Problem Statement- Part II

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

The optimal values of alpha for ridge and lasso regression were determined through cross-validation, resulting in the following values:

- Ridge: 1.5
- Lasso: 200

After doubling the alpha values for both Ridge and Lasso to 3 and 400, respectively, the R2 score started to decline. In the case of Lasso, the R2 score decreased by approximately 1%.

The most important predictor variables after implementing the change are as follows:

- House having a Second Floor
- Overall Condition of the House
- Bedroom Above Grade
- Type 1 finished square feet
- First Floor square feet

It's important to note that the reduction in the R2 score and the changes in the top predictor variables indicate that increasing the regularization strength (by doubling alpha) had an impact on the model's performance and feature importance.

Question 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?

- We choose Lasso as it provides a feature selection option. Lasso has the capability to remove unnecessary features from the model without significantly affecting its accuracy.
- This helps in creating a simpler and more interpretable model, making it a preferable choice in situations where feature sparsity and interpretability are important considerations.

Question 3: After building the model, you realised 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?

Top 5 features are:

- House having Second Floor
- Overall Condition of the House
- Bedroom Above Grade
- Type 1 finished square feet
- First Floor square feet

After dropping them model accuracy reduced to 69%.

Post dropping the first-choice columns the next top 5 are:

- Quality of the material on the exterior
- Overall material and finish of the house
- Hot water or steam heat other than gas
- Kitchens above grade
- Building Age

Question 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 model robust and generalizable 3 features are required:

1. Model accuracy should be > 70-75%:

2. P-value of all the features is < 0.05

3. VIF of all the features are < 5

It is also important to consider the values obtained for train and test, so that the model will perform well on unseen data. This means that the data should retain some outliers to help with predictions. As demonstrated in the assignment, accuracy of the model will vary, depending on the way data is processed and how features are selected. There may be no perfect model, but different steps are available to ensure that the model developed is fit for purpose for the specific context and the uniqueness of the business case. This is in line with Occam’s razor, that is, the model to be chosen should not be more complex than it needs to be.