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 of alpha:

Ridge: 0.2

Lasso: 50

Metrics for Ridge regression:

Train data:

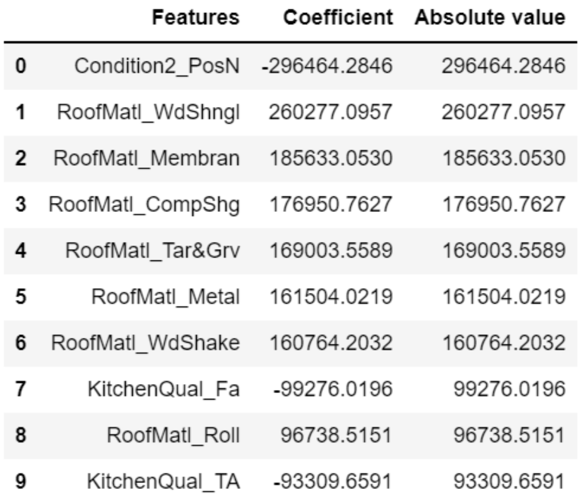
	Alpha = 0.2	Alpha = 0.4
R-Squared value	0.8074	0.7894
MSE	1203436688	1316250420
RMSE	34690	36280

Test data:

	Alpha = 0.2	Alpha = 0.4
R-Squared value	0.7524	0.7623
MSE	1593204589	1529950810
RMSE	39914	39114

Observation: There has been slight increase in R-Squared, MSE, RMSE values after doubling the alpha value.

Most important predictor variables after doubling alpha value:



	Features	Coefficient	Absolute value
0	Condition2_PosN	-296464.2846	296464.2846
1	RoofMatl_WdShngl	260277.0957	260277.0957
2	RoofMatl_Membran	185633.0530	185633.0530
3	RoofMatl_CompShg	176950.7627	176950.7627
4	RoofMatl_Tar&Grv	169003.5589	169003.5589
5	RoofMatl_Metal	161504.0219	161504.0219
6	RoofMatl_WdShake	160764.2032	160764.2032
7	KitchenQual_Fa	-99276.0196	99276.0196
8	RoofMatl_Roll	96738.5151	96738.5151
9	KitchenQual_TA	-93309.6591	93309.6591

Metrics for Lasso regression:

Train data:

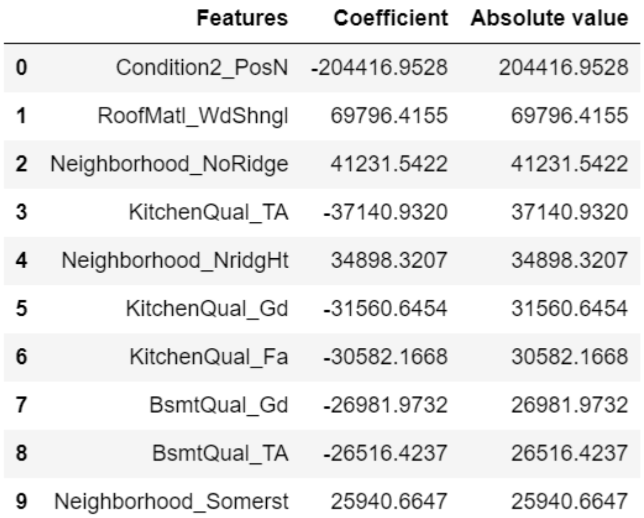
	Alpha = 50	Alpha = 100
R-Squared value	0.9062	0.8957
MSE	586253349	652057881
RMSE	24212	25535

Test data:

	Alpha = 50	Alpha = 100
R-Squared value	0.8330	0.8428
MSE	1074293069	1011280590
RMSE	32776	31800

Observation: Slight increase and decrease can be seen the above values after doubling the alpha value.

Most important predictor variables after doubling alpha value:



	Features	Coefficient	Absolute value
0	Condition2_PosN	-204416.9528	204416.9528
1	RoofMatl_WdShngl	69796.4155	69796.4155
2	Neighborhood_NoRidge	41231.5422	41231.5422
3	KitchenQual_TA	-37140.9320	37140.9320
4	Neighborhood_NridgHt	34898.3207	34898.3207
5	KitchenQual_Gd	-31560.6454	31560.6454
6	KitchenQual_Fa	-30582.1668	30582.1668
7	BsmtQual_Gd	-26981.9732	26981.9732
8	BsmtQual_TA	-26516.4237	26516.4237
9	Neighborhood_Somerst	25940.6647	25940.6647

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:

Optimal Values of alpha:

Ridge: 0.2

Lasso: 50

Metrics for Ridge regression:

