Subjective Questions and Answers

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 of alpha for ridge regression= **2.0**

Optimal Value of alpha for lasso regression= **0.0001**

As the value of alpha is doubled, the model output different set of predictors. It also changes variables coefficient value which impact their predicting power. Other changes are observed in performance metrics like R2 score, RSS, MSE and RMSE.

The most important predictor, after the value of alpha is doubled are given below:

Ridge Regression

Predictor- GrLivArea
Predictor Coefficient: 0.112

Lasso Regression

Predictor- GrLivArea
Predictor Coefficient: 0.2312

Parameter	Ridge Alpha = 2.0	Ridge Alpha = 4.0	Lasso Alpha = 0.0001	Lasso Alpha = 0.0002
Train R2 Score	0.9295	0.9256	0.9263	0.9198
Test R2 Score	0.8772	0.8796	0.884	0.8865
Train RSS	1.0721	1.1322	1.1216	1.2203
Test RSS	0.7932	0.7779	0.7494	0.733
Train MSE	0.0011	0.0012	0.0012	0.0013
Test MSE	0.0019	0.0019	0.0018	0.0018

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:

Parameter	Ridge Alpha = 2.0	Lasso Alpha = 0.0001
Train R2 Score	0.9295	0.9263
Test R2 Score	0.8772	0.884
Train RSS	1.0721	1.1216
Test RSS	0.7932	0.7494
Train MSE	0.0011	0.0012
Test MSE	0.0019	0.0018

It is clear from the metrics in the table above, that Lasso has improved R2Score on test dataset and slightly lower RSS (Residual Sum of Squares) and MSE (Mean Squared Error) as compared to Ridge.

Also, as Lasso help in feature reduction by making coefficients of few variables zero, it has better edge over Ridge.

Hence, the top variables predicted by Lasso can be selected as significant predictors to predict House price.

Question 3

After building the model, you realized 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 top 5 predictor variables of Lasso model are:

1. GrLivArea - Above grade (ground) living area square feet
2. OverallQual_10 - Rates the overall material and finish of the house
3. OverallQual_9 - Rates the overall material and finish of the house

4. TotalBsmtSF - Total square feet of basement area
5. OverallQual_8 - Rates the overall material and finish of the house

After dropping/excluding the five most important predictor, following are the new top five predictors:

1. BsmtFinSF1 - Type 1 finished square feet
2. 2ndFlrSF - Second floor square feet
3. OpenPorchSF - Open porch area in square feet
4. Fireplaces_2 - Number of fireplaces
5. OverallCond_7 - Rates the overall condition of the house

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?

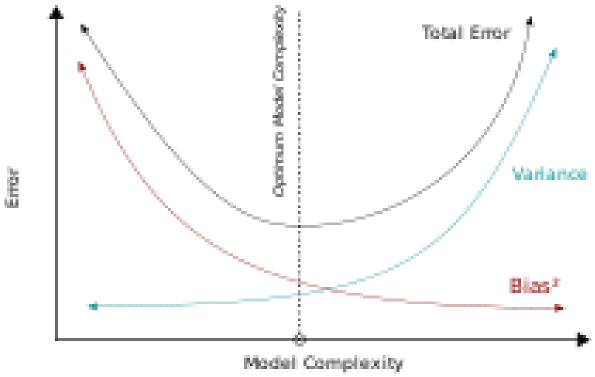
Answer:

Per, Occam's Razor— given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes fewer error on the test data due to following reasons: -

- Simpler models are usually more 'generic' and are more widely applicable
- Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
- Simpler models are more robust, whereas Complex models tend to change wildly with changes in the training data set
- Simple models have low variance, high bias and complex models have low bias, high variance
- Simpler models make more errors in the training set. Complex models lead to overfitting — they work very well for the training samples, fail miserably when applied to other test samples.

Therefore, to make the model more robust and generalizable, make the model simple but not simpler which will not be of any use.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model. Making a simple model leads to Bias-Variance Trade-off.



Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve, for e.g., one that gives same answer to all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high.

Variance refers to the degree of changes in the model itself with respect to changes in the training data. Thus, accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph