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 regression is 4 and for lasso regression it is 0.0001.
After doubling the alpha values the R2 score has decreased.

Initial R2 score was:

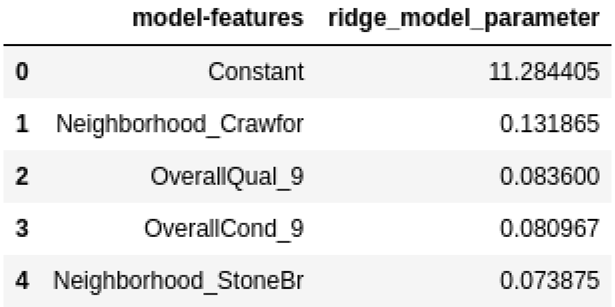
Ridge model: train set : 0.949 **Lasso model:** train set : 0.953
Ridge model: test set: 0.881 **Lasso model:** test set: 0.884

After doubling R2 score was:

Ridge model: train set : 0.945 **Lasso model:** train set : 0.950
Ridge model: test set: 0.880 **Lasso model:** test set: 0.882

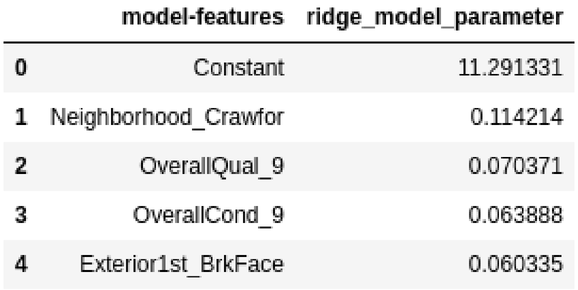
In ridge regression model, after implementing the change one of the top-5 model predictors changed. Previously it was Neighbourhood and after doubling it became Exterior1st.

Ridge optimal alpha coefficient



	model-features	ridge_model_parameter
0	Constant	11.284405
1	Neighborhood_Crawfor	0.131865
2	OverallQual_9	0.083600
3	OverallCond_9	0.080967
4	Neighborhood_StoneBr	0.073875

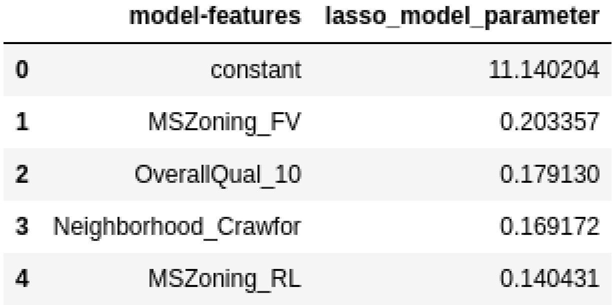
Ridge double alpha coefficient



	model-features	ridge_model_parameter
0	Constant	11.291331
1	Neighborhood_Crawfor	0.114214
2	OverallQual_9	0.070371
3	OverallCond_9	0.063888
4	Exterior1st_BrkFace	0.060335

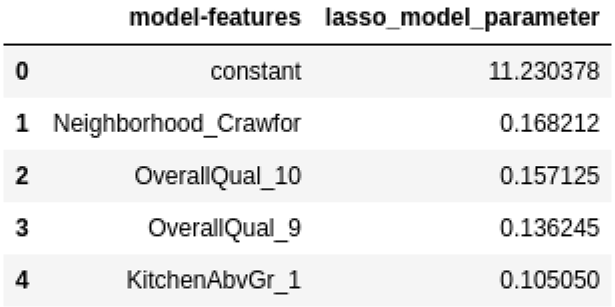
In ridge regression model, after implementing the change most of the top-5 model predictors changed. Photo attached for details.

Lasso optimal alpha coefficient



	model-features	lasso_model_parameter
0	constant	11.140204
1	MSZoning_FV	0.203357
2	OverallQual_10	0.179130
3	Neighborhood_Crawfor	0.169172
4	MSZoning_RL	0.140431

Lasso double alpha coefficient



	model-features	lasso_model_parameter
0	constant	11.230378
1	Neighborhood_Crawfor	0.168212
2	OverallQual_10	0.157125
3	OverallQual_9	0.136245
4	KitchenAbvGr_1	0.105050

The most important predictor variable for ridge model remains unchanged, but for lasso it changed from MSZoning to Neighborhood.

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:

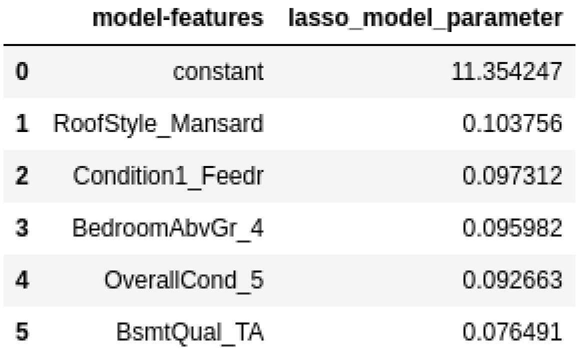
I will choose to apply the Lasso model. The reason for that is twofold, firstly, basis the R2 score the lasso model is tentatively performing better. Secondly, in the lasso model co-efficeint which were unimportant were driven to zero by the model hence giving a simplistic model.

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:

The new top-5 most important predictor variables now are : Roofstyle (Mansard), Condition1 (Feedr), BedroomAbvGr (4), OverallCond (5), BsmntQual (TA)



	model-features	lasso_model_parameter
0	constant	11.354247
1	RoofStyle_Mansard	0.103756
2	Condition1_Feedr	0.097312
3	BedroomAbvGr_4	0.095982
4	OverallCond_5	0.092663
5	BsmtQual_TA	0.076491

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

Robust means that a model can perform well on a broad set of inputs and generalisable means that the model also performs well on unseen data.

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. Along with that regularization helps by managing model complexity via addition of penalty terms in the model building process itself.

High accuracy on the train set and low accuracy on the test set means that the model has mugged up all the inputs and has thus overfitted thus the model is neither robust nor generalizable. The other case is of underfitting in which the model has low accuracy on both train and test set i.e. the model has not captured all the variance in the data.