Advanced Regression - Subjective Question

PRAVIN TAWADE

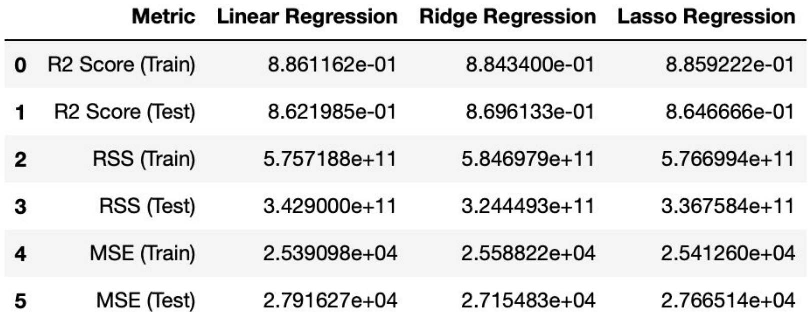
Question 1:

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?

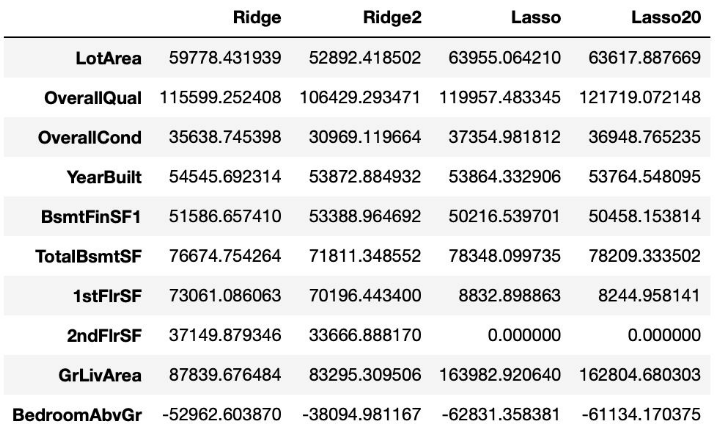
Answer:

- The optimal value of alpha for ridge and lasso regression
  - Ridge Alpha: **1**
  - Lasso Alpha: **10**
- R2score on training data has decreased but it has increased on testing data
- Predictors are same but the coefficient of these predictor has changed

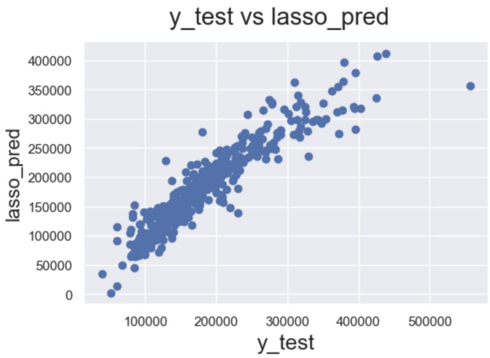
Question 1: (cont..)

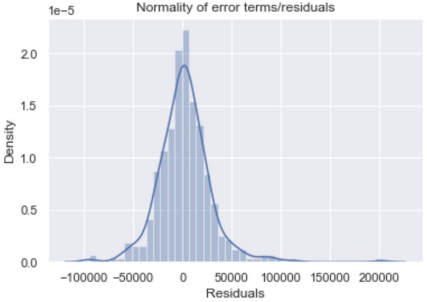


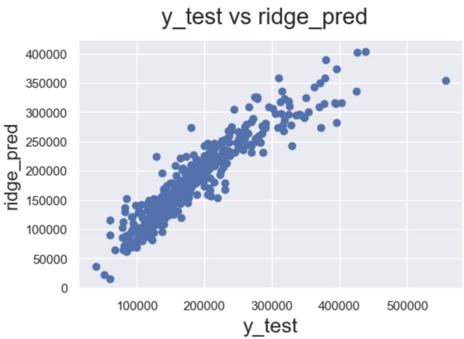
	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.861162e-01	8.843400e-01	8.859222e-01
1	R2 Score (Test)	8.621985e-01	8.696133e-01	8.646666e-01
2	RSS (Train)	5.757188e+11	5.846979e+11	5.766994e+11
3	RSS (Test)	3.429000e+11	3.244493e+11	3.367584e+11
4	MSE (Train)	2.539098e+04	2.558822e+04	2.541260e+04
5	MSE (Test)	2.791627e+04	2.715483e+04	2.766514e+04

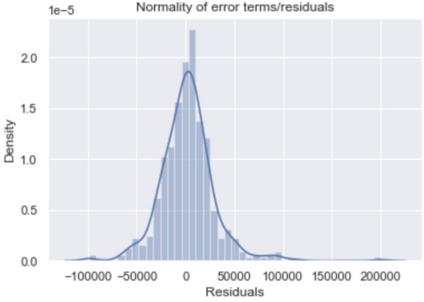


	Ridge	Ridge2	Lasso	Lasso20
LotArea	59778.431939	52892.418502	63955.064210	63617.887669
OverallQual	115599.252408	106429.293471	119957.483345	121719.072148
OverallCond	35638.745398	30969.119664	37354.981812	36948.765235
YearBuilt	54545.692314	53872.884932	53864.332906	53764.548095
BsmtFinSF1	51586.657410	53388.964692	50216.539701	50458.153814
TotalBsmtSF	76674.754264	71811.348552	78348.099735	78209.333502
1stFlrSF	73061.086063	70196.443400	8832.898863	8244.958141
2ndFlrSF	37149.879346	33666.888170	0.000000	0.000000
GrLivArea	87839.676484	83295.309506	163982.920640	162804.680303
BedroomAbvGr	-52962.603870	-38094.981167	-62831.358381	-61134.170375







