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?

When we plot the curve between negative mean absolute error and alpha, we can see that as the value of alpha grows from 0, the error term decreases, and the train error shows a rising trend as the value of alpha increases. We opted to use a value of alpha equal to 2 for our ridge regression because the test error is the smallest when alpha is equal to 2.

For lasso regression, I chose a very tiny value of 0.01, because as the value of alpha increases, the model tries to punish more and make the majority of the coefficient values zero. It started out as 0.4 in terms of negative mean absolute error and alpha.

When we double the amount of alpha for our ridge regression and set it to 10, the model will apply more penalty on the curve and try to generalise the model, making it simpler and eliminating the need to fit every data point in the data set. We can observe from the graph that when alpha is ten, both test and train have higher error.

When we raise the value of alpha for lasso, we are attempting to punish our model more and more coefficients of the variable will be reduced to zero; similarly, when we increase the value of our r2 square, we are attempting to penalise our model more.

The most important variable after the changes has been implemented for ridge regression are as follows: -

1. MSZoning_RL
2. MSZoning_RH
3. MSZoning_FV
4. MSZoning_RM
5. Neighborhood_Crawfor
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. Exterior1st_BrkFace
10. OverallQual

The most important variable after the changes has been implemented for lasso regression are as follows: -

1. OverallQual
2. GrLivArea
3. GarageArea
4. TotalBsmtSF
5. BsmtFinSF1
6. MSZoning_RL
7. Fireplaces
8. LotArea
9. HalfBath
10. OverallCond

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?

Regularizing coefficients and improving prediction accuracy, as well as reducing variance and making the model interpretable, are critical.

Ridge regression employs a tuning parameter known as lambda as the penalty, which is the square of the coefficients' magnitude as determined by cross validation. Using the penalty, the residual sum or squares should be modest. The penalty is lambda times the sum of squares of the coefficients, therefore higher-valued coefficients are punished. As the value of lambda is increased, the variance in the model decreases while the bias remains constant. Unlike Lasso Regression, Ridge Regression incorporates all variables in the final model.

Lasso regression incorporates a tuning parameter known as lambda as the penalty, which is the absolute magnitude of the coefficients as determined by cross validation. Lasso decreases the coefficient towards zero as the lambda value grows, making the variables precisely equal to 0. Lasso can also choose variables. When the lambda value is small, the model does straightforward linear regression; as the lambda value increases, the model shrinks, and variables with 0 values are ignored.

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?

1. OverallQual
2. GrLivArea
3. GarageArea
4. TotalBsmtSF
5. BsmtFinSF1

4. How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

The model should be as simple as feasible, as this will reduce its accuracy but increase its robustness and generalizability. The Bias-Variance trade-off can also be used to understand it. The simpler the model, the greater the bias, but the lower the variance and the more generalizable it becomes. Its accuracy implication is that a robust and generalizable model will perform equally well on both training and test data, i.e., the accuracy will not differ significantly between training and test data.

Bias: Bias is a model mistake that occurs when the model is unable to learn from the data. When a model has a high bias, it is impossible to learn details from the data. On training and testing data, the model performs poorly.

Variance: Variance is a model error that occurs when the model tries to learn too much from the data. High variance indicates that the model performs extraordinarily well on training data since it has been extensively trained on this type of data, but poorly on testing data because it was previously unknown to the model.

To avoid overfitting and underfitting of data, it is critical to maintain a balance in Bias and Variance.