GA DSI Project 2

(*Ames Housing Data and Kaggle Challenge*)

- By ***Solomon***

Contents:

1. Problem Statement
2. Executive Summary of Model
3. Possible Inferences
4. Recommendations

1. Problem Statement

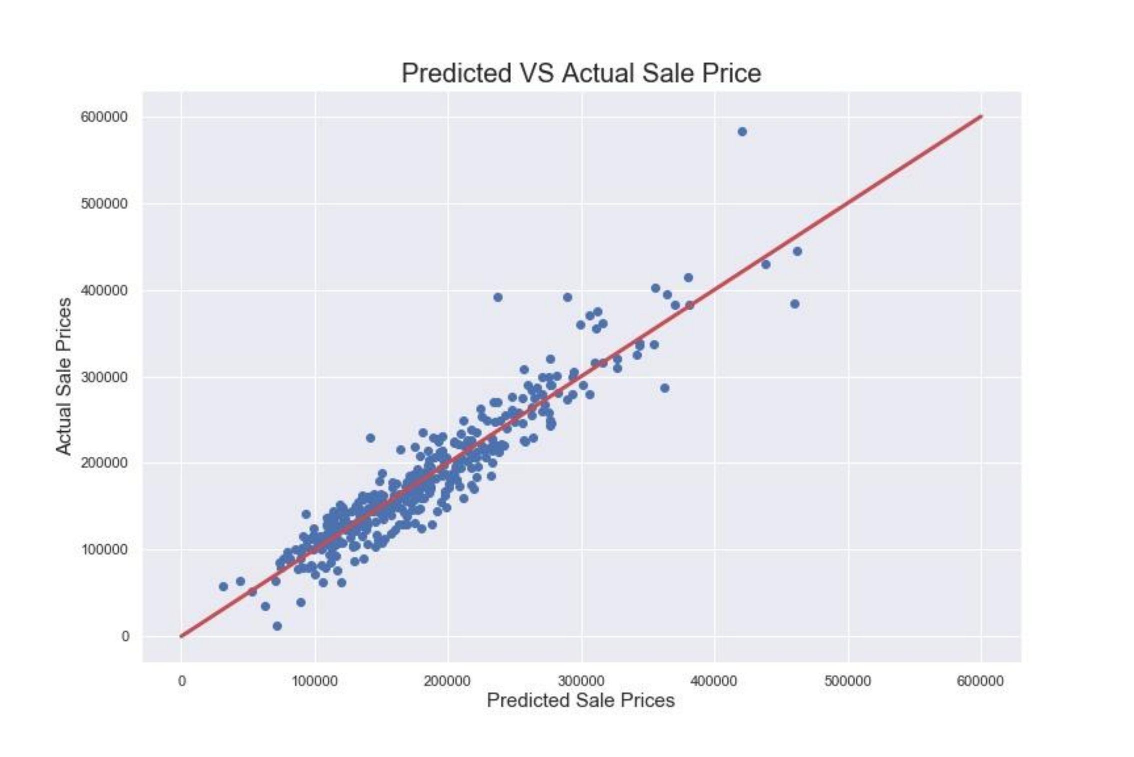
- Based on the dataset given, are we able to build a robust model which will help us predict/estimate the value (Sale Price) of a house in Ames, Iowa?

