

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 lambda value in case of Ridge and Lasso is as below:

Ridge - 2.0

Lasso - 0.0004

With the above lambda values of ridge and lasso, we were getting:

	r2_score	rss	rmse
Models			
Linear_Train	0.885430	1.918958	0.045073
Linear_Test	0.863485	1.023612	0.045073
Ridge_Train	0.883507	1.951158	0.045073
Ridge_Test	0.864062	1.019282	0.045073
Lasso_Train	0.879072	2.025449	0.045073
Lasso_Test	0.863906	1.020459	0.045073

When we double the values, we get:

	r2_score	rss	rmse
Models			
Linear_Train	0.885430	1.918958	0.047012
Linear_Test	0.863485	1.023612	0.047012
Ridge_Train	0.879895	2.011661	0.047012
Ridge_Test	0.860707	1.044443	0.047012
Lasso_Train	0.868440	2.203514	0.047012
Lasso_Test	0.852485	1.106093	0.047012

So, on increasing the lambda values, the r2_score for ridge and lasso regression decreases showing that lesser number of actual values of y are predicted correctly by the best-fit regression line. On the other hand, the rss value increases, showing that a larger error between y-actual and y-predicted values. The model starts underfitting.

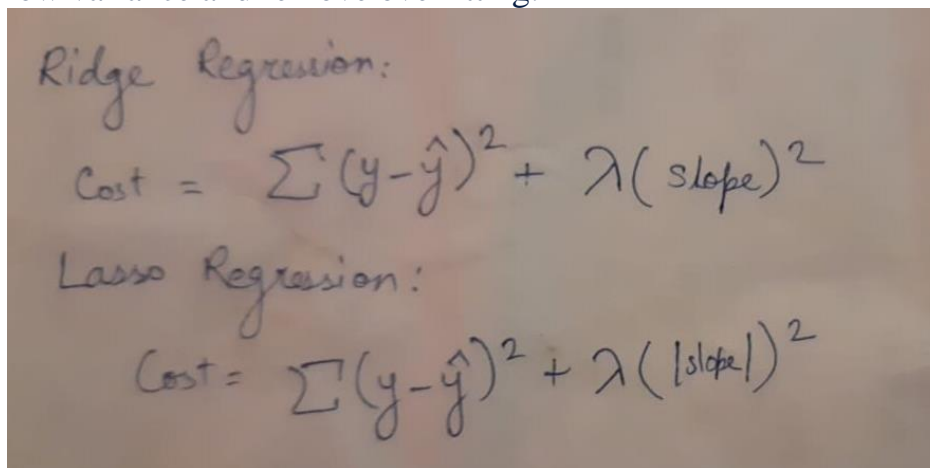
The features most important for house price prediction selected by Lasso Regression are:

- LotArea
- BsmtFinSF1
- TotalBsmtSF
- 2ndFlrSF
- FullBath
- TotRmsAbvGrd
- Fireplaces
- GarageArea
- HouseAge
- MSSubClass_DUPLEX - ALL STYLES AND AGES
- Neighborhood_Crawfor
- Neighborhood_NoRidge
- Neighborhood_StoneBr
- BldgType_Duplex
- OverallQual_Excellent
- OverallQual_Very Excellent
- OverallQual_Very Good
- OverallCond_Excellent
- SaleType_New

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: Ridge and Lasso regression is used to convert the high variance in a model to low variance and remove overfitting.



The image shows handwritten mathematical formulas for Ridge and Lasso regression. For Ridge Regression, the cost function is given as $\text{Cost} = \sum (y - \hat{y})^2 + \lambda (\text{slope})^2$. For Lasso Regression, the cost function is given as $\text{Cost} = \sum (y - \hat{y})^2 + \lambda (|\text{slope}|)^2$.

Normally, we always try to reduce the RSS. Here, we are also adding one penalty associated with the slope. In case of overfitting, slope is high, RSS is approx. 0 and therefore, penalty is high. In normal cases, RSS is of smaller value and slope is lower than the case of overfitting, so penalty is also less.

In our case,

Ridge RSS – 1.951158

Ridge Lambda - 2.0

Ridge Slope – 0.014799

Cost function value = 1.980756

Lasso RSS – 2.025449

Lasso Lambda - 0.0004

Lasso Slope - 0.019395

Cost function value = 2.025456758

Lesser the cost function, the better.

Also, the r^2 _score for Ridge regression is better than Lasso. So, in our case, Ridge regression is better with lambda 2.0.

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:

None. I'm getting coefficient 0 for all the rest with 0 r^2 score. Please refer to notebook for reference.

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:

We can create a generalized model if it has a low bias and a low variance.

Bias: Difference between predicted and actual values in a ML model due to unavoidable errors is called bias. Low bias means the ML model has made fewer assumptions about the target function. It will be a simple model.

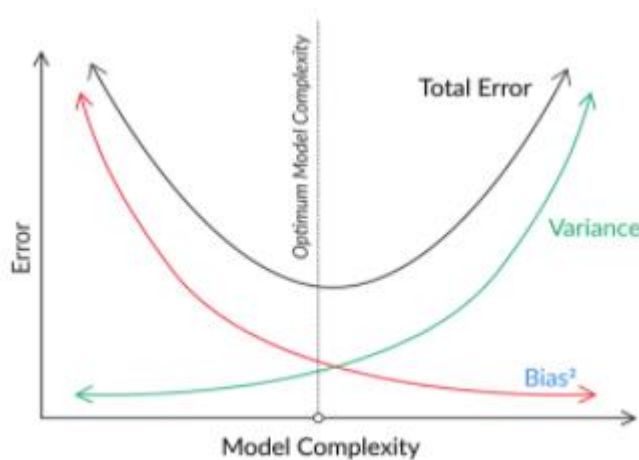
High bias on the other hand will make more assumptions and will be unable to capture important features from dataset.

Variance: *It means how much test data/unseen data differs from actual value. Low variance means there is a small variation in the prediction of the target function with changes in the training data set.*

At the same time, High variance shows a large variation in the prediction of the target function with changes in the training dataset.

There are four kinds of model :

1. High bias and high variance – this is the most error prone model.
2. High bias and low variance – this kind of model underfits the data. It also occurs due to a very high value of lambda.
3. Low bias and high variance – this kind of model overfits the data. It also occurs due to a low value of lambda.
4. Low bias and low variance model - this model fits perfectly with the test data.



The best model lies somewhere just before the point where variance and bias intersect.

Such a model will have a high accuracy as it indicates that the total error is least in such a model.