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 the regularization parameter, alpha, in Ridge and Lasso regression depends on the specific dataset and the goals of the modelling. In Ridge regression, alpha is a tuning parameter that controls the strength of the regularization, while in Lasso regression, it controls the amount of shrinkage applied to the coefficients.

Generally, the choice of alpha involves a trade-off between fitting the model well to the training data and keeping the model coefficients small to prevent overfitting. Cross-validation or other model selection techniques are often used to find the optimal alpha value for a given problem.

If you were to double the value of alpha in Ridge and Lasso regression, the regularization strength would increase. This would lead to stronger regularization, which, in turn, would result in more aggressive shrinkage of the coefficients. In both Ridge and Lasso, higher alpha values lead to more regularization and, consequently, simpler models with smaller coefficients.

The impact on the model would be a stronger penalty on the magnitude of the coefficients. In Ridge regression, this means that the sum of the squared coefficients would be more strongly penalized, and in Lasso regression, it means that some of the coefficients might be exactly set to zero, leading to feature selection.

As for the most important predictor variables after implementing the change, it's important to note that doubling the alpha value would generally result in a sparser model with more coefficients being close to zero. In Lasso regression, some coefficients may be exactly zero, effectively eliminating certain predictor variables from the model. In Ridge regression, the coefficients will still be shrunk towards zero, but none will be exactly zero.

The specific variables that remain important after the change would depend on the interplay of the features in your dataset and how they contribute to the model. Variables that have a stronger impact on the target variable and are less correlated with other variables are more likely to remain important even with higher regularization. It's advisable to perform feature importance analysis or examine the coefficients of the model to identify the most influential predictors after the regularization adjustment.

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:

It is crucial to apply regularization to coefficients to enhance prediction accuracy while reducing variance and maintaining model interpretability. Ridge regression employs the tuning parameter lambda, determined through cross-validation, to penalize the square of coefficient magnitudes. This penalty, lambda times the sum of squared coefficients, penalizes larger coefficient values. Increasing lambda decreases model variance while keeping bias constant. Unlike Lasso Regression, Ridge regression includes all variables in the final model.

Lasso regression also utilizes the lambda tuning parameter, penalizing the absolute value of coefficient magnitudes. Increasing lambda in Lasso shrinks coefficients towards zero, setting some variables exactly to 0, and performs variable selection. When lambda is small, Lasso behaves like simple linear regression, and with increasing lambda, shrinkage occurs, neglecting variables with a coefficient of 0.

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:

After constructing the model, it became apparent that the five crucial predictor variables identified by the Lasso model are absent in the new dataset. Consequently, a new model needs to be developed, this time excluding these five essential predictor variables. The predictor variables to be excluded are as follows:

1. GrLivArea

2. OverallQual

3. OverallCond

4. TotalBsmtSF

5. GarageArea

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 a model is robust and generalizable:

1. Use cross-validation and train-test splits to assess performance on different data subsets.

2. Consider a validation set for hyperparameter tuning.

3. Carefully choose and engineer features to capture essential patterns.

4. Implement regularization to prevent overfitting.

5. Strike a balance in model complexity to avoid overfitting.

Implications for Accuracy:

- Balance training and test accuracy to avoid overfitting.

- Ensure consistency across different datasets.

- Mitigate overfitting through regularization and cross-validation.

- Aim for a model that performs well on diverse datasets, promoting accuracy and generalization.