model using the same features in another area. This gives our model scalability, which makes it a great tool for us.

Business Recommendations:

The feature I find to add the most value to a house is the location. It is clear that the neighborhood in which a house is located will be the most important factor in pricing our house. This makes sense as school zones have a significant impact on

the demand of a location in which a home is located. Size of the home also has a strong impact, the more square footage and bathrooms a home contains, the higher the value.

My recommendations for our business:

- Purchase homes in Stone Brook, North Ridge Heights, and North Ridge as they have the highest price on average, in the $300,000 range.
- Strive to purchase houses whose overall quality is above Good, and has a large amount of square footage.

Recommendations for Home Owners:

- Renovating areas in your home will provide a boost in the overall quality and value, thus positively impacting price.
- Additions provide a way to increase square footage and bathroom count, both of which will significantly increase the price. Additions also provide an opportunity to renovate other areas in the home and therefore have the potential to give the most value in one solution.

Sources:

Ames School (outside research):
https://www.amestrib.com/news/20180817/ames-school-district-ranked-best-in-state-for-sixth-year

Data Description: http://jse.amstat.org/v19n3/decock/DataDocumentation.txt

Kaggle Challenge: https://www.kaggle.com/c/dsi-us-8-project-2-regression-challenge/leaderboard

Tableau Charts:
https://public.tableau.com/profile/eric8719#!/vizhome/AmesDash/Sheet2?publish=yes