

SUBJECTIVE 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?

Answer 1

The optimal value for ridge is 0.001 and lasso is 2.

The impact of doubling the value of alpha is mentioned below:

- a) Ridge:
R square of train data has decreased and test data has increased. The error terms of test data has decreased and train data has increased.
- b) Lasso:
R square of train and test data has decreased. The error terms of train and test data has increased.

The most important predictor variables after the change is implemented are:

- a) Ridge:

	Variable	Coeff
96	OverallQual_Very Excellent	0.084
9	2ndFlrSF	0.082
10	GrLivArea	0.081
118	RoofMatl_WdShngl	0.068
92	OverallQual_Excellent	0.066
54	Neighborhood_NoRidge	0.059
8	1stFlrSF	0.059
0	constant	0.058
12	FullBath	0.045
19	GarageCars	0.043
16	TotRmsAbvGrd	0.039

b) Lasso:

	Variable	Coeff
0	constant	0.171
10	GrLivArea	0.139
19	GarageCars	0.061
17	Fireplaces	0.051
92	OverallQual_Excellent	0.037
54	Neighborhood_NoRidge	0.035
16	TotRmsAbvGrd	0.034
96	OverallQual_Very Excellent	0.027
97	OverallQual_Very Good	0.026
169	BsmtExposure_Gd	0.025
55	Neighborhood_NridgHt	0.021
173	BsmtFinType1_GLQ	0.015
159	Foundation_PConc	0.008
227	SaleType_New	0.008

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

To choose between ridge and lasso regression we need to compare the values of various parameters which are provided below in a tabular format.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.884844	0.909615	0.832930
1	R2 Score (Test)	0.760791	0.866818	0.821738
2	RSS (Train)	1.417260	1.112388	2.056180
3	RSS (Test)	1.303531	0.725754	0.971410
4	MSE (Train)	0.037257	0.033008	0.044876
5	MSE (Test)	0.054491	0.040660	0.047040

From the table above we can see that Ridge Regression has highest R-square score and minimum error terms. So we will choose Ridge Regression over Lasso Regression.

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

The top 5 predictor variables in the lasso model are listed below:

	Variable	Coeff
10	GrLivArea	0.230
0	constant	0.118
96	OverallQual_Very Excellent	0.080
92	OverallQual_Excellent	0.077
19	GarageCars	0.061
54	Neighborhood_NoRidge	0.047

As per the question the incoming data does not have the above 5 variables. So we have to create another model excluding the above 5 variables.

After creating the new model the top 5 important predictor variables are '1stFlrSF', '2ndFlrSF', 'GarageArea', 'Fireplaces' and 'TotRmsAbvGrd'. The table below provides the coefficients for these variables.

	Variable	Coeff
8	1stFlrSF	0.145
0	constant	0.135
9	2ndFlrSF	0.094
18	GarageArea	0.069
16	Fireplaces	0.052
15	TotRmsAbvGrd	0.042

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

A robust model is one where variation in the data does not have much impact on accuracy.

A generalisable model is one that can easily adapt to previously unseen data (test data).

To make sure a model is robust and generalisable, we have to consider the below points:

1. The model should not overfit. That is the model should not be complex so that it has high train accuracy but low test accuracy. It means that the model should not have high variance.
2. The model should not underfit. That is the model should not be so generalized that the line does not pass through most of the points. It means the model should not have high bias.

If we look at it from the perspective of Accuracy, a too robust model will have very high train accuracy but low test accuracy. If a model is highly generalizable then it will have low accuracy. So, to make our model more robust and generalizable, we have to strike some balance between model bias and variance. Addition of bias means that accuracy will decrease but it will provide the optimum solution.