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:

- ➔ Optimal values for Ridge regression is 6.0
- ➔ Optimal values for Lasso Regression is 0.001

In case of Ridge Regression when the values of alpha has been doubled coefficient values has been changed as shown below ,Due to which model predictors will get impacted and perfomance decreases.

	Coef
GrLivArea	0.199889
AgeOfBuilding	0.179869
MSSubClass_1-STORY 1946 & NEWER ALL STYLES	0.142045
MSSubClass_2-STORY PUD - 1946 & NEWER	0.132949
MSSubClass_PUD - MULTILEVEL - INCL SPLIT LEV/FOYER	0.122511

➔

In case of Lasso Regression when the values of alpha has been doubled some more coefficient values becomes 0 as shown below ,Due to which model predictors will get impacted and perfomance decreases on unseen data and model may not be generalisable.

Neighborhood_Somerst	-0.000000e+00
Neighborhood_StoneBr	-0.000000e+00
Neighborhood_Timber	-0.000000e+00
Neighborhood_Veenker	0.000000e+00
Condition1_Norm	0.000000e+00
Condition1_RRAn	0.000000e+00
BldgType_Twnhs	0.000000e+00
OverallQual_Average	0.000000e+00
OverallQual_Below Average	-0.000000e+00
OverallQual_Excellent	-0.000000e+00
OverallQual_Fair	-0.000000e+00
OverallQual_Good	-0.000000e+00
OverallQual_Poor	-0.000000e+00
OverallQual_Very Good	-0.000000e+00
OverallQual_Very Poor	-0.000000e+00
OverallCond_Average	0.000000e+00
OverallCond_Below Average	-0.000000e+00
OverallCond_Fair	-0.000000e+00
Exterior1st_BrkFace	0.000000e+00

->When the alpha is doubled in both Ridge and Lasso Top 5 preditor varibales remains same but there coefficient values has changed in this dataset but this may not be the same with all the datasets.

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:

Below are the optimal values obtained from Ridge and Lasso

(a) Optimal lambda value for Ridge : 6.0

(b)Optimal lambda value for lasso : 0.001

And the metrics we obtained are:

1. Ridge regression:

(a) Training: 0.87

(b) Testing: 0.87

2.lasso regression:

(a) Training: 0.866

(b) Testing: 0.87

We have got good results by using both ridge and Lasso but I prefer to go with lasso Regression since the unwanted or less important variables are given with 0 coefficients resulting in better performance of the model on the unseen data.

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:

Below code is used in the notebook to remove the Top 5 features:

newdf.drop(['GrLivArea','AgeOfBuilding','MSSubClass_1-STORY 1946 & NEWER ALL STYLES','MSSubClass_2-STORY PUD - 1946 & NEWER','MSSubClass_PUD - MULTILEVEL - INCL SPLIT LEV/FOYER'],axis=1,inplace=True)

Ridge: We have obtained the below Top 5 new features using ridge Regression, which will be helpful in making house price predictions when we eliminate the earlier Top 5 features from the data.

```
## Top 5 features using Ridge
RidgeCoeff.head(5)
```

	Coef
BsmtFullBath	0.279486
FullBath	0.212750
HalfBath	0.211996
GarageCars	0.211607
MSSubClass_1-STORY 1945 & OLDER	0.167077

Lasso: We have obtained the below Top 5 new features using Lasso Regression, which will be helpful in making house price predictions when we eliminate the earlier Top 5 features from the data.

```
## Top 5 features using lasso
LassoCoeff.head(5)
```

	Coef
BsmtFullBath	0.306598
FullBath	0.255222
HalfBath	0.220663
GarageCars	0.220306
MSSubClass_1-STORY 1945 & OLDER	0.177215

As mentioned above both Ridge and Lasso are giving same features with high coefficient in the model: ‘bsmtFullbath’,’FullBath’,’HalfBath’,’GarageCars’,’Ms Subcalass_1-story 1945 & OLDEER’

