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 value of alpha for Ridge regression is 9.

Optimal value of alpha for Lasso regression is 0.001.

After doubling the value of alpha below are the changes in R2 score

- Ridge Regression R2 score for training set decreased from 0.939 to 0.934
- Ridge Regression R2 score for test set decreased from 0.912 to 0.911
- Ridge Regression RMSE for training set increased from 0.08 to 0.09
- Ridge Regression RMSE for test set doesn't change (0.10)
- Lasso Regression R2 score for training set decreased from 0.92 to 0.90
- Lasso Regression R2 score for test set decreased from 0.91 to 0.89
- Lasso Regression RMSE for training set increased from 0.09 to 0.10
- Lasso Regression RMSE for test set increased from 0.10 to 0.11

Below are the most important predictor variables after the change is implemented:

As per Ridge regression:

GrLivArea

Neighborhood_Crawfor

OverallQual_8

Functional_Typ

OverallQual_9

OverallCond_9

Exterior1st_BrkFace

TotalBsmtSF

SaleCondition_Normal

OverallCond_7

As per Lasso regression:

GrLivArea

Neighborhood_Crawfor

OverallQual_8

Functional_Typ

OverallQual_9

TotalBsmtSF

YearRemodAdd

FireplaceQu_Gd

OverallCond_7

BsmtFinSF1

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:

Model selection will depend on the requirement.

- If the requirement is to build a model using all the features and not to drop any of the feature, then we will use Ridge regression.
- If the requirement is to build a model using feature selection, then we will go for Lasso regression.

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?

Answer:

After removing the 5 most important predictor variables from the Lasso model and then building the new model, below are the 5 most important predictor variables now:

Condition2_PosA

OverallCond_9

SaleType_Oth

SaleCondition_Alloca

SaleCondition_AdjLand

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:

For a model to be robust and generalizable it should perform well on unseen data. Any variation in the data should not affect the model. We should train the model on diverse datasets so that the model learns the underlying patterns. The model should not overfit or underfit. Overfitting makes the model complex resulting in memorizing the training data and model will not be able to predict correctly on unseen data. We should use cross validation to assess the performance of model on multiple subsets of training data, this increases the generalizability. Techniques like regularization should be used to prevent the model from becoming complex. We should experiment with different hyperparameters to find the setting that yields the best performance. And finally, we should evaluate our model against entirely unseen data to ensure that its performance remains consistent across various datasets.

A model which can achieve high accuracy on training data might not generalize well on unseen data. Balancing accuracy on training data and performance on test data is crucial. A model that overfits the training data has high accuracy in training data but low in test data. On the other hand, a model that underfits has low accuracy on both. Bias vs Variance tradeoff is very crucial. Too much complexity can lead to overfitting (high variance), while too much simplification results in underfitting (high bias).