Subjective Questions: Advanced Regression
Submitted by: Nagendra J

Question 1: What is the optimum 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:

a) The hyperparameter alpha for ridge and Lasso is shown in figure 1.1 and figure 1.2

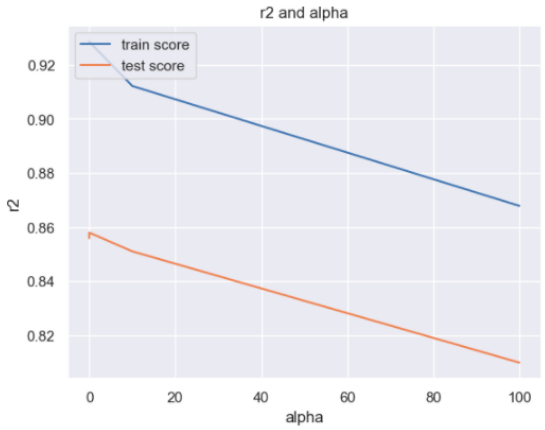


Figure 1.1: Ridge regression

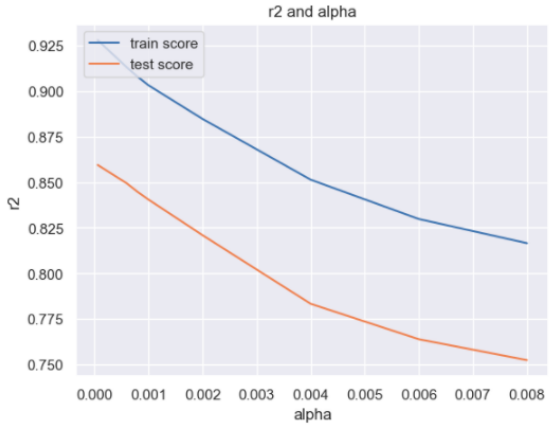
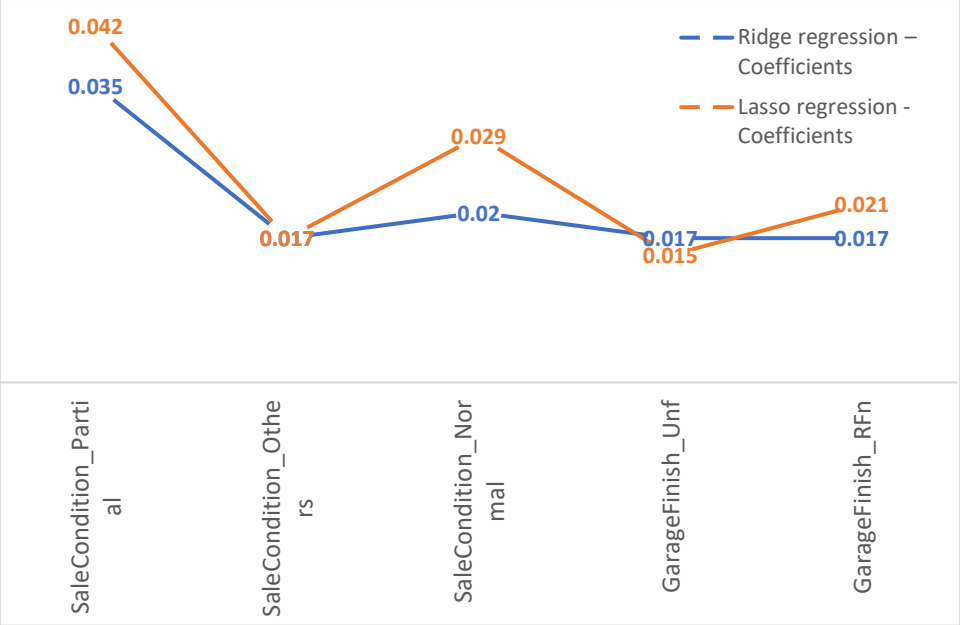


Figure 1.2: Lasso regression

From the above figure it is now noticed the r2 value decreases as and when the alpha value is increased. This narrow down (constant variance) is noticed when alpha is 10 and 0.001 for ridge regression and Lasso regression.