Question 4

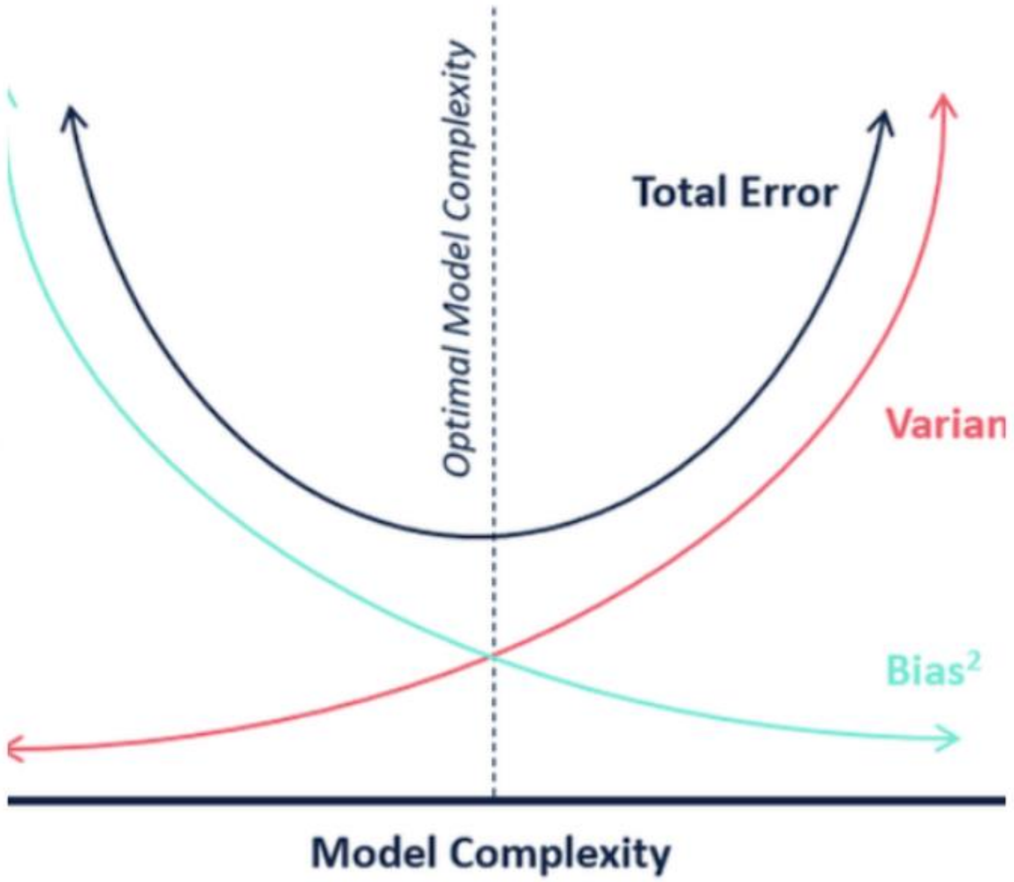
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

Lets see two major things to achieve generalizable model

Complex Models: If the model is not generalizable It will provide better results in Training data and results will drop when it comes to unseen data, This type of models are called high complexity models in which model complexity is more and even noise in the data is captured in training phase resulting in dropping the accuracy of Testing/unseen data. This type of models are not realistic and may not be used since it results in high error rate on unseen data. These type of models are low bias and high variance models

Lenient/Simple Models: If the model is too simple it will produce high error rate and has high bias low variance. These models will not perform well on training and even testing results are not accurate.



There are multiple ways to make a model more generalizable and suitable for different type of data:

As explained above we cannot stick to either Complex models or Simple models, Since they cannot generalise the results. So we need a trade-off to manage both,i.e., Models with low bias and low variance with less error rate in both training and test data/unseen data and modes are built by tanning them in such a way that this trade-off is achieved and provides a generalizable model which provides better results on unseen data as well.

Implications: If the model is robust and generalizable we can be confident on the results we obtain in Training, And model will give good accuracy on the unseen data which is important in most of the real world models and acceptable across industry standards.