Advance Regression House Price Prediction

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 Alpha values as detected by my model is

- Ridge – 3.0
- Lasso – 100.0

If we increase the alpha value, coefficients tend to become less and less significant, the best values are chosen based on scoring method which is `lowest mean absolute error`, so error terms will increase by changing alpha values.

Applying alpha of 6.0 on ridge model and alpha 200 on Lasso model, gives below Score metrics

	Metric	Ridge Regression (alpha 3)	Ridge Regression (alpha 6)	Lasso Regression (alpha 100)	Lasso Regression (alpha 200)
0	R2 Score (Train)	9.314815e-01	9.261018e-01	9.244527e-01	9.160205e-01
1	R2 Score (Test)	9.119658e-01	9.093350e-01	9.123151e-01	9.068332e-01
2	RSS (Train)	3.254867e+11	3.510425e+11	3.588761e+11	3.989318e+11
3	RSS (Test)	2.034505e+11	2.095305e+11	2.026434e+11	2.153121e+11
4	RMSE (Train)	1.910223e+04	1.983798e+04	2.005810e+04	2.114788e+04
5	RMSE (Test)	2.304783e+04	2.338968e+04	2.300207e+04	2.371018e+04

We can see R2 score on both training and test model have reduced slightly in both Ridge and Lasso models using double alpha values.

After doubling the alpha, most important predictor variable still remains as `**GrLivArea**`, however we can see change in the predictor variables order down the line, and coefficients of all features have definetly changed.

Lasoo_predictors_alpha_100

	Feature Name	Coefficient	Absolute Coefficient
14	GrLivArea	163767.732454	163767.732454
3	OverallQual	68344.702927	68344.702927
11	TotalBsmtSF	61115.409910	61115.409910
8	BsmtFinSF1	42700.036245	42700.036245
5	YearBuilt	32117.606789	32117.606789
4	OverallCond	30908.014078	30908.014078
60	Neighborhood_NridgHt	25022.776939	25022.776939
130	ExterQual_TA	-24609.509546	24609.509546
66	Neighborhood_StoneBr	23754.287788	23754.287788
129	ExterQual_Gd	-21995.772736	21995.772736

Lasoo_predictors_alpha_200

	Feature Name	Coefficient	Absolute Coefficient
14	GrLivArea	165582.046101	165582.046101
3	OverallQual	71497.362545	71497.362545
11	TotalBsmtSF	60637.440024	60637.440024
8	BsmtFinSF1	39777.374491	39777.374491
60	Neighborhood_NridgHt	25290.810040	25290.810040
130	ExterQual_TA	-24015.407662	24015.407662
4	OverallCond	23972.994385	23972.994385
21	GarageCars	21897.670116	21897.670116
204	SaleCondition_Partial	20951.229435	20951.229435
5	YearBuilt	20821.417388	20821.417388

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:

Both models are showing very close R2 scores on test data prediction, however we can see Lasso scores are slightly better. So I will use Lasso model in this scenario.

Lasso has another advantage that it actually reduces coefficients to absolute 0, where as Ridge never eliminates the features (all coefficients are reduced but never to absolute 0 value)

So we can say Lasso model will be less complex due to low number of predictor variables.
We would be able to make predictions using fewer variables.

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:
Rebuiding the model without best 5 predictors, we get below 5 most important predictor variables

- 1stFlrSF
- 2ndFlrSF
- houseAge
- OverAllCond
- Neighborhood_ StoneBr

Best_Predictors_with_all_data

	Feature Name	Coefficient	Absolute Coefficient
14	GrLivArea	163767.732454	163767.732454
3	OverallQual	68344.702927	68344.702927
11	TotalBsmtSF	61115.409910	61115.409910
8	BsmtFinSF1	42700.036245	42700.036245
5	YearBuilt	32117.606789	32117.606789
4	OverallCond	30908.014078	30908.014078
60	Neighborhood_NridgHt	25022.776939	25022.776939
130	ExterQual_TA	-24609.509546	24609.509546
66	Neighborhood_StoneBr	23754.287788	23754.287788
129	ExterQual_Gd	-21995.772736	21995.772736

Next_best_Predictors_when_best_5_are_missing

	Feature Name	Coefficient	Absolute Coefficient
8	1stFlrSF	184786.684071	184786.684071
9	2ndFlrSF	98585.671229	98585.671229
23	houseAge	-42437.601326	42437.601326
3	OverallCond	36080.918671	36080.918671
61	Neighborhood_StoneBr	32435.437593	32435.437593
55	Neighborhood_NridgHt	31647.026049	31647.026049
125	ExterQual_TA	-31032.668538	31032.668538
142	BsmtExposure_Gd	26834.670833	26834.670833
16	GarageCars	26612.162758	26612.162758
124	ExterQual_Gd	-24657.439454	24657.439454

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?

Ans:
To make the model robust, we should

- reduce the complexity of model, use less predictor features
- make sure training data scores are as close to test data
- drop more correlated feature, and not relying on lasso only
- have more data available so that outliers treatment can be significant, here I have only dropped 0.5 to 0.95 Inter Quantile range, due to less data, ideally we should use 0.25 & 0.75 quantile range
- Apply regularization
- Apply bias, variance trade off rules