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 alpha:

Optimal value of alpha for ridge and lasso:

Optimal alpha for Ridge: 4.0

Optimal alpha for Lasso: 100

Effect of doubling alpha:

Doubling the alpha value increases the regularization strength, which typically reduces the magnitude of the coefficients, making the model simpler and potentially less prone to overfitting.

Most important predictor variables:

Most important variables after doubling alpha for Ridge:

OverallQual	54290.650332
GrLivArea	43647.780360
2ndFlrSF	43169.109739
TotRmsAbvGrd	36058.232078
Neighborhood_NoRidge	35183.266540
Neighborhood_StoneBr	34020.779712
GarageCars	32228.499555

1stFlrSF 31705.653374

FullBath 29369.293058

Neighborhood_NridgHt 24634.891544

Most important variables after doubling alpha for Lasso:

GrLivArea 172398.097251

OverallQual 100772.757649

GarageCars 46381.228461

Neighborhood_NoRidge 40344.906036

Neighborhood_StoneBr 37426.596004

Neighborhood_NridgHt 31060.237171

TotRmsAbvGrd 24429.253459

BsmtFullBath 23555.913715

Fireplaces 21298.538641

BsmtExposure_Gd 20547.874211

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:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.75	0.90	0.89
1	R2 Score (Test)	0.62	0.87	0.88
2	RSS (Train)	1736000291646.84	729049310188.63	763423313686.70
3	RSS (Test)	856453961018.30	292067677438.38	272954682056.98
4	MSE (Train)	38552.58	24983.72	25565.91
5	MSE (Test)	54157.75	31626.44	30574.11

Analysis:

Both Ridge and Lasso have significantly higher R2 scores on the training set compared to Linear Regression, indicating better fit during training.

- Both Ridge and Lasso show a significant improvement in the R2 score on the test set compared to Linear Regression, with Lasso slightly outperforming Ridge
- Ridge has the lowest RSS on the training set, closely followed by Lasso, indicating a better fit with less residual sum of squares.
- Lasso has the lowest RSS on the test set, indicating it has the smallest residual errors.
- Ridge has the lowest MSE on the training set, indicating better predictive accuracy during training.
- Lasso has the lowest MSE on the test set, indicating better predictive accuracy on unseen data.

Given the metrics, Lasso Regression appears to be the better model for the following reasons:

- It has the highest R2 score on the test set, indicating better explanatory power on unseen data.
- It has the lowest RSS and MSE on the test set, indicating lower prediction errors and better generalization performance.

Therefore, I would choose to apply Lasso Regression because it provides the best balance of high explanatory power (R2 Score) and low prediction error (MSE and RSS) on the test set, making it more robust and reliable for new, unseen data.

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 5 most important features in Lasso: 'GrLivArea', 'OverallQual', 'GarageCars', 'Neighborhood_StoneBr''Neighborhood_NoRidge'

New top 5 most important features in Lasso after excluding the original top 5: '1stFlrSF', '2ndFlrSF', 'GarageArea', 'RoofMatl_WdShngl', 'Exterior2nd_ImStucc'

Note: Please refer the assignment section in ipynb file for the code.

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:

To ensure that a model is robust and generalizable, We can follow several best practices in the modeling process. Robustness means that the model performs well under different conditions and is not overly sensitive to variations in the data. Generalizability refers to the model's ability to perform well on unseen data, beyond the training dataset.

Best Practices to Ensure Robustness and Generalizability:

Cross-Validation:

Use techniques like k-fold cross-validation to assess the model's performance across different subsets of the data. This helps ensure that the model is not overly dependent on a particular portion of the data.

Implication: Cross-validation helps provide a more reliable estimate of model performance and reduces the risk of overfitting.

Regularization:

Apply regularization techniques like Ridge or Lasso regression to prevent overfitting by penalizing large coefficients.

Implication: Regularization reduces the model complexity, which helps in improving the generalization of the model to new data.

Feature Selection:

Select relevant features that contribute to the model's predictive power. Remove irrelevant or redundant features to avoid overfitting.

Implication: Proper feature selection improves the model's interpretability and performance on unseen data.

Hyperparameter Tuning:

Perform hyperparameter tuning using cross-validation to find the best set of parameters for the model. This ensures the model is well-calibrated and performs optimally.

Implication: Proper hyperparameter tuning helps in finding a balance between bias and variance, leading to better generalization.