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 alpha = 1.0

- Lasso alpha = 0.0001

- There's very little decrease in the R2 values (train: 0.003934 and test: 0.001582) after increasing the alpha.

	Metric	RidgeRegression	RidgeRegression - Double Alpha	LassoRegression	LassoRegression - Double Alpha
0	R2 Score (train)	0.896550	0.892616	0.891965	0.892616
1	R2 Score (test)	0.898744	0.897162	0.897623	0.897162
2	RSS (train)	1.734025	1.799962	1.810880	1.799962
3	RSS (test)	0.796095	0.808531	0.804912	0.808531
4	RMSE (train)	0.041191	0.041967	0.042094	0.041967
5	RMSE (test)	0.042633	0.042965	0.042868	0.042965

Top predictors:

	Ridge(2.0)
OverallQual	0.221143
GrLivArea	0.219838
TotalBsmtSF	0.133826
OverallCond	0.122564
GarageCars	0.096789
LotArea	0.058411
MSZoning_FV	0.052945
RoofMatl_WdShngl	0.047310
RoofMatl_CompShg	0.036082
MSZoning_RL	0.030856

Lasso:

- There’s very little decrease in the R2 values (train: 0.000651and very little increase in test: 0.000461) after increasing the alpha.

- Top predictors:

	Lasso(0.0002)
OverallQual	0.240348
GrLivArea	0.222429
TotalBsmtSF	0.134366
OverallCond	0.128366
GarageCars	0.091773
LotArea	0.051744
MSZoning_FV	0.031868
MSZoning_RL	0.013791
RoofMatl_CompShg	0.006227

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:

Ridge & Lasso models have very little difference in values of R-squared in both test & train data sets. If Compare Lasso and Ridge in values of RMSE and RSS, Ridge is better fit.

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:

Before drop:

L
OverallQual
GrLivArea
TotalBsmtSF
OverallCond
GarageCars

After drop:

LotArea
RoofMatl_WdShngl
GarageCond_Po
MSZoning_FV
BsmtUnfSF

After drop top 5

	Metric	RidgeRegression	RidgeRegression - Drop Top5	LassoRegression	LassoRegression - Drop Top5
0	R2 Score (train)	0.896550	0.558139	0.891965	0.556541
1	R2 Score (test)	0.898744	0.538138	0.897623	0.528765
2	RSS (train)	1.734025	7.406474	1.810880	7.433253
3	RSS (test)	0.796095	3.631261	0.804912	3.704955
4	RMSE (train)	0.041191	0.085130	0.042094	0.085283
5	RMSE (test)	0.042633	0.091052	0.042868	0.091972

Question 4

Answer:

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:

A model is robust when any variation in the data does not affect its performance much.

A generalizable model is able to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.

To make sure a model is robust and generalizable, we have to take care it doesn't overfit. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data.

In other words, the model should not be too complex in order to be robust and generalizable.

If we look at it from the perspective of Accuracy, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.

In general, we have to find strike some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge Regression and Lasso.