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 values of alpha for ridge and lasso regression are:

Ridge : 0.2

Lasso : 0.0001

After doubling the values for ridge and lasso difference between r2 square for train and test is reduced. For lasso prediction has improved a lot. Accuracy of train dataset has gone down a little for both ridge and lasso.

Below statistics with initial alpha values predicted by GridSearch and after doubling the values:

Ridge : 0.2

Lasso : 0.0001

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.893264	0.872780	0.888817
1	R2 Score (Test)	0.711076	0.842955	0.789306
2	RSS (Train)	15.986099	19.054057	16.652030
3	RSS (Test)	23.916343	12.999748	17.440676
4	MSE (Train)	0.125068	0.136543	0.127646
5	MSE (Test)	0.233674	0.172278	0.199547

Ridge : 0.4

Lasso : 0.0002

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.893264	0.861702	0.875452
1	R2 Score (Test)	0.711076	0.857904	0.840628
2	RSS (Train)	15.986099	20.713133	18.653764
3	RSS (Test)	23.916343	11.762307	13.192394
4	MSE (Train)	0.125068	0.142363	0.135101
5	MSE (Test)	0.233674	0.163874	0.173550

After implementing the change, parameters have remained same just their coefficients have changed. Important parameters are:

1. LotArea - Lot size in square feet
2. OverallQual - Rates the overall material and finish of the house
3. OverallCond - Rates the overall condition of the house
4. YearBuilt - Original construction date
5. BsmtFinSF1 - Type 1 finished square feet
6. 1stFlrSF - First Floor square feet

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:

I will choose ridge regression and the prediction power for this model in train and test are similar.

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:

After removing top 5 predictors, below are new top 5:

1. TotalBsmtSF
2. 1stFlrSF
3. GarageArea
4. 2ndFlrSF
5. SaleType_Con

Please refer the notebook for the steps followed.

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:

Model is generalized when its performance on unseen/test data is similar to the training data and values predicted should be under the variance defined by business. To build robust model as part of data analysis treatment of missing values and outliers is critical as it may impact the overall predictability of the model. Additionally its important to select features which are positively correlated to the target variable.