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

Optimum value for alpha ridge is 20 and for lasso is 0.0001.
Doubling the value of alpha in both Ridge and Lasso regression will strengthen the re gularization, leading to shrinkage of the coefficients, causing the coefficients to be pus hed more toward zero and causing the problem of underfitting.
Important predictor would be:

- GrLivArea: Above grade (ground) living area square feet,
- OverallQual: Rating as very excellent(10) and excellent(9) for the overall mater ial and finish of the house as very excellent and excellent,
- RoofMatl_WdShngl: Roof marterial as wood shingles,
- FullBath_3: full bathroom as 3,
- LotArea: Lot area in square feet,
- Condition2_PosN: Proximity to various condition as PosN,
- Fireplaces_3 : Number of fireplaces as 3,
- KitchenQual_TA: Kitchen Quality as TA,
- TotRmsAbvGrd_11: Total rooms above ground.

Predictor variables GrLivArea, OverallQual_10, RoofMatl_WdShngl impacts housing price positively while predictor variables Condition2_PosN, Fireplaces_3, KitchenQu al_TA, otRmsAbvGrd_11 impacts housing price negatively.

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

I would choose lambda for lasso regression because with optimum value of lambda la sso regression's R square value is higher than ridge and also since all the columns wer e taken into consideration (grid search cross validation is not performed) lasso regressi

on makes many non-important predictor variables coefficients zero and gives only im portant predictor variables which impacts house price.

Question 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?

Answer 3

Five most important predictor variables are

- GrLivArea: Above grade (ground) living area square feet,
- OverallQual: Rating as very excellent (10) and excellent (9) for the overall mat erial and finish of the house as very excellent and excellent,
- RoofMatl: Roof marterial as wood shingles,
- Condition2: Proximity to various condition as PosN,
- Fireplaces: Number of fireplaces as 3.

Predictor variables GrLivArea, OverallQual_10, RoofMatl_WdShngl impacts housing price positively while predictor variables Condition2_PosN, Fireplaces_3 impacts housing price negatively .

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?

Answer 4

To ensure model is robust and generalizable we can perform below strategies:

1) Feature engineering: Choose relevant features and transform them appropriatel y.
2) Regularization Techniques: Regularization methods (e.g., Ridge, Lasso, Elastic Net) helps prevent overfitting by adding penalty terms to the regression coeffici ents. It encourages the model to generalize well to new, unseen data.

3) Cross validation: Perform cross-validation (e.g., k-fold cross-validation) to asse ss the model's performance on different subsets of the data.
4) Outlier handling: Outliers can disproportionately influence the model, leading t o poor generalization.
5) Data Scaling: Scaling features ensures that the coefficients are on a similar scal e, preventing the model from giving too much weight to features with larger ma gnitudes.
6) Hyperparameter tuning: Optimize hyperparameters using techniques like grid s earch helps find the best configuration for improved generalization.

Implication for Accuracy: Implementing these strategies enhances the model's ability t o generalize well to new, unseen data and prevents overfitting.