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

Ridge - Optimal Alpha: 1
Lasso - Optimal Alpha: 0.0005

After doubling the alpha for both
Ridge: R-sq drops by 1.5% in training dataset while 0.5 in test dataset, hence accuracy drops.
Lasso: R-sq drops by 2% in training dataset while 1.5 in test dataset, hence accuracy drops.

Ridge: First five predictors remain the same. Others change order.
Lasso:
Before

Positive Factors: The 5 most positively contributing factors to the price of a property

```
In [476]: df.head()
```

Out[476]:

	Coeff	Predictor
0	0.3062	GrLivArea
1	0.1064	TotalBsmtSF
2	0.1049	OverallQual_9
3	0.0938	OverallQual_10
4	0.0721	BsmtFinSF1

Negative Factors: The 5 most negatively contributing factors to the price of a property

```
In [477]: df.tail()
```

Out[477]:

	Coeff	Predictor
34	-0.0200	ExterQual_TA
35	-0.0207	OverallCond_3
36	-0.0208	BsmtQual_Gd
37	-0.0227	BsmtQual_TA
38	-0.0448	HouseAge

After

Positive Factors: The 5 most positively contributing factors to the price of a property

```
df.head()
```

	Coeff	Predictor
0	0.2858	GrLivArea
1	0.1111	TotalBsmtSF
2	0.0887	OverallQual_9
3	0.0788	BsmtFinSF1
4	0.0546	GarageArea

Negative Factors: The 5 most negatively contributing factors to the price of a property

```
df.tail()
```

	Coeff	Predictor
22	-0.0121	MSZoning_RM
23	-0.0168	BsmtQual_Gd
24	-0.0181	ExterQual_TA
25	-0.0206	BsmtQual_TA
26	-0.0243	FireplaceQu_NA

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

In real world assignment Lasso is always preferred, since it can assign 0 to coefficients, but if the penalty on the coefficients is high enough, Ridge can also be used. In our case both models provide similar R-sq and values for other metrics. Here Lasso can be better as penalty is quite small and most coefficients are 0, hence complexity of the model decreases.

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

Five most important right now are:
Positively:
0 0.3062 GrLivArea
1 0.1064 TotalBsmtSF
2 0.1049 OverallQual_9

3 0.0938 OverallQual_10
4 0.0721 BsmtFinSF1
Negatively:
34 -0.0200 ExterQual_TA
35 -0.0207 OverallCond_3
36 -0.0208 BsmtQual_Gd
37 -0.0227 BsmtQual_TA
38 -0.0448 HouseAge

After removing them and recalculating the model, these are the 5 most important variables:
Positively:
0 0.2920 1stFlrSF
1 0.1252 2ndFlrSF
2 0.0581 GarageArea
3 0.0546 GarageCars_3
4 0.0459 Neighborhood_StoneBr

Negatively:
22 -0.0204 KitchenQual_Gd
23 -0.0214 BsmtQual_Gd
24 -0.0265 KitchenQual_TA
25 -0.0265 FireplaceQu_NA
26 -0.0308 ExterQual_TA

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

1. Data Cleaning and preparation of derived variables based on business domain. This is very important as the model is built for a specific usecase and sometimes a derived variable can play a very important role in prediction, as seen above with HouseAge.
2. Model is trained using KFold mechanism inorder that all parts of data are counted as training and test datasets atleast once, so that robustness and generalisability increases.
3. Feature elimination is done properly so that slight variation in features does not affect the model much and it can easily generalized for more data. This is done so that **Variance** remains in check.
4. Optimum number of features need to be selected otherwise model will not be robust and will have high **Bias.**

Variance increases the fluctuations of the model to very slight change in test data. So there can be very high accuracy in training data but it will lose it on test data.

Bias increases the simplicity of the model and if not kept in check, high bias will make the model too simple to predict the patterns in the data.

Hence it is essential to keep these in check for a model to be Robust and Generalized.