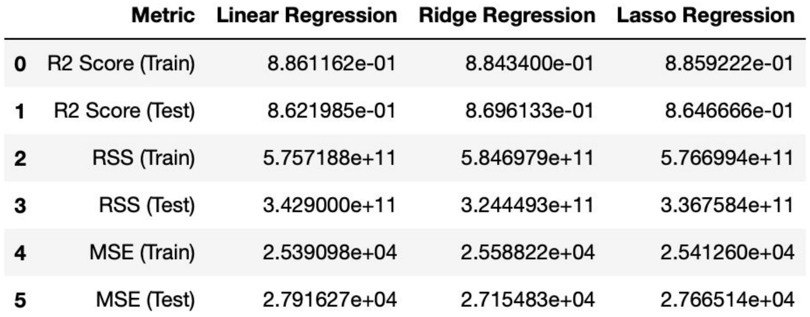


Question 2:

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 **r2_score** of lasso is slightly **higher** than lasso for the test dataset compare to ridge score, so we will **choose lasso regression** to solve this problem



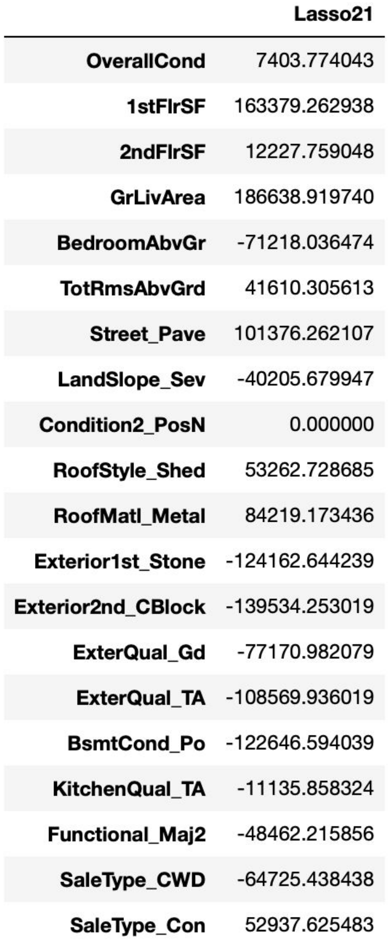
	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.861162e-01	8.843400e-01	8.859222e-01
1	R2 Score (Test)	8.621985e-01	8.696133e-01	8.646666e-01
2	RSS (Train)	5.757188e+11	5.846979e+11	5.766994e+11
3	RSS (Test)	3.429000e+11	3.244493e+11	3.367584e+11
4	MSE (Train)	2.539098e+04	2.558822e+04	2.541260e+04
5	MSE (Test)	2.791627e+04	2.715483e+04	2.766514e+04

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:
With the references in changes in the coefficients after regularization, below mentioned top 5 variables are significant in predicting the price:

- **11stFlrSF** : First Floor square feet
- **GrLivArea** : Above grade (ground) living area square feet
- **Street_Pave** : Pave road access to property
- **RoofMatl_Metal** : Roof material_Metal
- **RoofStyle_Shed** : Type of roof(Shed)



	Lasso21
OverallCond	7403.774043
1stFlrSF	163379.262938
2ndFlrSF	12227.759048
GrLivArea	186638.919740
BedroomAbvGr	-71218.036474
TotRmsAbvGrd	41610.305613
Street_Pave	101376.262107
LandSlope_Sev	-40205.679947
Condition2_PosN	0.000000
RoofStyle_Shed	53262.728685
RoofMatl_Metal	84219.173436
Exterior1st_Stone	-124162.644239
Exterior2nd_CBlock	-139534.253019
ExterQual_Gd	-77170.982079
ExterQual_TA	-108569.936019
BsmtCond_Po	-122646.594039
KitchenQual_TA	-11135.858324
Functional_Maj2	-48462.215856
SaleType_CWD	-64725.438438
SaleType_Con	52937.625483

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 generalized so that the test accuracy is not lesser than the training score.
- The model should be accurate for datasets other than the ones which were used during training.
- Too much importance should not given to the outliers so that the accuracy predicted by the model is high.
- To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset.
- If the model is not robust, It cannot be trusted for predictive analysis.

Conclusion

Based on our analysis and as per our final Model and with the references in changes in the coefficients after regularization, the top 10 predictor variables that influences the House sell price are:

- **LotArea** ----- Lot size in square feet
- **OverallQual** ----- Rates the overall material and finish of the house
- **OverallCond** ----- Rates the overall condition of the house
- **YearBuilt** ----- Original construction date
- **BsmtFinSF1** ----- Type 1 finished square feet
- **TotalBsmtSF** ----- Total square feet of basement area
- **GrLivArea** ----- Above grade (ground) living area square feet
- **TotRmsAbvGrd** ----- Total rooms above grade (does not include bathrooms)
- **Street_Pave** ----- Pave road access to property
- **RoofMatl_Metal** ----- Roof material_Metal

Thank you