Train data:

	Alpha = 0.2	Alpha = 0.4
R-Squared value	0.8074	0.7894
MSE	1203436688	1316250420
RMSE	34690	36280

Test data:

	Alpha = 0.2	Alpha = 0.4
R-Squared value	0.7524	0.7623
MSE	1593204589	1529950810
RMSE	39914	39114

Metrics for Lasso regression:

Train data:

	Alpha = 50	Alpha = 100
R-Squared value	0.9062	0.8957
MSE	586253349	652057881
RMSE	24212	25535

Test data:

	Alpha = 50	Alpha = 100
R-Squared value	0.8330	0.8428
MSE	1074293069	1011280590
RMSE	32776	31800

Observation: Due to improper methods of feature elimination, I have got unusual values for the above metrics.

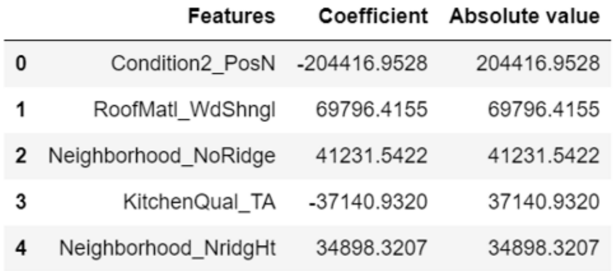
Although, if performed properly, I believe Lasso regression will render a good model as it involves feature selection in the process and gives good R-Squared value, MSE and RMSE values compared to Ridge regression. It is said to perform better on unseen data and therefore it has a higher hand over ridge regression. Hence, Lasso regression becomes the better choice to predict the price of houses.

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?

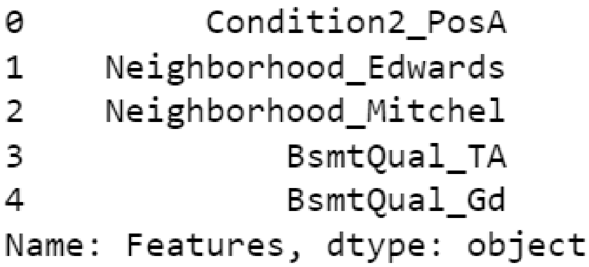
Answer:

Initial Top 5 variables :



	Features	Coefficient	Absolute value
0	Condition2_PosN	-204416.9528	204416.9528
1	RoofMatl_WdShngl	69796.4155	69796.4155
2	Neighborhood_NoRidge	41231.5422	41231.5422
3	KitchenQual_TA	-37140.9320	37140.9320
4	Neighborhood_NridgHt	34898.3207	34898.3207

Top 5 variables after dropping the above one and building the model:



```
0          Condition2_PosA
1     Neighborhood_Edwards
2     Neighborhood_Mitchel
3              BsmtQual_TA
4              BsmtQual_Gd
Name: Features, dtype: object
```

Question 4

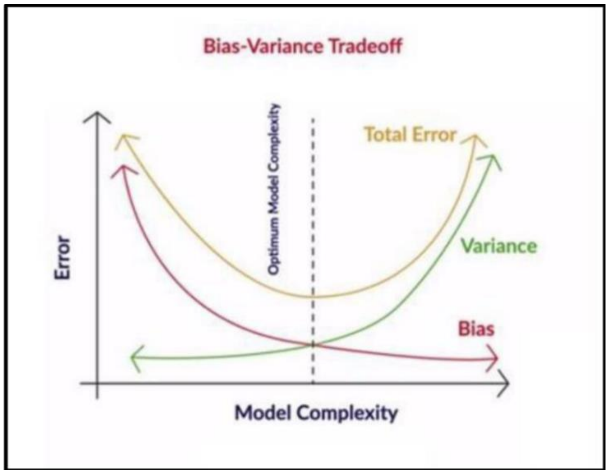
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

The robustness and generalizability of the model is the measure of how much the model can relate with unseen data and the accuracy with which it predicts. A simple model is usually expected to be robust as it learns the underlying patterns in the data instead of memorizing the training data which leads to overfitting.

Regularization is one approach which cuts down the coefficients using different methods making the model simpler. It makes sure that the model has ideal level of complexity. We need to make sure that the model is not too naïve nor highly complex. It should be balanced.

Bis-variance trade off is also a good visualizer to understand the data and how it is behaving when fitted to the model. The model is said to be optimum when the bias, variance and total errors hit the optimal value making the model balanced.


Bias-Variance Tradeoff
Optimum Model Complexity
Total Error
Variance
Bias
Error
Model Complexity