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?

Optimal value of alpha for

- Ridge: 1.5
- Lasso: 200

After doubling alpha for Ridge and Lasso to 400 and 3 respectively the R2 Score starting going down and in Lasso's case it went down by around 1%.

New Top Features post doubling alphas:

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

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 its giving feature selection option also. It has removed unwanted features from model without affecting the model accuracy and making the model simple.

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
- Membrane Roof Material
- Bedroom Above Grade
- First Floor square feet

After dropping them model accuracy reduced from 80 and 80% to 74% and 72%.

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

- Masonry veneer area in square feet
- Wood Shakes Roof material
- Overall material and finish of the house
- Kitchens
- Lot size in square feet

Question 4: How can you make sure that a model is robust and generalisable? 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.