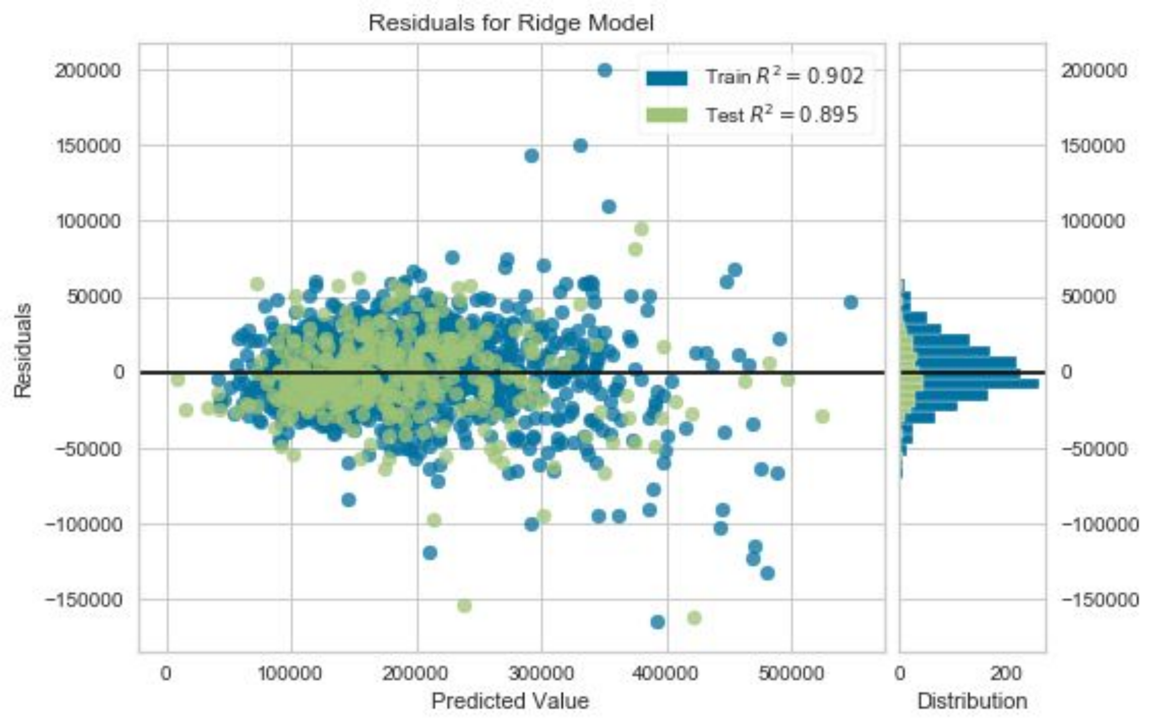
- From the final model, what are we able to infer about the major factors influencing the value of a house in Ames, Iowa?

- What other possible methods of getting a more robust prediction?

2. Executive Summary of Model

Checking of Assumptions:





From the graphs above, it seems that normality of residual errors assumption is not violated.

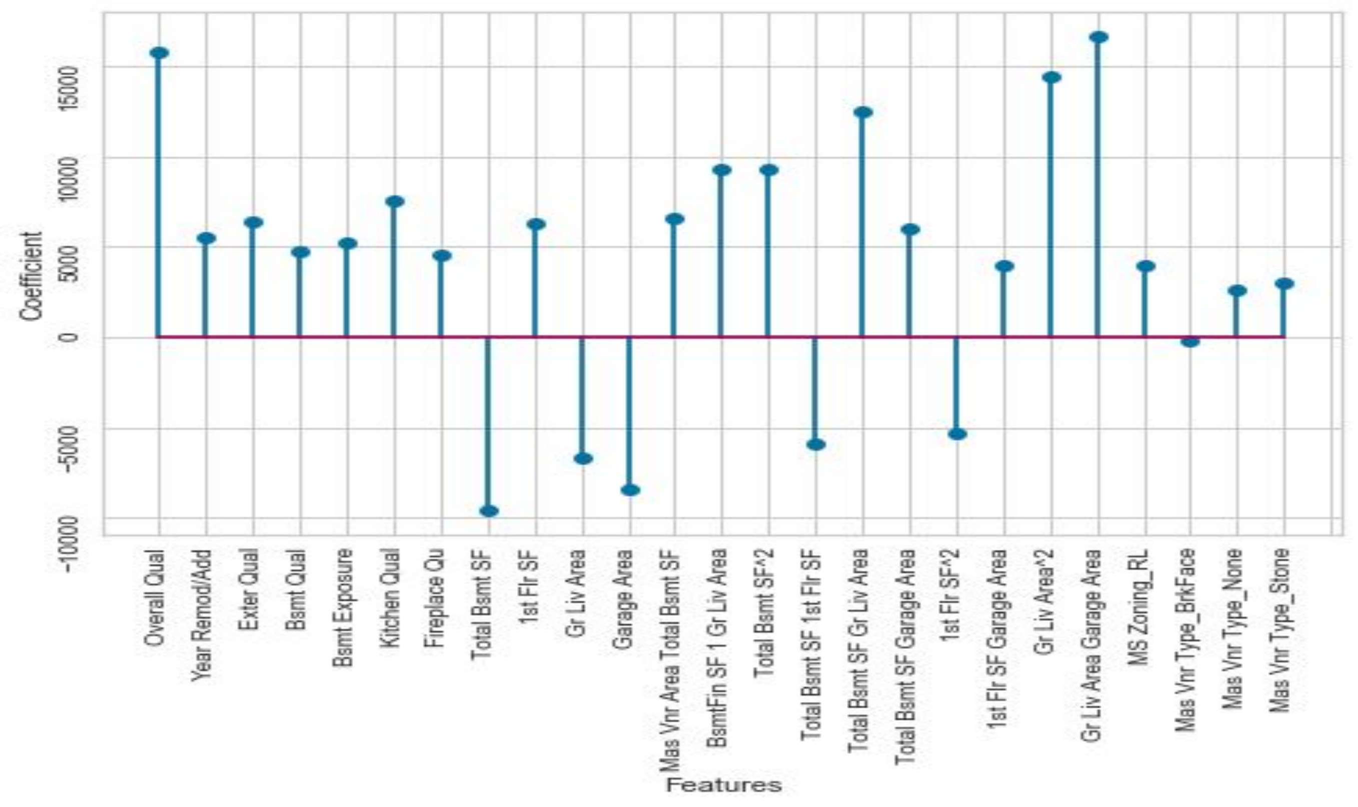
However, the variance of the errors seem to increase for higher Sale Prices, residual errors are not homoscadastic

Perhaps a higher level model like Neural Networks & Random Forests would be able to handle the change in variance at the higher Sale Prices better. Or maybe introduce even more complex features (i.e. polynomial features to higher powers)

The regression model seem to be decent (for houses below approx. $250,000-$300,000) but I believe not to be the best possible.

3. Possible Inferences

Coefficient Plot:



3. Possible Inferences

It will be very difficult to explain the higher polynomial features, as such, we will only be focusing on the first order features for the purpose of this project

1. The feature with the greatest impact on the value of the house is Overall Quality.

2. In general, it seems that a huge area (*Total Bsmt Area, Gr Liv Area, Garage Area*) for the house does not guarantee a high Sale Price.

   In fact, the regression model itself seem to suggest otherwise in some cases

3. The model seem to suggest that having a higher quality(*Overall, Exter, Bsmt, Fireplace, Kitchen Qual*) in general will have a positive effect on the Sale Price. A logical relationship.

4. A later year of remodification/addition (*Year Remod/Add*) seemed to have a positive impact on the Sale Price. A logical relationship (a newly fitted/renovated house would be worth more)

Once again, bear in mind that the increase in power of the features has made the inference portion highly complex and hence much harder to explain. To have a better explanation of things, we can do a regression model without so many unexplained polynomial features.

4. Recommendations

A point which I find interesting from this model is that in general, the ***area of the house seemed to have an inverse impact on its value***.

Further studies can be done on the repercussions of a house having a huge area. Perhaps after exceeding a certain threshold area, house maintenance would be increasingly difficult, resulting in a reduction of house quality.

Another interesting study could be on the trade-off between having a huge area and the general quality maintained.