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

The optimal value of alpha for ridge and lasso regression for chosen model (Model1 in notebook) is,

Regression	Alpha Value
Ridge	100
Lasso	0.001

The optimal value has been found using GridCV method running 5 fold operations on the dataset.

If we double the values of alpha (at the risk of overfitting), following changes in the model observed for Ridge and Lasso.

Ridge Model

When we run Ridge model with alpha as 200, following are the R2 scores of the model,

Alpha	Train R2 Score	Test R2 Score
100	0.8808953549868707	0.7754219439361587
200	0.8777301025703378	0.7781874577033928

As shown in the above table, Train R2 score has reduced marginally and that could be the sign of slight underfitting. However, the test R2 score is maintained.

Following table shows the top features when alpha is 100 and 200,

Alpha = 100	Alpha = 200
OverallQual	OverallQual
GrLivArea	GrLivArea
YearRemodAdd	Fireplaces
GarageArea	1stFlrSF
Fireplaces	YearRemodAdd

Lasso Model

When we run Lasso model with alpha as 0.002, following are the R2 scores of the model,

Alpha	Train R2 Score	Test R2 Score
0.001	0.8826796954041278	0.7671908809378312
0.002	0.8808105592232676	0.7683508126914869

After doubling the alpha, in case of Lasso model, the Train and test R2 scores have remained pretty much same.

Following table shows the top features when alpha is 0.001 and 0.002,

Alpha = 0.001	Alpha = 0.002
OverallQual	OverallQual
GrLivArea	GrLivArea
GarageArea	GarageArea
YearRemodAdd	YearRemodAdd
TotalBsmtSF	TotalBsmtSF

It is interesting to note that Lasso model is quite stable and there are no changes to top 5 predictors when the alpha is changed.

In my experiments I found that when alpha goes beyond 0.05 the R2 scores get deteriorated.

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 have applied GridSearchCV model selection technique to determine the value of alpha for Ridge and Lasso individually.

The dataset has been divided into 5 folds and the GridSearchCV would run the algorithm finding the optimum alpha value from the given set of alphas.

The higher the alpha, the coefficients will be minimized and that may result into underfitting and lower the alphas would result into overfitting.

The GridSearch algorithm would perform train-test process for each k-folds with k-1 as train and one set as test data for each alpha values.

We have taken 32 alpha values and therefore it performs 160 fits for each estimator (i.e. ridge and lasso).

The alpha given in the best parameter, as returned by the GridSearch algorithm is what I would apply to final models, as that would be the optimum value of alpha for the given estimator.

For example, in case of Ridge model the alpha's value is 100 whereas in case of Lasso it's 0.001

I tried other nearby alpha values to the found ones and the output R2 scores remained pretty much same, however once the alpha became 10x bigger the underfitting was evident.

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

As per the original Lasso model, following are the top 5 predictors,

Top 5 predictors
OverallQual
GrLivArea
GarageArea
YearRemodAdd
TotalBsmtSF

Removing them from the model and rebuilding Lasso model with the rest of the variables (count : 51), following is the results.

Please refer to **Question 3** section in the accompanying notebook.

Statistics	**Train Data**	**Test Data**
R2 Score	0.8588706795545306	0.7540074331152243
RSS	18.038971167760728	15.205447012136808
MSE	0.019272405093761463	0.03782449505506669

So it can be seen that R2 score went down because of no presence of the top 5 predictors of the previous model.

Following are the top 5 predictors of this new model,

Top 5 predictors (new model)
1stFlrSF
2ndFlrSF
KitchenQual
GarageType_Attchd
Fireplaces

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

The Model1 built using Lasso is robust as well as generalizable. It is robust because even if we removed the top 5 predictors (variables) from the dataset, it performed well on R2 scores.

Secondly it's able to reduce the predictor variables count from 56 to 44, which in itself a great deal because it would make the model simpler and less prone to variances in the target variables for the unknown data. Having many variables in the model would result into uptick of variances when introduced with unseen data.

The accuracy of model on train data is roughly 0.88 and on test data roughly 0.76, which are decent scores.

However, in case of model2 with many more variables (154 vs 56 in model1) , the accuracy R2 went upto 0.91 for train data and 0.80 for test data. So there was a difference of 0.04 in the accuracy but at a much greater cost of making model extremely complex.

Although Lasso was able to reduce the variables from 154 to 108 (by finding their coefficients 0), the model2 is still a very complex model and even a smaller change in these 108 variables would result into unexpected variances in the outcome.

Hence it is important to note that if there is not much difference in accuracy then choosing the simpler model is a wise decision.