Surprise Housing Assignment

1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented

In case of Ridge, we are having similar results across different alpha values and in case of Lasso, our optimal alpha value is 0.01.

If we double the value of the alpha, in case of Ridge there is not change in terms of accuracy or coefficients and in case of Lasso, we don't see any significant change in R2 and some minor change in coefficients.

The most important predictor features after the change from Lasso Regression is are as follows:-

-51.4623330053
Full bathrooms above grade

-2933.7174537111
Remodel date

377.8403524516
Masonry veneer area in square feet

77.7153991444
Fireplaces

45092.9320448458
Basement full bathrooms

21580.6780969653
Bedrooms above grade

16564.3025726093
Second floor square feet

11.102186201
Kitchen

2. You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

From the tested lambda values for Ridge and Lasso, I will choose 0.01 to apply for Lasso as it is giving us a model with better accuracy as compared to other and also the model coefficients shows significant impact on dependent or target variable. As, in case of Ridge I'm seeing similar impact across all the models and coefficients are quite reduced towards 0.

3. After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

The most important 5 predictor variables as per the problem statement and business understanding are:-

- House Style
- Garage Type
- Central Air Condition
- Lot Size
- Sale Type

4. How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

To make sure the model is robust and generalizable, we make sure that our model follows the following assumptions: -

- There is a linear relationship between dependent and independent features.

- The error term is normally distributed with mean equal to 0 (not X, Y)If you just want to fit the line and no further interpretation is required, it is okay if the error term is not normally distributed.
  However, if we want to make inferences about the model we have built, we need to know the distribution of the error term. A particular consequence of non-normally distributed error terms is that the p-values obtained during hypothesis testing to determine the significance of coefficients become unreliable. The normality assumption is made because it has been observed that the error term is usually normally distributed with a mean of 0 in most cases, but it must have a mean of 0. The only necessary condition is that the residuals are normally distributed.

- Error terms are independent of each other, and error terms should not be dependent on each other (like time series data where the next value depends on the previous value).

- The error term has constant variance (homogeneous variances). The variance must not increase (or decrease) as the error value changes. Also, the variance must not follow a pattern in which the error term changes.