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?

Answer

After log transforming target variable, and performing GridSearchCV, following optimal alpha values were found for ridge, and lasso regression:

Ridge Regression: 10

Lasso Regression: 0.001

As performed at the end of the python notebook, if the value of the optimal alpha is increased or doubled, then error term will increase for both Ridge and Lasso Regression. Also, the value of the adj. R2 is decreased when we increase the value of optimal alpha. So, it resulted in underfitting of the data.

Top 20 imp variables post this change are:

1. GrLivArea
2. OverallQual
3. House_Age
4. SaleType_New
5. Neighborhood_Crawfor
6. OverallCond
7. MSZoning_RL
8. Neighborhood_Somerst
9. TotalBsmtSF
10. MSSubClass_160
11. Foundation_PConc
12. Condition1_Norm
13. CentralAir_Y
14. GarageCars
15. BsmtFinSF1
16. Functional_Typ
17. SaleCondition_Normal
18. GarageArea
19. LotArea
20. HeatingQC_TA

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 found following alpha values for ridge, and lasso regression:

Ridge Regression: 10

Lasso Regression: 0.001

Well, as entire concept behind regularization is to create an optimal complex model, i.e. model which is simple, and is good at prediction too. Simpler models are usually more generic and need fewer examples for training.

We will choose to apply Lasso Regression because r2, Mean Squared Error (MSE) of both Lasso, and Ridge are approximately same and on top of that Lasso Regression did feature selection resulting in a simpler model. Lot of coefficients in case of Lasso are zero, and that resulted in feature selection. So, since Lasso regression resulted in simpler model, and that too while maintaining good prediction parameters, so that's why we have chosen Lasso Regression.

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

Top 5 predictor variables in Lasso model are as following:

1. GrLivArea
2. Neighborhood
3. OverAllQual
4. SaleType
5. House_Age (Derived from YrSold and YearBuilt columns)

After building Lasso Model again with good R2 score, and without above mentioned top 5 imp predictor variables, following are the next top 5 imp predictor variables:

1. MSZoning
2. 2ndFlrSF
3. SaleCondition
4. CentralAir
5. 1stFlrSF

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

We can make model more robust and generalizable by adding Regularization to the model, as it adds cost to the coefficients of the model. Regularization creates an optimally complex model, i.e. a model which is simple, and has good prediction value.

Simpler models are usually more generic, and more widely applicable. They also require fewer training examples for effective training as compared to complex models. If we don't use the regularization, then the model can overfit on the training set, and prove less reliable on predictions on the test set. Accuracy of the model with no regularization would be more on training set, as compared to test set. On the other hand, regularization decreases the accuracy on the training set, and try to increase it over the test set.

Since, along with reducing the Mean squared error (MSE) of the training set, model also must make sure that coefficients of the predictor variables are also simple, it decreases the overall accuracy on the training set.

Whereas if we don't introduce regularization, then model is only focused towards reducing MSE, and hence it results in higher accuracy of the model