Subjective Questions

In any of the following questions, for more information, please refer to the last section of the notebook "Additional Questions".

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?

Best alpha for Ridge: 1
Best alpha for Lasso: 0

For Ridge
When we double the best alpha for Ridge,

	param_alpha	mean_test_score
0	0	0.870482
1	0.5	0.870526
2	1	0.870557
3	2	0.870587
4	3	0.870582
5	4	0.870549
6	5	0.870492
7	6	0.870415
8	7	0.870321
9	8	0.870212
10	9	0.870088
11	10	0.869953

R2-score for best alpha: 0.870587
R2-score for alpha=4: 0.870549
The R2-score decreases by a negligible amount.

As for the coefficients:

For the best alpha:

Coefficients:

	Feature	Coefficient
0	Neighborhood|NridgHt	0.132842
1	OverallCond|low_rat	-0.128209
2	Functional|other	-0.117582
3	BsmtExposure|Gd	0.114976
4	ExterQual|low_rating	-0.111992
5	OverallCond|ex_rat	0.110986
6	BldgType|other	-0.108265
7	MSZoning|RM	-0.102732
8	SaleCondition|Partial	0.097118
9	GarageQual|other_rating_r_no_rating	-0.089207
10	YearRemodAdd|2000-till date	0.068555
11	SaleCondition|other	-0.063730
12	Foundation|PConc	0.060890
13	Condition1|other	-0.053292
14	GarageType|Attchd	0.052201
15	FireplaceQu|low_rating	0.048199
16	HeatingQC|low_rating	-0.031838
17	YearBuilt|1950-1999	0.028826
18	Neighborhood|NAmes	-0.028714
19	OverallCond|gd_rat	0.022387
20	GrLivArea	0.000386
21	BsmtFinSF1	0.000146

For alpha=4:

Coefficients:

	Feature	Coefficient
0	Neighborhood|NridgHt	0.128265
1	OverallCond|low_rat	-0.125193
2	Functional|other	-0.115140
3	BsmtExposure|Gd	0.112005
4	ExterQual|low_rating	-0.111743
5	BldgType|other	-0.106074
6	MSZoning|RM	-0.101379
7	OverallCond|ex_rat	0.098516
8	SaleCondition|Partial	0.095495
9	GarageQual|other_rating_r_no_rating	-0.087641
10	YearRemodAdd|2000-till date	0.069169
11	SaleCondition|other	-0.063065
12	Foundation|PConc	0.060892
13	GarageType|Attchd	0.053045
14	Condition1|other	-0.052784
15	FireplaceQu|low_rating	0.048494
16	HeatingQC|low_rating	-0.032289
17	YearBuilt|1950-1999	0.027800
18	Neighborhood|NAmes	-0.027696
19	OverallCond|gd_rat	0.021805
20	GrLivArea	0.000387
21	BsmtFinSF1	0.000147

As can be observed, the magnitude changed a bit, but the order of their impact has not.

For Lasso
No change, since the best alpha for the same is 0.

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?

Ridge is preferred.

The reason being the best alpha value for Lasso is 0. This means that Lasso regularization is not helping the model.

But for Ridge the alpha value is 2, and it shows at least a slight improvement in R2-score compared to the model with no regularization.

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?

The 5 most important predictor variables where:

- Neighborhood|NridgHt
- OverallCond|low_rat
- Functional|other
- OverallCond|ex_rat
- BsmtExposure|Gd

Removing them and creating a new model, we get the following coefficients:

	Feature	Coefficient
0	ExterQual|low_rating	-0.131568
1	SaleCondition|Partial	0.123730
2	MSZoning|RM	-0.106017
3	BldgType|other	-0.092031
4	GarageQual|other_rating_r_no_rating	-0.085464
5	YearRemodAdd|2000-till date	0.080794
6	Foundation|PConc	0.076517
7	SaleCondition|other	-0.065185
8	FireplaceQu|low_rating	0.060912
9	GarageType|Attchd	0.058314
10	Condition1|other	-0.055383
11	OverallCond|gd_rat	0.036140
12	HeatingQC|low_rating	-0.033379
13	Neighborhood|NAmes	-0.023019
14	YearBuilt|1950-1999	0.021271
15	GrLivArea	0.000381
16	BsmtFinSF1	0.000178

As you can see, now the 5 most important predictor variables are:

- ExterQual|low_rating
- SaleCondition|Partial
- MSZoning|RM
- BldgType|other
- GarageQual|other_rating_r_no_rating

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?

To make a model robust and generalizable, it should explain the observed data very well with the lowest possible complexity.

The model becomes robust and generalizable when it ignores all the noise patterns in the data and explains the underlying pattern. If the model is more complex than needed, it also learns the noise in the data. This means that if a new pattern of noise is encountered in a new set of data, the complex model will perform poorly. Moreover, if we train a new model for the new set of data, it might look entirely different from the old model.

But if you had created a simpler model which only learned the underlying pattern, it would not be deterred by any kind of new noise patterns. And even if you must train for a new model, it will not be very different from the old one.

Make the simplest model that explains the data. That is how you build a robust and generalizable model.

The simpler models will have lower accuracy. You are trading complexity for accuracy. The best model is the one which has both in the right balance.