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 for Ridge regression: 400

Optimal alpha for Lasso regression: 0.04

If the value of alpha is doubled for ridge and lasso then the model will regularize more, which means that the insignificant features will be penalized further. The optimal value of alpha has been chosen based on consideration of minimum RMSE on train data, thus with a change in the alpha the RMSE will increase and the R-squared value data will reduce – thus deteriorating the model performance. The table below depicts the change in the R-Squared value on doubling the alpha for the Ridge and Lasso

Alpha	**R-Squared on test data**	
	Ridge	Lasso
Optimal	0.8265	0.8197
Double of Optimal	0.8135	0.8017

After doubling the alpha value, the top 6 most important predictor variables with their coefficients are as below:

Ridge regression (with alpha 800):

Feature	Coefficient
OverallQual	0.122795
GrLivArea	0.111260
TotRmsAbvGrd	0.083564
GarageCars	0.082483
ExterQual	0.077165
KitchenQual	0.074888

Lasso regression (with alpha 0.08)

Feature	Coefficient
OverallQual	0.280572
GrLivArea	0.250756
GarageCars	0.119492
KitchenQual	0.080568
ExterQual	0.077978
TotalBsmtSF	0.052535

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:

For the base model without regularization, we have an R-squared value of **0.84** on the training data as per the stats model summary.

Based on the evaluation of the regularization models for the optimal value of lambda, we have the following:

Ridge:

R-squared on test data: 0.8265

RMSE: 0.405

Lasso:

R-squared on test data: 0.8197

RMSE: 0.413

Thus, I will choose to apply the Ridge regularization model over the Lasso model, because the Ridge model provides a better R-squared value (closer to the training R-squared value) and also has a slightly lower root mean squared error.

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 eliminating the five most important predictor variables from the Lasso model and then rebuilding it, the new five most important predictor variables are as follows:

Feature	Coefficient
TotRmsAbvGrd	0.227860
TotalBsmtSF	0.159949
Fireplaces	0.123012
BsmtQual	0.113267
MasVnrArea	0.109640

The R-squared value for the test data reduced from 0.8197 to 0.7355.

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?

Answers:

In order to make sure that the model is robust and generalisable it is important not to over complicate the model. As part of the model building, it is important to achieve the right balance between bias and variance. The model should be complex enough that it optimizes the bias and it should be simple enough that it optimizes variance across different datasets. A too simple model may cause underfitting while a too complex model may lead to overfit to the train data. A generalized model is one that learns the pattern in the training data and is able to perform well on unseen data.

Cross is a technique that can be used to assess the performance of the model on various sets/folds of the data considered as train and test alternatively.

While trying to generalize a model it may cause a reduction in its accuracy. This is caused because the regularization techniques that are used for the process of generalizing employ an approach of penalizing the less significant predictor variables – this may cause a reduction in the contribution of those variables towards the prediction thus reducing the accuracy to some extent. However, this is a fair trade off as long as the impact on the accuracy is small (<5%).