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) optimal value of alpha is 5 for both ridge regression
2) Lasso regression optimal value is – 0.0009
3) After doubling the alpha values R2 score of Ridge regression is 0.85 and Lasso regression R2 value is 0.86

After changes are implemented, there are few dissimilarities in ridge coefficients but there is no change in top influencers.

	Ridge Coef		Ridge Coef - DOUBLE ALPHA
OverallQual	0.407	OverallQual	0.322
GrLivArea	0.256	GrLivArea	0.199
TotRmsAbvGrd	0.202	TotRmsAbvGrd	0.192
FullBath	0.176	FullBath	0.171
GarageCars	0.169	GarageCars	0.154
BsmtFullBath	0.169	BsmtFullBath	0.145
OverallCond	0.159	OverallCond	0.123
BedroomAbvGr	0.113	GarageArea	0.115
GarageArea	0.109	BedroomAbvGr	0.109
TotalBsmtSF	0.108	TotalBsmtSF	0.095
TotalPorchAreaSF	0.097	CentralAir_enc	0.089
TotalBasementFinAreaSF	0.088	TotalPorchAreaSF	0.085
LandSlope_enc	0.087	Fireplaces	0.083
CentralAir_enc	0.084	HalfBath	0.081
WoodDeckSF	0.08	WoodDeckSF	0.076
LotArea	0.08	TotalBasementFinAreaSF	0.075
HalfBath	0.078	BsmtUnfSF	0.073
BsmtUnfSF	0.077	LandSlope_enc	0.069
SaleCondition_enc	0.074	SaleCondition_enc	0.067
Fireplaces	0.069	MasVnrArea	0.063

Similarly there are no noticeable changes in lasso model results

Column1	Lasso Coef
OverallQual	0.641
GrLivArea	0.636
GarageCars	0.222
BsmtFullBath	0.168
TotRmsAbvGrd	0.153
OverallCond	0.15
FullBath	0.127
CentralAir_enc	0.093
LandSlope_enc	0.067
SaleCondition_enc	0.064
WoodDeckSF	0.055
PavedDrive_enc	0.041
HalfBath	0.041
TotalPorchAreaSF	0.033
Fireplaces	0.03
Fence_enc	0.03
BedroomAbvGr	0.03
GarageArea	0.029
LandContour_enc	0.026
BsmtUnfSF	0.025

	Lasso Coef - DOUBLE ALPHA
OverallQual	0.669
GrLivArea	0.495
GarageCars	0.221
TotRmsAbvGrd	0.184
BsmtFullBath	0.136
FullBath	0.129
CentralAir_enc	0.102
OverallCond	0.072
SaleCondition_enc	0.049
Fireplaces	0.041
HalfBath	0.04
WoodDeckSF	0.037
PavedDrive_enc	0.036
GarageArea	0.027
Fence_enc	0.019
MasVnrArea	0.016
LandSlope_enc	0.005
RoofStyle_enc	0.005
KitchenAbvGr	0
Heating_enc	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?

The best lambda values for ridge and lasso regression are 5 and 0.0009

Mean square error of Ridge and lasso regression are: 0.019 and 0.018 and it's very minimal.

The R2 score of two models is almost same. So, I prefer lasso regression because we can use this for feature selection.

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 top five important predictor variables in lasso model are:

TotalBsmtSF

TotRmsAbvGrd

GarageArea

FullBath

CentralAir_enc

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?

We will check the mode robustness by measuring the overfit against the test data.

Make sure model should be simple enough and understand the core pattens of test data instead of memorizing it .

We can use VIF to check multicollinearity in the predictor variables. Makes sure VIF factor < 5 (Depends on problem) so that model will not become complex with multilinearity data. If any such multi colinear attributes are there better to drop them