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

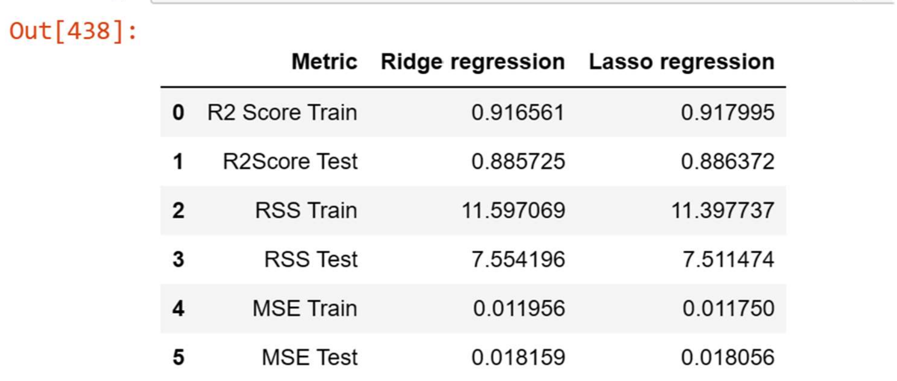
The optimal value for alpha as fetched through the analysis is as follows:

Ridge Model Alpha Value	3.0
Lasso Model Alpha Value	0.0001

Changing the alpha values to double i,e

New Ridge Model Alpha Value	6.0
New Lasso Model Alpha Value	0.0002

The Comparison results between the models with the doubled alpha values is:

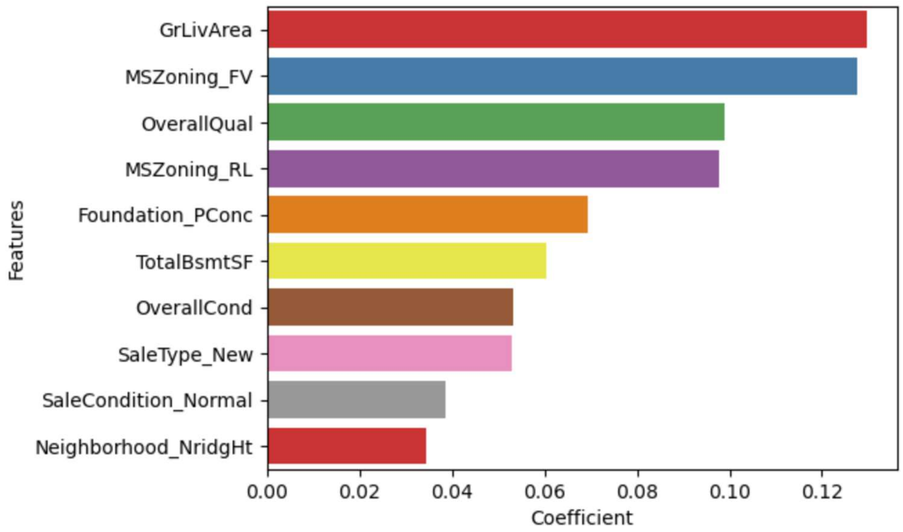


Out[438]:

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.916561	0.917995
1	R2Score Test	0.885725	0.886372
2	RSS Train	11.597069	11.397737
3	RSS Test	7.554196	7.511474
4	MSE Train	0.011956	0.011750
5	MSE Test	0.018159	0.018056

Based on the above result we notice that the better one is still Lasso.

With the new change implementation, the most important predictors variables remain the same:



- GrLivArea:Above grade (ground) living area square feet also has significant increase in the sales price.
- MSZoning_FV: Floating Village Residential also has significant effect in the sales price.
- OverallQual: Quality of over all house also has significant increase in the sales price.
- MSZoning_RL : Residential Low Density is good then it will also has significant effect on sales price.
- Foundation_PConc: Concrete foundation has also significant impact on the sales price
- TotalBsmtSF:Total square feet of basement area is also a reason to increase in salesprice
- OverallCond : If the Overall Condition is Excellent the SalePrice is higher
- SaleType_New : If New is sold then higher sale price
- SaleCondition_Normal: Normal Sale when compared to othersales has an improvement in the sales price
- Neighborhood_NridgHt: Sale price is high if in Northridge heights neighborhood

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

Based on the results obtained and the below observations made, Lasso regression better performing compared to Ridge

1) The some of the coefficients become 0, thus resulting in model selection and, hence, easier interpretation, particularly when the number of coefficients is very large.
2) The slightly lower R2 Scores and the Mean Square Error values indicate that Lasso is a better performing model.

Question 3

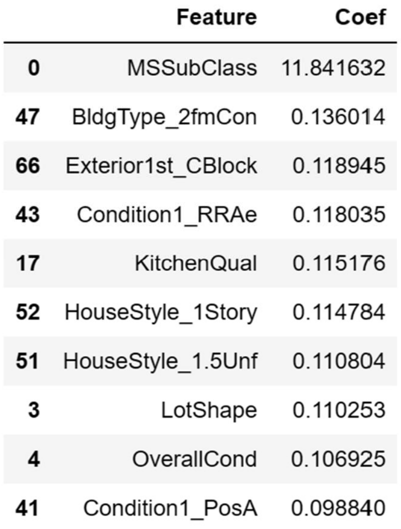
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer 3

After removing the earlier 5 most important predictor variables in the lasso model and then creating another model and evaluating it(using the python notebook), the below predictor variables emerge as the five most important ones.

1) Overall Condition- Overall condition of the house
2) Condition1_PosA – Proximity to various condition - Adjacent to positive off-site feature
3) LotShape – General Shape of the Property
4) HouseStyle_1.5Unf – Style of dwelling - One and one-half story: 2nd level unfinished

5) HouseStyle_1Story - Style of dwelling – One story



	Feature	Coef
0	MSSubClass	11.841632
47	BldgType_2fmCon	0.136014
66	Exterior1st_CBlock	0.118945
43	Condition1_RRAe	0.118035
17	KitchenQual	0.115176
52	HouseStyle_1Story	0.114784
51	HouseStyle_1.5Unf	0.110804
3	LotShape	0.110253
4	OverallCond	0.106925
41	Condition1_PosA	0.098840

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 4:

Model's robustness & generalization can be determined by the below

1) Model is simple
2) Cross validation techniques like k-fold is applied to assess the model's performance on different subset of data. This helps in generalizing the model.
3) Splitting Train and Test data such that training data is used only for training and then test data is used to test to remove any bias and model can perform better on unforeseen data.
4) Applying regularization to prevent overfitting
5) Predicted variables are significant
6) The test accuracy is not very less compared to training score

The implications for accuracy revolve around mitigating overfitting, finding an optimal model complexity, and ensuring that the model's performance on testing data is indicative of its real-world effectiveness.