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

1.1) Optimal value of alpha for ridge regression=10 and for Lasso=0.001
1.2) The accuracy of train and test scores have reduced slightly for lasso and in return the error terms have increased a bit. For ridge, train score decreased while test score increase very slightly. Also, there error term remain almost the same.
Original Model:

Lasso
R2 train : 0.9206709153542387
R2 test : 0.9067784241823188
RMSE : 0.11546718823765784

Ridge
R2train : 0.9364594823911133
R2 test : 0.9077597079466584
RMSE : 0.11485785595060977

New Model with double alpha:
Lasso
R2 Test : 0.9122051184016314
R2 Train : 0.9031718080740088
RMSE : 0.11767962655896999

Ridge
R2 Test : 0.9272757940609145
R2 Train : 0.9100087351460887
RMSE : 0.11344896766981008

Also, number of predictor variables reduced for Lasso from 95 to 48.

1.3) The predictor variables after increase in alpha

- Lasso
  - GrLivArea
  - OverallQual
  - SaleCondition_Partial
  - PropAge
  - Neighborhood_Crawfor
  - OverallCond
- Ridge
  - OverallQual
  - Neighborhood_Crawfor
  - GrLivArea
  - PropAge
  - OverallCond
  - Neighborhood_IDOTRR

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 2
The test accuracy is near about similar for Lasso & Ridge, though train accuracy was slightly lesser compared with Ridge.
With the similar accuracy, I would choose Lasso since it brings and assigns a zero value to insignificant features. It is always advisable to choose simpler model.

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:
Solution in the python notebook. The top 5 most important predictor variables are:

1. ScreenPorch
2. PropAge
3. PoolArea
4. MSZoning_FV
5. Neighborhood_Crawfor

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 model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.
Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.
Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.
It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.

-