House Price Prediction Assignment Subjective Questions

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 regression is 3 and for Lasso Regression is 0.001
- For Lasso, on doubling the alpha value, the test set R-Squared goes down from 88.02% to 86.58%
- For Ridge, on doubling the alpha value the test set R-Squared goes down from 87.71% to 87.18%

Lasso

Before

	Feature	Coef
0	constant	10.700
4	OverallQual	0.804
9	GrLivArea	0.672
14	GarageCars	0.335
13	TotRmsAbvGrd	0.213
5	OverallCond	0.200
11	FullBath	0.180
10	BsmtFullBath	0.169
32	Neighborhood_StoneBr	0.123
54	CentralAir_Y	0.111

After

	Feature	Coef
0	constant	10.754
4	OverallQual	0.885
9	GrLivArea	0.605
14	GarageCars	0.356
13	TotRmsAbvGrd	0.187
11	FullBath	0.167
10	BsmtFullBath	0.143
5	OverallCond	0.142
54	CentralAir_Y	0.121
15	WoodDeckSF	0.091

Ridge

	Feature	Coef
0	constant	11.213
4	OverallQual	0.313
17	ScreenPorch	0.218
12	HalfBath	0.209
19	Alley_Pave	0.169
5	OverallCond	0.162
11	FullBath	0.149
16	EnclosedPorch	0.136
14	GarageCars	0.133
51	BsmtFinType1_Unf	0.129

	Feature	Coef
0	constant	10.810
4	OverallQual	0.505
9	GrLivArea	0.318
13	TotRmsAbvGrd	0.302
14	GarageCars	0.287
11	FullBath	0.234
5	OverallCond	0.203
7	TotalBsmtSF	0.176
32	Neighborhood_StoneBr	0.160
10	BsmtFullBath	0.145

- There is a slight change in the important predictor variables in case of Lasso, but there are noticeable differences in case of Ridge in terms of important predictor variables.

House Price Prediction Assignment Subjective Questions

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:

Although we found the optimal values of alpha for both Ridge and Lasso, Lasso produces the best score on the test set. Moreover, the difference in R2 Score between training and testing is not too much and it can also help in feature selection.

```
Training Lasso Model with optimal alpha......

Results: {'model': 'Lasso', 'training_r2': 0.884072788407257, 'testing_r2': 0.8802728184834513}
```

```
Training Ridge Model with optimal alpha......

Results: {'model': 'Ridge', 'training_r2': 0.8955799622587268, 'testing_r2': 0.8771027545882633}
```

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:

After rebuilding the model without the top 5 predictor variables, the test score now decreased to 81.77 %

Now the top variables are **TotalBsmtSF, FullBath, 2ndFlrSF, Neighborhood_StoneBr, Neighborhood_NridgHt.**

Question 4

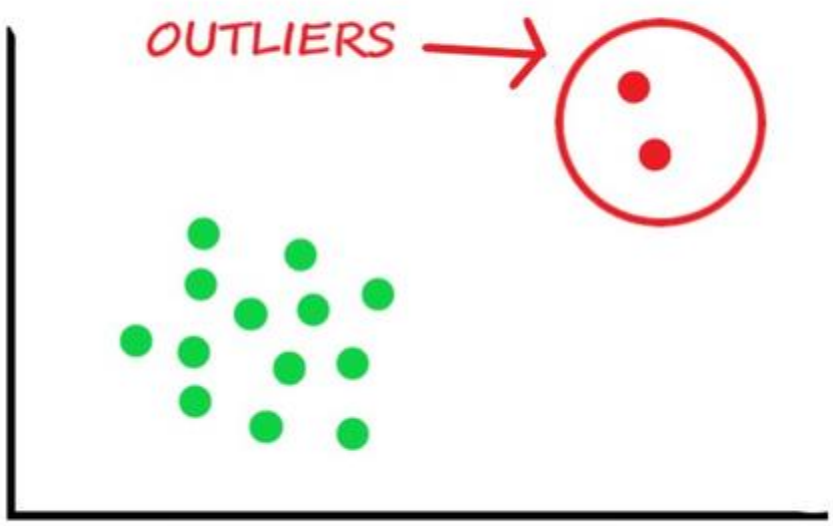
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

House Price Prediction Assignment Subjective Questions

Answer:

When we say a model is robust, it means that it needs to be immune to outliers. When we train on data that has outliers, it will skew the results of the model. It can potentially harm the predictive power of the model.

Outliers are the points that are distanced from other observations.



Now, when we try to fit a best fit line through the points, the coefficients will be much smaller in case of a regularized regression as it will try to penalize for the high error.

This can make the coefficients of the model unreliable hence making the predictions on unseen data inaccurate. This might make someone think that deleting outliers would be the right strategy, but we must treat them carefully instead. Now, Outliers can be misread observations or can be intended.

We can use several methods to detect outliers like using a Box plot or a Scatter plot. After detecting we can use any suitable methods like Inter Quartile Range or Z-Score (Standard Deviation) method to remove outliers. We can even replace the outliers with a more suitable 99^{th} percentile or 25^{th} percentile value.

This will help in increasing the prediction power of the model on unseen data as well. Hence making it more generalizable.

We can also use regularization techniques of Ridge and Lasso that penalize model for overfitting the training data also minimizing the effects of outliers.