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?

Answer:

Optimal value of alpha for Ridge: **20**

Optimal value of alpha for Lasso: **0.001**

After doubling the value for alpha:

- ✓ For both Ridge and Lasso the **r2 score dropped** and **"mse" increased** but at the same time it seems to **correct the overfitting issue** with **Lasso** model
- ✓ Based on **Ridge model**, **Basement Full Bathrooms now becomes a significant variable** instead of Overall Condition type "Poor"
- ✓ There are changes in coefficients values and in their significance order

Top 5 features are:

- Overall Quality type "Very Excellent"
- Neighbourhood type "Old Town"
- Garage Area
- Basement Full Bathrooms
- Neighbourhood type "Northridge Heights"

Please refer the notebook for solution.

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?

Answer:

I chose to apply Ridge because r2 score on training set is ~.89 and on test set it is ~.86 which shows a generalize model compared to Lasso which has r2 score of ~.92 on training set and ~.83 on test set which shows overfitting. Hence, I chose Ridge even though it has slightly less r2 score on training set compared to Lasso.

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?

Answer:

Top five features now are:

- Fence type "Good Wood"
- Basement Half Bathrooms
- Neighbourhood type "Old Town"
- Neighbourhood type "Edwards"
- Sale type "Other"

Please refer the notebook for solution.

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?

Answer:

One of the ways to make sure that model is robust and generalisable is by looking at r2 score for training and test sets. A robust and generalisable model will perform with similar accuracy on both training (seen) dataset and test (unseen) dataset and the r2 score for such model will be similar. Such model will not be sensitive to changes in training set.

One of the Implications on accuracy is that if model is not robust and generalisable then accuracy on test data set will drop significantly compared to the accuracy on training data set. The reason would be that model is overfitting in this case and have remembered the training data set hence accuracy is high on training data set but on test data set, which is unseen, model will fail to perform with the same accuracy as it has remembered the data set instead of capturing the pattern in data set.