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 Regression Model is 1 and optimal value of alpha for lasso is 0.001.

ridge = Ridge(alpha=2) so initially the optimal value was 1 we need to do double so the alpha value becomes 2 we then fit the model then predict the features.

Calculation of r-square value for train and test data which shows us that it has decreased in case of training data by and increased for test data.

In case of lasso regression model lasso = Lasso(alpha=0.002) initially the optimal value which we choose for lasso model is 0.001 when we double the value it becomes 0.002.

Again, we fit the model make necessary predictions then calculate r2 value for train and test data and make comparison. on comparison both the train set and test set has increased eventually.

The most Important predictor variables are

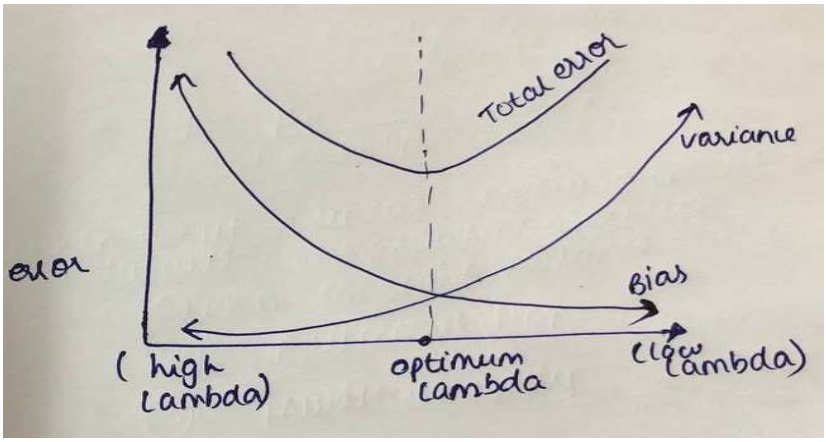
```
MSZoning_RL
MSZoning_RM
MSZoning_FV
RoofMatl_WdShngl
Neighborhood_NoRidge
HouseStyle_2Story
BsmtFinType1_GLQ
Neighborhood_StoneBr
Neighborhood_NridgHt
Foundation_PConc
```

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?

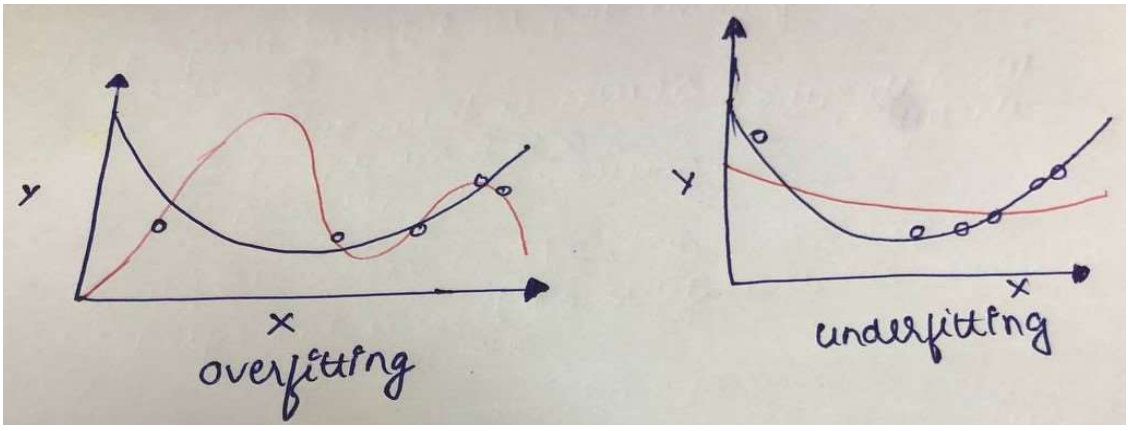
Answer:

The bias is high when model fails on training data and high variance when model fails on test data. Simple model has high bias and low variance where as a complex model has low bias and high variance in which both the model gives us a increased total error.



We need a low total error where both variance and bias are low this can be achieved only through Regularization which helps in managing model complexity by shrinking the coefficients to 0.

Ridge regression we estimate the coefficients with the help of a cost function which adds a penalty term. penalty is added so that it helps in shrinking the terms. So, lambda plays a very important role, choosing a correct lambda number is also difficult if it is too small overfitting occurs and too large lambda number leads to underfitting.



Lasso regression is very much similar to ridge regression but the difference occurs in case of penalty term. Both the regression tries to shrink towards 0 but in lasso it tries to make coefficients to exact 0 which helps in feature selection.

It depends on the data given to us to solve when it goes under data preparation, we then can decide on the above points to choose Lasso or Ridge. I will choose lasso because mean square and r2 are less when compared to ridge.

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:

Top 5 predictors for the lasso model are- MSZoning_RL, MSZoning_RM ,MSZoning_FV

RoofMatl_WdShngl , Neighborhood_NoRidge

We create a list of features which we want to delete it, we should drop from both train and test data then print the test and train data with .head() using the Lasso model first choose the alpha value then fit the model. Then calculate the metric values like r-square or mean square error. Then by observing the data we can see on r2 testing and training data is reduced.

After that the top 5 predictor variables are-

Street_Pave, Foundation_PConc,11stFlrSF,GrLivArea, SaleType_WD

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:

In a robust model what actually happens is when some new data is added or removed or some kind of change is occurring in that data it doesn't affect the output a lot. A generalizable model is where a new data comes into picture it should be stable enough to handle it have a good intuition (mathematical reasoning) to predict if it's a good one or not.

For both these conditions to work we should eradicate the overfitting concept. what exactly is overfitting is when the data gets trained well on train data but not well on test data.

Model shouldn't be too complex; model complexity depends on magnitude of coefficients and no of coefficients. Accuracy also plays a role in case of model but when the model is having high accuracy it will lead to complexity model with low bias and high variance, when we reduce the variance eventually the bias increases and it becomes a large model.in both

the cases the test error is high. we need to reduce this error by using regularization techniques which is Ridge and Lasso which helps in balancing the complexity by shrinking the coefficients towards 0.