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 :

Alpha for Ridge: 5.0
Alpha for Lasso: 100

If we choose to double the value of alpha for both ridge and lasso, A few things will happen :

1) For Ridge regression model, Coefficients will further shrink towards 0, For Lasso more coefficients will become 0
2) Model will start to underfit, this can lead to model's accuracy.
3) Model will lose be understanding to some underlying relationships and become less sensitive to underlying quirks in the training data.

For Ridge most important features will be: OverallQual , GrLivArea , GarageCars

For Ridge most important features will be: OverallQual, OverallCond and YearRemodAdd

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?

So as per my analysis there are 3 factors where we can judge the two :

1) R2 Scores :
   For Ridge: Train = 0.893, Test = 0.860
   For Lasso: Train = 0.897, Test = 0.872
2) RSS :
   For Ridge: Train = 655662528159.61, Test = 428190679992.92
   For Lasso: Train = 633892743882.41, Test = 391100974448.14
3) MSE :
   For Ridge: Train = 641548461.99, Test = 977604292.22
   For Lasso: Train = 620247303.21, Test = 892924599.20

In conclusion, since in all the three parameters Lasso is better and Lasso makes simpler models (low #features) , I will prefer Lasso over 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?

For the original Alpha value i.e 100, for a lasso regression model , Top 5 most important predictor variables(after removing previous top 5) are :

1) TotalBsmtSF
2) 2ndFlrSF
3) Neighborhood_StoneBr
4) Neighborhood_NoRidge
5) RoofMatl_WdShngl

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?

There are few factors that helps ensure the model is robust and generalizable

- Diverse data sets : A Large enough and diverse dataset ensures captyring the underlying patterns and reduces bias
- Proper feature selection : It increase predicting power and keeps the model simple
- Cross validation : It reduces risk of overfitting and ensures consistency of model's performance
- Regularization techniques : It prevent overfitting and makes the model simple

Implications on accuracy :

- By maintaining a balance between variance and bias we ensures model is robust and generalizable, If we ensure that model's accuracy is high on test as well train data , we can balance bias and variance.
- Regularization methods like Lasso and Ridge are essential. They keep the model simple enough to generalize well but complex enough to capture key patterns. This balancing act is crucial for maintaining accuracy on both training and new, unseen data.
- Cross-validation enhances model accuracy by ensuring it performs consistently across different data subsets, thus reducing the risk of overfitting. This method provides a more reliable estimate of the model's performance on unseen data.
- Proper feature selection focuses the model on the most relevant variables, significantly improving its predictiveness and reducing unnecessary complexity. This leads to a more accurate and interpretable model, especially when applied to new, unseen datasets.