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 and lasso regression

Ridge Alpha 2
lasso Alpha 100

R2 score Ridge when Alpha value is 2

R2score(train) 8.796016e-01
R2score(test) 8.311042e-01

R2 score Ridge when Alpha value is 4
R2score(train) 8.751450261474036
R2score(test) 8.322099713928739

lasso Alpha 100
lasso R2 score when Alpha value is 100

R2score(train) 8.759332e-01
R2score(test) 8.328141e-01

Lasso R2 score when Alpha value is 200

R2score(train) 8.736374064073695
R2score(test) 8.359991365566337

R2score of training data has decreased and it has increased on testing data for both Lasso and

important predictor variables

1. **GrLivArea**
2. **OverallQual**
3. **YearBuilt**
4. **TotalBsmtSF**
5. **OverallCond**
6. **LotArea**
7. **BsmtFinSF1**
8. **RoofMatl_CompShg**
9. **SaleType_New**
10. **Exterior2nd_CmentBd**

Predictors are same but the coefficient of these predictor has changed

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:

The r2_score of lasso is slightly higher than lasso for the test dataset so we will choose lasso regression to solve this problem

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and making the model interpretably.

Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude ofcoefficients which is identified by cross validation. Residual sum or squares should be small by using thepenalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that havegreater values gets penalized. As we increase the value of lambda the variance in model is dropped andbias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitudeof coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection.When lambda value is small it performs simple linear regression and as lambda value increases,shrinkage takes place and variables with 0 value are neglected by the model.

Question 3

After building the model, you realized 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:

LotArea,OverallQual,YearBuilt,BsmtFinSF1,TotalBsmtSF are the top 5 important predictors.

Before removing

	Lasso
GrLivArea	146491.5579
OverallQual	127507.0651
YearBuilt	58277.31106
TotalBsmtSF	46920.16529
OverallCond	46579.34058

Therefore these columns are dropped.

	without removing top 5 predictors	after removing top 5 predictors
R2 Train Score	8.759331	6.44795203
R2 Test Score	8.32814108	6.06293914

Those 5 most important predictor variables that will be excluded are :-

After Removing the top 5 indicators

	Lasso21
LotArea	107752.7979
BsmtFinSF1	58876.69354
SaleType_New	38842.56552
RoofMatl_WdShake	30884.88119
Condition2_PosA	16037.23173

five most important predictor variables

1. LotArea
2. BsmtFinSF1

3. SaleType_New
4. RoofMatl_WdShake
5. Condition2_PosA

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:

In order to ensure that the test accuracy is equal to the training score, the model needs to be generalized. When applied to datasets other than the ones used for training, the model ought to yield reliable results. To ensure that the model's predicted accuracy is high, the outliers shouldn't be given undue weight. Only those outliers that are pertinent to the dataset should be kept after an outliers analysis has been completed to make sure this is not the case. The dataset has to have the outliers that don't make sense kept eliminated. A model cannot be relied upon for predictive analysis if it lacks robustness.

The simpler the model, the more resilient and generalizable it will be, even if its accuracy will decline. The Bias-Variance trade-off can also be used to understand it. There is more bias but less variation and greater generalizability in simpler models. It implies that a reliable and generalizable model will function similarly on training and test data, meaning that accuracy will not significantly vary between the two sets of data.

Bias: When a model is unable to adequately learn from the data, it exhibits mistake. When a model has a high bias, it cannot extract information from the data. The model does not perform well on training or testing data.
.
Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.
It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data