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:

The Optimal Value of alpha varies based on the dataset and problem being solved. Alpha is the hyperparameter that controls the amount of regularization to be applied to the model. Best value of alpha can be determined using techniques like gridsearch or cross validation.

If you choose to double the value of alpha, the model would become more regularized with coefficients shrinking towards zero. In Ridge regression, the coefficients values would shrink and becomes very small overall. In Lasso regression, it will lead to more coefficients being set to zero, reducing the predictor variables in the model.

In Lasso regression, the most important predictor variables are those that survive after the feature selection process. In ridge, while the coefficients may not be exactly zero , those with smaller absolute values will have less importance and others would have higher importance.

This has been answered in the notebook too with regards to the dataset.

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:

After determining the optimal value for ridge and lasso, the choice between them would depend on various factors:

If your dataset has many features, some of which may be correlated, and you're primarily interested in prediction accuracy while still retaining some interpretability, ridge regression might be the better choice. Ridge regression tends to perform well when there are many small to medium-sized effects and can handle multicollinearity effectively without sacrificing too much predictive power. Additionally, if computational efficiency is a concern and you prefer a model with non-zero coefficients for all features, ridge regression would be preferable, as it doesn't produce sparse solutions.

On the other hand, if you suspect that many of your features are irrelevant or redundant and you prioritize model interpretability or sparsity, lasso regression might be more appropriate. Lasso regression automatically performs feature selection by setting some coefficients to zero, resulting in a sparse model. This can be advantageous if you want to identify the most important predictors or if computational efficiency is crucial for your application.

Ultimately, the choice between ridge and lasso regression depends on your specific requirements, including the trade-off between prediction accuracy, interpretability, sparsity, and computational efficiency.

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:

We dropped the top 5 most important predictor variables in the lasso model and again created again model and got the below five most important predictor variables:

1. TotalBsmtSF
2. TotRmsAbvGrd
3. OverallCond
4. Total_Bathrooms
5. LotArea

```
TotalBsmtSF         0.325641
TotRmsAbvGrd        0.126619
OverallCond         0.093923
Total_Bathrooms     0.086192
LotArea             0.067803
```

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:

The model aims for simplicity, as simpler models are perceived as more 'generic', albeit sacrificing some accuracy for increased robustness.

- Understanding the Bias-Variance trade-off is key: simpler models exhibit higher bias but lower variance, enhancing generalizability, while complex models display higher variance and lower bias.
- Underfitting and overfitting are common model challenges, underscoring the need for balance in Bias and Variance, achievable through "Regularization".
- Regularization manages model complexity by shrinking coefficients towards zero, curbing over-complexity and reducing the risk of overfitting.
- It's essential to use Regularization to maintain an optimum level of simplicity in the model, penalizing unnecessary complexity.
- Regularization facilitates achieving a balance in the Bias-Variance trade-off, compromising by increasing bias to a position where Total Error is minimized.
- This equilibrium point, termed Optimum Model Complexity, ensures the model is sufficiently simple to generalize effectively yet complex enough to remain robust.
- Simplifying the model entails navigating the Bias-Variance trade-off.