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 regularization model is built by removing the multi-collinearity features from the data set.

Different models are built like taking all the features post removal for multicollinearity, using RFE and selecting different numbers of feature selection. On all the models the model A (build taking all the features) comes out be the best one both for Ridge and Lasso

1. The optimal value **alpha** for **Ridge regression** is **2**
2. The optimal value **alpha** for **LASSO regression** is **100**

When we double the alpha values as derived above –

a. For Ridge regularization

RIDGE	Train r2 score	Test r2 score	predictor variables
alpha -2	0.9483	0.8926	GrLivArea OverallQual TotalBsmtSF Neighborhood_StoneBr LotArea YearBuilt
double alpha -4	0.9437	0.8946	GrLivArea OverallQual TotalBsmt Neighborhood_StoneBr LotArea GarageArea

No such significant difference has been observed in Ridge Regression

The r2 score for both train and test set is nearly same.

b. For Lass regularization

LASSO	Train r2 score	Test r2 score	predictor variables	Total features
alpha -100	0.9391	0.9088	GrLivArea OverallQual TotalBsmtSF YearBuilt Neighborhood_StoneBr BsmtFinSF1	113
double alpha - 200	0.9299	0.9057	GrLivArea OverallQual TotalBsmtSF BsmtFinSF1 SaleType_New Neighborhood_StoneBr	80

The feature selection has been reduced to 8 when alpha is doubled

R2 score for both train and test are reduced a bit

The important predictor for double alpha is –

For Lasso: **GrLivArea hold the most critical predictor**

Features	Coefficients
GrLivArea	143677.7308
OverallQual_Very Excellent	72510.02131
OverallQual_Excellent	70609.93302
TotalBsmtSF	49834.30396
OverallQual_Very Good	30687.77902
BsmtFinSF1	27846.30744
SaleType_New	24692.083
Neighborhood_StoneBr	22515.97474
GarageArea	21414.38194
LotArea	17649.26023

Foe Ridge : **GrLivArea hold the most critical predictor**

Features	Coefficients
GrLivArea	70853.4449
OverallQual_Excellent	49865.57612
OverallQual_Very Excellent	47381.66491
TotalBsmtSF	46903.82524
BsmtFinSF1	37928.83433

Neighborhood_StoneBr	31851.60119
LotArea	25572.1849
GarageArea	23901.04977
OverallQual_Very Good	23341.87713
SaleType_New	23092.25569

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:

Lasso Regression is selected over Ridge

a. Total number of features post modeling is 269 in case of Ridge and 113 in case of Lasso. Lasso able to remove the high collinearity features and removes the features having no impact
b. The measuring metrics r2 score for training and test set data is better for Lasso as compared to Ridge. The difference of r2 score in train and test data is less in Lasso.
    a. r2 score for ridge for train and test data is 0.9483 and 0.8926
    b. r2 score for Lasso for train and test is 0.9391 and 0.9088
c. The computation time also reduces for the model as the features and the coefficients value are reduced as compared to Ridge

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 5 most important predictor in the lasso model is - **'GrLivArea', 'OverallQual_Very Excellent', 'Overall Qual_Excellent', 'TotalBsmtSF' and 'YearBuilt'**

Removing these features from the data set of model input and rebuilding the lasso model

1. Optimal value of alpha is 50

2. New important features –
    a. **BsmtFinSF1**
    b. **BsmtUnfSF**
    c. **Neighborhood_StoneBr**
    d. **2ndFlrSF**
    e. **BsmtFinSF2**

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:
For a model to be more generic and robust it needs to be simple and not complex. By simple it means mi nimum number of driving factors that can provide an impact on the decision.

1. The model may not trace all the data points in the training data set however it has a higher chan ce of detecting the test data.
2. There must be a good trade off between bias and trade off
3. The model should not be too simple else it will fall under high bias and high variance. Not able to perform both on training and test data set
4. Regularization can help us to achieve a simpler model

The robust model has a good accuracy value which satisfies a trade off between bias and variance.
The complex model tends to have a good accuracy score however it tends to overfit the model and fail to predict the test data.