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 1

The optimal value of ridge is 10 and for lasso regression is 0.001.

If we double the value of ridge and lasso alpha then changes in value of the coefficients is shown in the cell 87. Based on R2 and RMSE , when we double the parameter the RIDGE model perform better than lasso.

Metrics Comparison

Ridge

R2 score of train set around the same pointer with no difference from 0.916 to 0.916

R2 score of test set is difference by one pointer from 0.870 to 0.871

Lasso

R2 score of train set decreased by one pointer from 0.915 to 0.914.

R2 score of test set is increased by 2 pointer from 0.874 to 0.876.

1stFlrSF is the most important predictor variable after the change.

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?

Answer2

I will use Lasso as its RMSE is low .The RMS for Lasso is - 0.15124 in comparison to Ridge lamda . Also since it do feature selection as coefficient is zero , this model is calculated to perform better than ridge in this dataset. However in general if we want to do feature selection then we should use Lasso otherwise Ridge Regression. Ridge also helps in keeping the coefficients to a minimum and hence avoid any irregularity in model with large positive values of coefficients.

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?

A***nswer 4***

The five most important predictors before the drop are below.

'1stFlrSF', '2ndFlrSF', 'OverallQual', 'OverallCond', 'SaleCondition_Partial'

The code for dropping and creating another model is shown from cell 90 to 98. After the new model is created below is the five important predictors and their coefficients.

After dropping the top features are below:

	Lasso
GarageArea	***0.079492***
KitchenQual	***0.065497***
LotArea	***0.063554***
Fireplaces	***0.061321***
BsmtQual	***0.047436***

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 4

The ability to perform best on unseen data make the model robust and generalizable. Model should not learn too much from test set to mimic it completely i.e. it should not overfit or underfit the given data. So, we should really decrease the variance and bias just enough to reach the optimal values of low variance and low bias to attain a perfect mode. SO model should be simple and yet accurate. Through regularization this can be attained.

Accuracy will be high for complex models but it's may not be perfect model due to increased bias and hence it's important to balance that out to absorb some decrease in bias to attain optimal performing, robust model.

Yellow highlighters show the optimal point which will balance off variance and bias and accurate enough to get us the robust model for any problem.