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 value for alpha for Ridge - 20

Optimal value for alpha for Ridge - 0.0001

After doubling the alpha for both , we see that metrics for both haven't changed much. While looking at the coefficients, we can see that it has reduced slightly for all the predictors for both Ridge and Lasso

For Lasso, with alpha 0.0001, the model had made 59 out of 230 features to 0. However, after doubling the alpha, 92(35 more features) were 0 , making the model more simpler

The most important predictor variables after doubling the alpha are as follow

For Ridge:

Features	Coefficients
TotalFlrSF	0.119017
OverallQualCond	0.072648
Neighborhood_Edwards	-0.055379
property_age	-0.053012
Condition1_Norm	0.047115
Neighborhood_NridgHt	0.042577
Neighborhood_Crawfor	0.042269
LotArea	0.041707
Neighborhood_Somerst	0.037913
TotalBsmtSF	0.037841

For Lasso:

Features	Coefficients
Condition2_PosN	-0.758349
MSZoning_RH	0.225659
MSZoning_RL	0.224444
MSZoning_FV	0.214951
TotalFlrSF	0.204466
MSZoning_RM	0.196522
RoofMatl_WdShngl	0.187679
FullBath_3	0.125956
Neighborhood_NoRidge	0.117115
SaleType_ConLD	0.115759

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 metrics for both are quite similar

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.925868	0.947932
R2 Score (Test)	0.897015	0.891404
RSS (Train)	11.897649	8.356521
RSS (Test)	7.421995	7.826347
MSE (Train)	0.107949	0.090469
MSE (Test)	0.130174	0.133673

But since Lasso puts more penalty and has made around 59 predictors coefficient to 0, I would choose the Lasso model for prediction

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 the 5 most important features which were –

Condition2_PosN, RoofMatl_WdShngl, RoofMatl_CompShg, RoofMatl_Roll, RoofMatl_Tar&Grv

The five most important features now are

Features	Coefficients
TotalFlrSF	0.2045
Neighborhood_NoRidge	0.0916
Neighborhood_Somerst	0.089
OverallQualCond	0.0767
Neighborhood_NridgHt	0.0761

Question 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?

Answer :

The model can be made more robust and generalizable by handling the outliers, getting more data to train on, by feature engineering and transforming and deriving new features/information from the existing features and selecting the best features that explains the relationship of the predictors with the features.

When the model is more robust, it will be able to predict correctly even if the dataset is changed. There will be no overfitting problem. In general there will be less bias and less variance