1. 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 regressions are two techniques used to regularize coefficients in order to improve the prediction accuracy and interpretability of a model. Ridge regression applies a penalty to the sum of squares of the coefficients, using a tuning parameter called lambda, which is determined through cross validation. This penalty helps to reduce variance in the model, while keeping the bias constant. All variables are included in the final model in Ridge regression. On the other hand, Lasso regression uses the absolute value of the coefficients as the penalty, also with a tuning parameter called lambda determined through cross validation. As the lambda value increases, Lasso shrinks the coefficients towards zero and may even set some coefficients exactly to zero, thereby performing variable selection. When the lambda value is small, Lasso performs simple linear regression, but as the value increases, shrinkage occurs and variables with zero values are excluded from the model. Both Ridge and Lasso regressions aim to regularize coefficients and improve the prediction accuracy of a model, but they differ in the type of penalty applied and the way they handle the inclusion or exclusion of variables.

2. 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. There are several predictor variables that have been identified as the most important ones to be excluded from the model. These include:

1. GrLivArea: This variable represents the above grade (ground) living area in square feet, and is often used to predict the sale price of a property.
2. OverallQual: This variable represents the overall material and finish quality of the property, and is typically considered a key factor in determining the sale price.
3. OverallCond: This variable represents the overall condition of the property, and is also often used in predicting the sale price.
4. TotalBsmtSF: This variable represents the total square footage of the basement in the property, which can be an important factor in determining the sale price.
5. GarageArea: This variable represents the total square footage of the garage in the property, which can also be a significant predictor of the sale price.

Excluding these variables from the model may impact the accuracy of the predictions, as they are considered to be important factors in determining the sale price of a property. It is important to carefully consider the impact of excluding these variables and weigh the trade-offs between model accuracy and interpretability.

3. 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?

Answer. Simplicity is important in a model because it can improve the model's robustness and generalizability. This is known as the bias-variance trade-off, where a simpler model has more bias but less variance, and is therefore more generalizable. In terms of accuracy, a model that is robust and generalizable will perform equally well on both training and test data, meaning that the accuracy does not change significantly between the two.

Bias refers to error in the model when it is weak at learning from the data. A model with high bias is unable to capture the details in the data, and will perform poorly on both training and test data. On the other hand, variance refers to error in the model when it tries to overlearn from the data. A model with high variance will perform exceptionally well on training data, but poorly on test data because it has not been exposed to this unseen data.

It is important to find a balance between bias and variance in order to avoid overfitting and underfitting of data. Overfitting occurs when a model is too complex and has learned too much from the training data, leading to poor performance on test data. Underfitting occurs when a model is too simple and has not learned enough from the training data, also leading to poor performance on test data. By finding the right balance between simplicity and complexity, a model can achieve good performance on both training and test data.

4. 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. In the process of implementing ridge regression, the negative mean absolute error was plotted against the alpha value. It was observed that as the alpha value increased from 0, the error term decreased and the train error showed an increasing trend. When the alpha value was 2, the test error reached a minimum, leading to the decision to use an alpha value of 2 for the ridge regression model. The most important variables after implementing these changes for ridge regression are:

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

10. Exterior1st_BrkFace

For lasso regression, a very small alpha value of 0.01 was chosen. As the alpha value was increased, the model attempted to apply more penalties and set more coefficient values to zero. Initially, the negative mean absolute error was 0.4 for this alpha value. When the alpha value for the ridge regression model was doubled to 10, the model applied a stronger penalty and tried to make the model more generalized by simplifying it and not trying to fit every data point in the dataset. From the graph, it was observed that when alpha is 10, both the test and train errors increase. The most important variables after implementing these changes for lasso regression are:

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