

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:

For both the models, optimal alpha value got was 500.

While doubling, R2 value of training data was reduced slightly while the test data R2 value increased to some extent. In Ridge, the was more pronounced than in lasso.

Most important predictor variables remain the same.

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 is better compared to lasso model built.

Because the test data predictability (R-squared value) in ridge model is better than lasso model. Training data R-squared is more or less the same in both the cases

	Ridge	Lasso
R2 Train	0.9326114641079967	0.9436243125166998
R2 Test	0.6858512805908459	0.5349333134386924

Question 3:

After building the model, you realized 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?

The top five attributes are:

GrLivArea

OverallQual

TotalBsmtSF

YearBuilt

BsmtFinSF1

New model not evaluated.

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?

The model can be made more robust and generalizable by reducing the overfitting of training data which results in high variance, while at the same time optimizing (reducing) the square error. i.e. lowering the bias.

This is done in the Lasso and Ridge models by introducing hyperparameters which will limit the coefficients to optimal value depending on hyperparameter tuning. The parameter λ - λ^2 / λ will ensure this in Ridge and Lasso regression.

Implications:

The accuracy of prediction would be slightly reduced. Actually, it becomes optimal. It lies close to the intersection of bias and variance.

Also, during EDA, any outlier data has to be removed from the data which will help in better predictability and the model will be more robust.