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 1

Optimal values for Ridge: 50

R2 score for train and test are: 90.14 and 90.04

Optimal values for Lasso: 0.01

R2 score for train and test are: 87.04 and 88.94

If doubled,

Ridge: R2 score for train and test are: 89.21 and 89.91

Lasso: R2 score for train and test are: 85.77 and 88.34 => the number of selected features drop from 38 to 28

Updated Important Variables when doubled

Ridge		**Lasso**	
Lambda = 50	**Lambda = 100**	**Lambda = 0.01**	**Lambda = 0.02**
Overall Quality	Overall Quality	Overall Quality	Overall Quality
Above ground living area sq ft	Overall Condition	Above ground living area sq ft	Above ground living area sq ft
Overall Condition	Above ground living area sq ft	Garage Car Capacity	Age of the Property
Garage Car Capacity	Garage Car Capacity	Age of the Property	Garage Car Capacity
If Neighbourhood is Northridge	Condition 1 being normal	Overall Condition	Overall Condition

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 2:

Even though the Ridge model's R2 score is higher than Lasso model we will suggest the Lasso model since it eliminates more than 70% of the variables and thereby creating a simpler model. As per Occam's Razor we always go with the simpler model. Therefore, the company can focus on just 39 features, rather than over 250 variables as provided by the ridge model.

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 3

If the top 5 contributing features are not available, we simply cannot choose the next 5. We will have to build a new model without them and then retrieve the top 5 contributing features. After deploying the model we find them to be:

1. 1st Floor Sq Ft
2. Presence of Central Air Conditioning
3. 2nd Floor Sq Ft
4. Garage Area
5. Years Since Remodel

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 4

A model is robust when it isn't greatly affected by variations in the data. A generalized model can adapt well to new, unseen data. To ensure both robustness and generalizability, we must avoid overfitting. This is because an overfit model, with high variance, is sensitive to small data changes, struggling with unseen test data. So, the model shouldn't be overly complex for robustness and generalizability ensuring appropriate bias and variance.

Considering accuracy, an overly complex model may show high accuracy. To enhance robustness and generalizability, we must reduce variance, introducing some bias, which reduces accuracy. Striking a balance between accuracy and complexity is crucial, often achieved through techniques like Ridge Regression and Lasso as seen in this assignment.