b) When we double the value of alpha values, the score changes from "0.9095 to 0.9010" for "train data" and from "0.8743 to 0.8668" for "test data" in Ridge regression. Similarly, for Lasso regression the score changes from "0.8978 to 0.8803" for train data and from "0.8641 to 0.8493" for test data. The coefficients are listed in the table below:

Variable	Ridge regression – Coefficients		Lasso regression - Coefficients	
	Alpha = 10	Alpha = 20	Alpha = 0.001	Alpha = 0.002
SaleCondition_Partial	0.143	0.108	0.197	0.155
SaleCondition_Others	0.106	0.089	0.121	0.104
SaleCondition_Normal	0.098	0.078	0.099	0.070
GarageFinish_Unf	0.094	0.077	0.082	0.067
GarageFinish_RFn	0.082	0.065	0.080	0.059

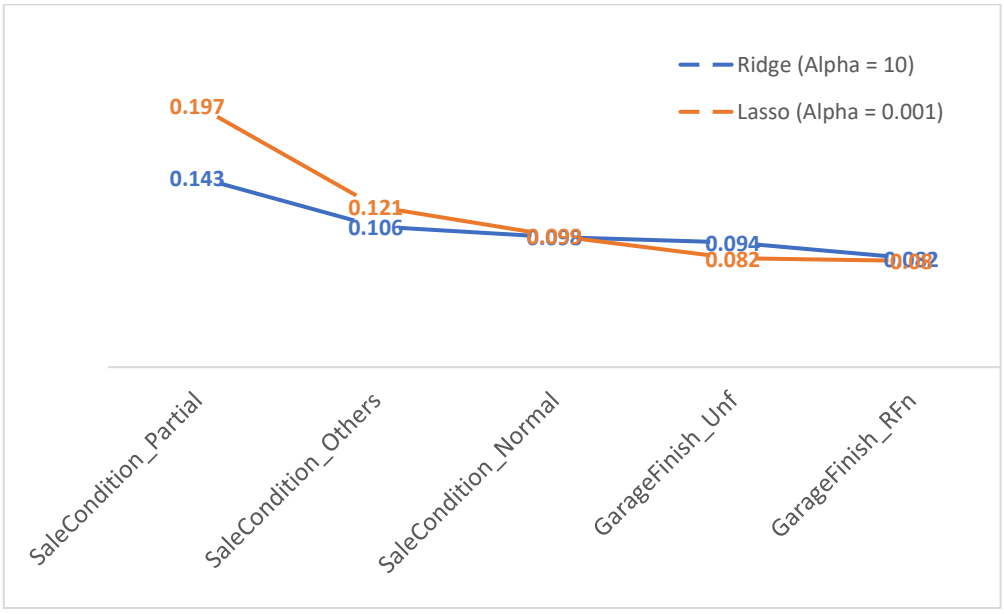


The above graph shows the trend in the difference value that we get when doubling the alpha values under both Ridge and Lasso regression.

c) The most important and significant predictor variable is SaleCondition_Partial.

Question 2: You determine the optimum value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Ans: Lasso regression helps us in feature selection. The below graph shows the coefficient values for the influencing features (top five).



From the graph, we notice that the coefficients are slightly higher in Lasso as compared to ridge. This minimises the risk of the model. Lasso regression is preferrable than ridge regression. However, the both the values are close to each other.

Question 3: After building the model, you realise 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: The next influencing variables after dropping the previously identified variables are:

- 'GarageFinish_No Garage' with coefficient value = 0.192
- 'GarageType_Others' with coefficient value = 0.119
- 'GarageType_No Garage' with coefficient value = 0.093
- 'GarageType_Detchd' with coefficient value = 0.084
- 'GarageType_BuiltIn' with coefficient value = 0.084

Note: We can find the previous data by refereeing the table in question 1.

Question 4: How can you make sure that a model is robust and generalisable? What are the implication of the same for the accuracy of the model and why?

Answer: The implications of having a robust and generalizable model and significant, as the notice very small difference between the predicted and the actual data (test data and train data). The model results are very much within the acceptable range as we have also done the hyper-parameter tuning. Hence we can say that the model is robust and generalizable.

Bias: The model is not biased and hence the model is acceptable. Bias means how well the model fits in to the training data.

Variance: This defines the sensitivity of the model even for a small change in the data. In the present model the variables are taken care in order to make the model overfit.