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:

Lasso Regression:

1. The optimal value of alpha with Lasso regression is 0.001
2. When the value of alpha is doubled, we see following changes:
   a) The R2 score reduced from 0.82 to 0.78.
   b) The number of predictor features has reduced from 24 to 22.
   c) Although the top 5 predictor variables did not change, their coefficients changed and hence the order of prefered features changed

alpha = 0.001		alpha = 0.002	
Top 5 predictor features	**Coefficient**	**Top 5 predictor features**	**Coefficient**
GrLivArea	0.239808	OverallQual	0.173505
OverallQual	0.169949	GrLivArea	0.115436
Neighborhood_NoRidge	0.062858	KitchenQual	-0.057848
GarageCars	0.058208	GarageCars	0.05752
KitchenQual	0.049976	Neighborhood_NoRidge	0.047273

Ridge Regression:

1. The optimal value of alpha with Ridge Regression is 10.0
2. When the value of alpha is doubled, we see following changes:
   a. The R2 score reduced from 0.86 to 0.84.
   b. No of predictor variables remains the same i.e. 183
   c. The coefficents of the top 5 predictor variables have changed. BsmtExposure is now 5th most where as 1stFlrSF is now not featuring in top 5

alpha = 10.0		alpha = 20.0	
Top 5 predictor features	**Coefficient**	**Top 5 predictor features**	**Coefficient**
OverallQual	0.082357	OverallQual	0.065683
GrLivArea	0.065476	Neighborhood_NoRidge	0.052820
2ndFlrSF	0.062813	GrLivArea	0.050848
Neighborhood	0.060228	2ndFlrSF	0.046252
1stFlrSF	0.049093	BsmtExposure	-0.045109

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:

Lasso or Ridge - which model to select:

In model building we captured the stats around accuracy and error. Same are mentioned below:

	Lasso Regression		Ridge Regression	
	Training Data	**Test Data**	**Training Data**	**Test Data**
R2 Score	0.82	0.81	0.86	0.85
Residual Sum of Squares (rss)	2.19	1.05	1.7	0.83
Mean Squared Error (mse)	0.002	0.002	0.001	0.002
Root Mean Squared Error (rme)	0.05	0.05	0.04	0.04
No features with >0 coefficient	24		183	

Based on the inferences captures, Ridge regression seems to be a better choice compared to Lasso Regression. Here's why:

1. **Higher R2 Score on Test Data:** Ridge regression has a higher R2 score on both the train and test sets compared to Lasso regression. A higher R2 score indicates that the model explains more variance in the target variable and performs better in generalization to unseen data.
2. **Better Error Metrics on Test Data**: Ridge regression has lower error metrics (RSS, MSE, RMSE) on the test data compared to Lasso regression. Lower error metrics indicate that the model's predictions are closer to the actual values, demonstrating better performance.
3. **Retained Features:** While Lasso regression reduced the number of dependent features from 185 to 24, Ridge regression retained all 183 features. In some cases, reducing the number of features can lead to oversimplification and loss of important information. Therefore, retaining more features in Ridge regression may provide a more comprehensive representation of the data.
4. **Similar Assumptions:** Both models satisfy the assumptions of linear regression, such as a linear relationship between predictors and the target variable, normal distribution of error terms with a mean of 0, independence of error terms, and constant variance of error terms.

Considering these factors, Ridge regression appears to be a better choice as it provides higher predictive accuracy, retains more features, and exhibits similar adherence to regression assumptions.

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:

The top 5 predictor variables in lasso regression are GrLivArea, OverallQual, Neighborhood, GarageCars & KitchenQual. After dropping these variables from the base X_train and X_test data and build the model again.

Top 5 predictor feature as per new model are:

1. 1stFlrSF - 1stFlrSF: First Floor square feet
2. 2ndFlrSF - Second floor square feet
3. ExterQual - Evaluates the quality of the material on the exterior
4. GarageArea - Size of garage in square feet
5. Fireplaces - Number of fireplaces

Few important observations from the new model are as below:

- The model accuracy as well as robustness has reduced from model 3.
    - R-squared on the training set: 0.768
    - R-squared on the test set: 0.769
- However the test accuracy is at par with training set. Hence we can say that model is still robust

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:

Ensuring that a model is robust and generalizable is crucial for its effectiveness in real-world applications. Here are some strategies to achieve robustness and generalizability:

1. Cross-Validation: Use techniques like k-fold cross-validation to assess the model's performance on multiple subsets of the data. This helps in evaluating how well the model generalizes to unseen data and reduces the risk of overfitting.
2. Regularization: Regularization techniques like Ridge and Lasso regression can prevent overfitting by penalizing large coefficients. These techniques help in building simpler models that generalize better to new data.
3. Feature Selection: Selecting relevant features and eliminating irrelevant or redundant ones can improve model robustness. It reduces the risk of overfitting and focuses the model on the most important predictors.
4. Hyperparameter Tuning: Optimize model hyperparameters using techniques like grid search or randomized search. Tuning hyperparameters helps in finding the optimal balance between bias and variance, leading to a more robust model.

The implications of building a robust and generalizable model for its accuracy are significant. While it's essential to strive for high accuracy on the training data, the ultimate goal is to build a model that performs well on unseen data. A highly accurate model on the training data may indicate overfitting, where the model learns noise or patterns specific to the training set but fails to generalize to new data. On the other hand, a robust and generalizable model may sacrifice some accuracy on the training data to achieve better performance on unseen data, leading to more reliable predictions in real-world scenarios. Therefore, prioritizing robustness and generalizability over training accuracy is crucial for building models that are truly effective and trustworthy.