Surprise Housing

Advanced Regression Assignment

upGrad

Assignment-based 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?

The optimal value for alpha are as follow :

For Ridge - 1.0 For Lasso - 0.001

On doubling the value for alpha , here are the changes in r2 score train and test and the predictor variable :

	Ridge (Optimal Alpha=1.0)	Ridge(Doubling Alpha = 2.0)	Lasso(Optimal Alpha=.001)	Lasso(Doubling Alpha=.002)
R2 Score Train	0.9188	0.9179	0.9130	0.9052
R2 Score Test	0.8919	0.8923	0.8927	0.8861
Most Important Predictor Variable	GrLivArea	GrLivArea	GrLivArea	GrLivArea

upGrad

Assignment-based 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?

R2 Score	Ridge (Optimal Alpha=1.0)	Lasso(Optimal Alpha=.001)
Train	0.9188	0.9130
Test	0.8919	0.8927

Based on the above , we have chosen lasso as our final model because of two reasons :

- It is performing slightly better in test data (unseen Data) as compared to Ridge
- It performs feature selection by driving coefficients to exactly zero thereby simplifying the model

upGrad

Assignment-based Subjective Questions

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?

In our current model the top 5 Features are : GrLivArea , TotalBsmtSF , LotArea ,propertyAge ,Neighborhood_Somerst

On Removing these Features , if we create our model again,
the new 5 features are :

- BsmtFinSF1
- BsmtUnfSF
- GarageArea
- 2ndFlrSF
- FullBath

Would also like to mention that r2 score for train and test are now 0.8868 and 0.8688 respectively

upGrad

Assignment-based Subjective Questions

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?

A model is robust and generalisable if it meets below criterias :

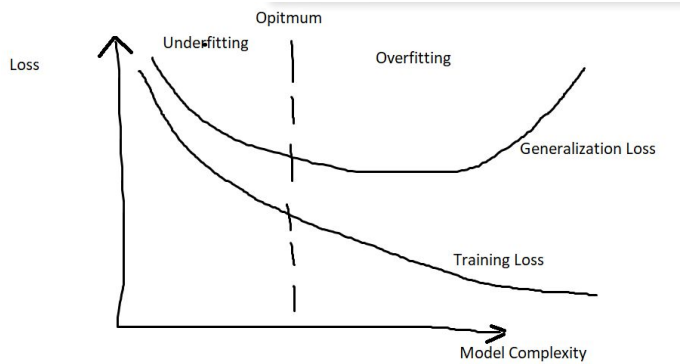
a) It should **perform well not only on training data but also on test and unseen data.** It should generalize its learning to new, unseen data and capable of identifying patterns and not just work on memorizing data.

b) Should not be overfitting which means that its predicting very high in training data and low in testing data and should not be underfitting which means that its predicting low in training as well in testing data**.**

c) The model should be kept simple and **not have more parameters** or features than necessary to solve the problem at hand.A model with high complexity has a large number of parameters and can potentially fit the training data very well, even capturing noise and random fluctuations. However, such a complex model may not generalize well to unseen data, leading to overfitting.

We can understand the above with the help of this chart that represents model complexity and data prediction errors :

Assignment-based Subjective Questions



- As the model complexity increases, the training error decreases and the test error increases
- When the model is very complex, the gap between training and test error is very high. This is overfitting
- When the model is very simple (less complex), the model will have high training error. The model is said to be underfitting.

Ensuring models generalization and robustness is essential for the success in real world application. Failing to do so will entirely defeat the purpose of creating model. Once the model is generalized and robust , it can effectively learn from the training data and make accurate prediction on the unseen data. To ensure that this achieve , we can make a checklist based on accuracy of the model performance on training and testing data. It includes :

a) Significant gap between training and testing data performance results.
b) Wide variance in its performance metrics during evaluation.
c) Testing models on new data and if it performs poorly on an entire new dataset , it lacks generalization.
d) The model's sensitivity to minor data changes, resulting in significant prediction variations, suggests overfitting and a lack of robustness.