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 value:

Ridge: 4.0

Lasso: 100

If we choose double the value of alpha for both ridge and lasso it will increase regularization strength and hence magnitude of coefficients will reduce which simplify the model and chances of overfitting will be less.

Most Important variables after doubling alpha for Ridge:

OverallQual 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:

Ridge and Lasso have higher R2 scores on the training set compared to Linear Regression. It indicates better fit during training.

Below is the Analysis:

- Lasso slightly outperforming in R2 score of test set.
- Lower the residual sum squares better the model, ridge has lowest RSS on training set.
- 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.

Considering above factor, Lasso Regression appears to be the better model because

It has the highest R2 score on the test set, indicating better explanatory power on unseen data.

And also it has the lowest RSS and MSE on the test set, indicating lower prediction errors and better generalization performance.

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'

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

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 make sure that a model is robust and generalizable, we can follow best practices that will ensure robustness and generalization.

In case of robustness model performs well irrespective of conditions and is not very much sensitive to variations of data. Generalizability performs well on unseen data which is not part of training set.

Some of the best practices are:

Regularization: Regularization help to reduce the model complexity, which improves generalization of the model to new data.

Cross-Validation: Cross-validation reduces the risk of overfitting.

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

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