Surprise Housing Assignment

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 Value of alpha

- Optimal Value of alpha for Ridge: 10
- Optimal Value of alpha for Lasso: 0.001

If we choose to double the value of alpha:

- The coefficients value will become lower for Ridge.
- The less valued feature coefficients will become zero for Lasso.

Most important predictor variables remain the same.

- 'SaleCondition_Partial'
- 'SaleCondition_Others'
- 'SaleCondition_Normal'
- 'GarageFinish_Unf'
- 'GarageFinish_RFn'

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 Value of lambda

- Optimal Value of lambda for Ridge: 10
- Optimal Value of lambda for Lasso: 0.001

We see that with both Ridge and Lasso we are getting a decent score. We can use either for the final model and because of Feature selection less important feature coefficients become zero, so we will choose Lasso regression in this case.

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

GarageType_BuiltIn
GarageType_Detchd
GarageType_NoGarage
GarageType_Others
GarageFinish_NoGarage

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

A model needs to be made robust and generalizable so that they are not impa cted by outliers in the training data. The model should also be generalisable s o that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much weightage should not given to the outliers so that the accuracy pred icted by the model is high. To ensure that this is not the case, the outlier analy sis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be rem oved from the dataset. This would help increase the accuracy of the prediction s made by the model. Confidence intervals can be used (typically 3-5 standard deviations). This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis.