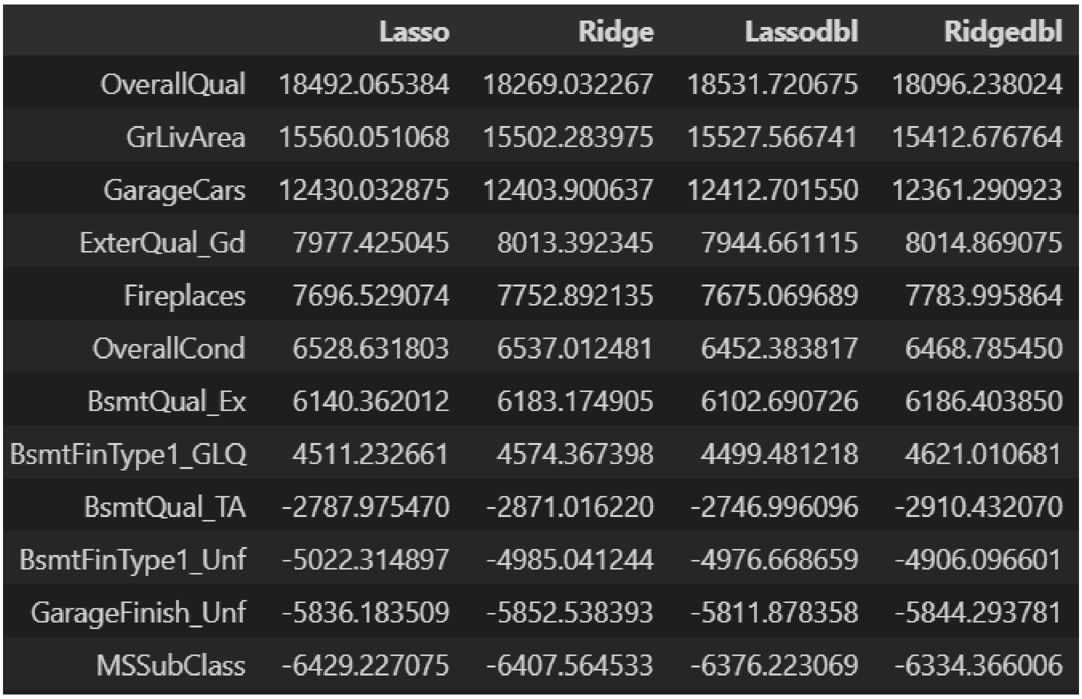
Question 1

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?

Ans: Optimal value of alpha for ridge is 10.001 and for lasso is 50.001. After doubling alpha, not much change was observed in the model coefficients (slight decrease). The performance and the important predictor variables remain the same.

R2 score	optimal alpha	2* optimal alpha
Lasso	0.79	0.789
Ridge	0.789	0.787

Performance on test set



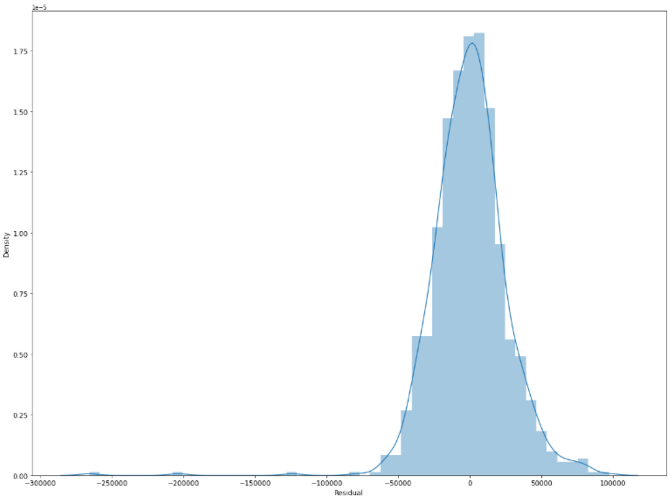
	Lasso	Ridge	Lassodbl	Ridgedbl
OverallQual	18492.065384	18269.032267	18531.720675	18096.238024
GrLivArea	15560.051068	15502.283975	15527.566741	15412.676764
GarageCars	12430.032875	12403.900637	12412.701550	12361.290923
ExterQual_Gd	7977.425045	8013.392345	7944.661115	8014.869075
Fireplaces	7696.529074	7752.892135	7675.069689	7783.995864
OverallCond	6528.631803	6537.012481	6452.383817	6468.785450
BsmtQual_Ex	6140.362012	6183.174905	6102.690726	6186.403850
BsmtFinType1_GLQ	4511.232661	4574.367398	4499.481218	4621.010681
BsmtQual_TA	-2787.975470	-2871.016220	-2746.996096	-2910.432070
BsmtFinType1_Unf	-5022.314897	-4985.041244	-4976.668659	-4906.096601
GarageFinish_Unf	-5836.183509	-5852.538393	-5811.878358	-5844.293781
MSSubClass	-6429.227075	-6407.564533	-6376.223069	-6334.366006

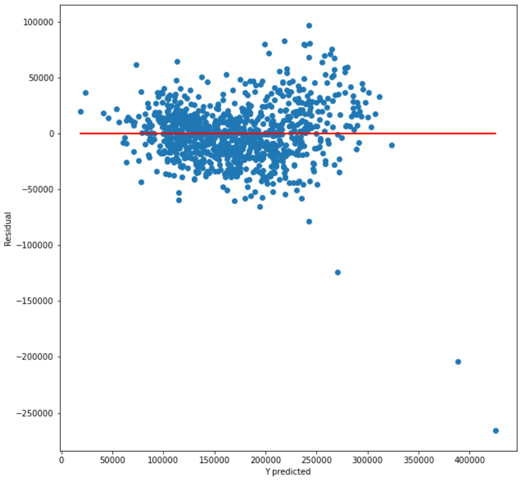
Model Coefficients

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?

Ans: Lasso model is selected for this dataset as lasso is giving slightly higher r2 than ridge for this dataset. Also lasso is satisfying all the assumptions of linear regression such as homoscedasticity and normal residual distribution with mean 0. In general, optimum model is selected by checking the scores on the validation set.





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?

Ans: From the metrics we can see that the model performance is widely effected by removing the top 5 features (R2 on test set -0.35). Probably selecting more than 12 features in the original model might help to improve the performance after removing top 5 features. Cannot comment on the top five features as the model is under fitting.

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Ans: The model should not under fit or over fit the training data. This can be ensured by using a reasonably good model and selecting right features. The assumptions of the model used should be satisfied. There should not be much difference in train and test metrics (<=5%). Compromise in accuracy can be done to make the model more generalizable as using a complex model can lead to poor test data metrics.