Assignment-based Subjective Questions

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 = **7.5**

The optimal value of Alpha for Lasso = **0.5**

All the necessary details could be found from the below table.

Ridge Model Alpha=7.5	Ridge Model Alpha=15	Lasso Model Alpha=0.5	Lasso Model Alpha=1
GarageCars	OverallQual	YearBuilt	YearBuilt
OverallQual	GarageCars	YearRemodAdd	YearRemodAdd
BsmtFullBath	BsmtExposure_Gd	GarageArea	GarageArea
OverallCond	OverallCond	GrLivArea	GrLivArea
FullBath	FullBath	ScreenPorch	ScreenPorch

On doubling the value of alpha for both ridge and lasso, the respective most important predictor variables didn't change indeed their coefficients values had decreased.

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:

Will choose the Lasso regression. The model metrics for Lasso is better than Ridge. The R2 values of test and train data are closer. Also Lasso does the feature selection, and coefficients for many features are made 0 and hence lesser features are used compared to Ridge on larger datasets.

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:

The five most important predictor variables in the lasso model are **YearBuilt, YearRemodAdd, GarageArea, GrLivArea, ScreenPorch**

The five most important predictor variables *in the new lasso model excluding the above ones* are **OverallQual, GarageCars, BsmtExposure, OverallCond, Fireplaces**

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:

To make sure that the model is robust and generalisable we need to make sure that model is neither overfitting or underfitting. That way, model will be able to provide good results on the unseen data in the future also.

If the model is too complex, the bias will be low, but the variance will be high, i.e any slight changes in the input there will be large changes in the output. If keep the model very simple then bias will be high then the accuracy of the model will be low on the training data itself.

We will need to strike a balance between the bias and variance.

We will use techniques like Lasso and Ridge regressions to make sure that we don't overfit and thereby making the model more robust and generalisable.

A model should be generalisable so that the test accuracy is not lesser than the training score. A machine learning model should be able to perform well on unseen data with similar accuracy as on training data.

Since outliers are unavoidable in real time data, the outlier's analysis and treatment is needed to ensure the accuracy predicted by the model is good. To ensure this, unnecessary erratic data points need to be removed. This would help standardize the predictions, making it robust.

Simpler models are the best models, even though this impacts accuracy, but it will be more robust and generalisable.

It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.