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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Increasing alpha value, the model will become underfitting it means model is generalized (model complexity is less). Low variance and high bias for model underfitting

When alpha value is increase it maximize the RMSE value

Ridge = 0.046544372008500826
Lasso = 0.05203395988143083

The most important predictor variables are as follow

[77]:

	Ridge Doubled Alpha Co-Efficient
OverallQual	0.235515
TotRmsAbvGrd	0.141937
LotArea	0.116859
MasVnrArea	0.105338
Total_Bathrooms	0.102107
LotFrontage	0.073532
GarageArea	0.067169
Fireplaces	0.059936
Total_porch_sf	0.036773
GarageCars	0.034363
OverallCond	0.027760
Heating_Floor	0.000000
Exterior2nd_AsphShn	0.000000
Exterior1st_AsphShn	0.000000
RoofMatl_Membran	0.000000
HeatingQC_Po	0.000000
MiscVal	-0.000083
MoSold	-0.000287
YrSold_Old	-0.001772
YearRemodAdd_Old	-0.024295

[78]:

	Lasso Doubled Alpha Co-Efficient
OverallQual	0.270385
Total_Bathrooms	0.086175
Fireplaces	0.066102
TotRmsAbvGrd	0.059513
GarageArea	0.055202
GarageCars	0.043245
MasVnrArea	0.042675
Total_porch_sf	0.004987
LotFrontage	0.000000
YrSold_Old	-0.000000
Heating_Floor	0.000000
Exterior2nd_AsphShn	0.000000
Exterior1st_AsphShn	0.000000
RoofMatl_Membran	0.000000
MiscVal	-0.000000
GarageYrBlt_Old	-0.000000
MoSold	0.000000
LotArea	0.000000
KitchenAbvGr	-0.000000
BedroomAbvGr	0.000000

When the alpha value in the ridge regression penalty component in the cost function is doubled, the variance decreases by compromising bias, resulting in greater model generalisation.

When the alpha value in Lasso regression is doubled, the number of features decreases even further.

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?

▪ The optimum lambda value in case of Ridge and Lasso is as follows:-

• Ridge – 1 • Lasso – 0.0009

▪ The Mean Squared Error in case of Ridge and Lasso are:

• Ridge - 0.0021396938125939455 • Lasso - 0.0021246459184499694

▪ The Mean Squared Error of both the models are almost same.

▪ Lasso has a greater advantage over Ridge and should be chosen as the final model since it helps in feature reduction (as the coefficient value of some of the features becomes zero).

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: Five most important predictor variables from Lasso regression

- OverallQual
- Total_Bathrooms
- Fireplaces
- TotRmsAbvGrd
- GarageArea

Metrics after removing 5 most important predictor variables

- The R2 Score of the model on the test dataset is 0.6683695043584168
- The MSE of the model on the test dataset is 0.00373686262951211

Model generated from Five most important predictor variables

	Lasso Co-Efficient
MasVnrArea	0.220553
GarageCars	0.191974
LotArea	0.185439
LotFrontage	0.133628
Total_porch_sf	0.103080

R2square value is decrease after removing 5 most important predictor variables

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?

The model should be generalised to ensure that the test accuracy does not fall below the training score. The model should be accurate for datasets other than those used for training. Give no weight to outliers so that the model's projected accuracy is high. That is why we need to do an outliers analysis and save just the values that are relevant to the dataset. Outliers that are irrelevant to the dataset must be deleted. The model cannot be trusted for predictive analysis if it is not robust. The model should be as basic as feasible, as this will reduce accuracy while increasing robustness and generalizability.

It is also understandable in terms of the Bias-Variance trade-off. The simpler the model, the greater the bias, but the lower the variance and the greater the generalizability. Its accuracy implication is that a resilient and generalizable model will perform equally well on both training and test data, implying that accuracy does not differ significantly between training and test data. Bias: A model mistake occurs when the model is unable to learn from the data. A high bias indicates that the model is unable to learn details from the data. On training and testing data, the model performs poorly.

Variance: Variance is an inaccuracy in the model that occurs when the model attempts to overlearn from the data. High variance indicates that the model works extraordinarily well on training data since it has been very well trained on this type of data, but performs very poorly on testing data because it was previously unknown data for the model. To avoid overfitting and underfitting of data, Bias and Variance must be balanced.