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: In the case of ridge regression: - When we plot the curve between negative mean absolute error and alpha we see that as the value of alpha increase from 0 the error term decrease and the train error is showing increasing trend when value of alpha increases. When the value of alpha is 2 the test error is minimum so we decided to go with value of alpha equal to 2 for our ridge regression.
For lasso regression used small value that is 0.01, on increasing the value of alpha the model try to make mostly the coefficient value 0. Earlier it was 0.4 in negative mean absolute error and alpha.
When we double the value of alpha for our ridge regression we take the value of alpha equal to 10 the model., will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set .from the graph we can see that when alpha is 10 we get more error for both test and train.
Similarly when we increase the value of alpha for lasso we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases.
The most important variable after the changes are implemented are :

1. MSZoning_FV - Floating Village Residential is the general zoning classification of the sale.
2. MSZoning_RL - Residential Low Density is the general zoning classification of the sale.
3. Neighborhood_Crawfor- Crawfor is the Physical locations within Ames city limits
4. MSZoning_RH - Residential High Density is the general zoning classification of the sale.,
5. MSZoning_RM - Residential Medium Density is the general zoning classification of the sale.,
6. SaleCondition_Partial - Condition of sale - Home was not completed when last assessed,
7. Neighborhood_StoneB -- StoneB is the Physical locations within Ames city limits ,
8. GrLivArea - Above grade (ground) living area square feet ,
9. SaleCondition_Normal- Condition of sale -Normal Sale,
10. Exterior1st_BrkFace - Exterior covering on house is Brick Face

Though the model performance by Ridge Regression was better in terms of R2 values of Train and Test, it is better to use Lasso, since it brings and assigns a zero value to insignificant features, enabling us to choose the predictive variables.

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: The r2_score of lasso is slightly higher than lasso for the test dataset so we will choose lasso regression to solve this problem. Ridge Regression R2 values of Train and Test given below respectively,model performance by Ridge Regression was better in terms of R2 values of Train and Test:
0.93645622972421
0.9134742831765208.
Using Lasso is better as it brings and assigns a zero value to insignificant features, enabling us to choose the predictive variables.

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and making the model interpretably.
Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.
Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

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: The 5 most important predictor variables

- OverallQual
- OverallCond
- GarageArea
- GrLivArea
- TotalBsmtSF

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:
The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much importance should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis.
