Subjective Questions and Answers
Advanced Regression

Submitted by,

Mohana Raja Manohar

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 is **1** and lasso is **0.001**.

Ridge

There is not much difference in R2 score and mean squared error after doubling the alpha value to 2. Except few coefficients (OverallQual, YearBuilt, GrLivArea, MSSubClass_70, MSSubClass_70) all the other coefficient values are increased. The most important predictor variable remains to be 'Condition2_PosN' even after doubling the alpha value.

Lasso

There is a very slight decrease in R2 score and slight increase in mean squared error after doubling the alpha value to 0.002. The coefficient values are significantly reduced which makes two coefficient value to become zero. The most important predictor variables changes from 'Condition2_PosN' to 'GrLivArea.

Optimal value of alpha:

R2 score	***Ridge(alpha=1)***	***Lasso(alpha=0.001)***
Train Score	84.73%	83.45%
Test score	84.90%	84.18%

After Doubling the alpha value:

R2 score	***Ridge(alpha=2)***	***Lasso(alpha=0.002)***
Train Score	84.44%	82.34%
Test score	84.76%	83.64%

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?

I prefer to choose 'Lasso Regression'. The reasons are

1. The optimal value of alpha is very less compared to ridge which makes the model less complex.
2. Coefficient of a few features tend to become zero which helps in feature selection.
3. There is not much difference in r2 score and in mse when compared with 'Ridge Regression'.

4. The r2 score of test is approximately 0.75% more than the train r2 score which makes the model more generalizable.

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

The five most important predictor variables are 'GrLivArea','OverallQual','MSSubClass_160','YearBuilt' and 'MSZoning_RL'.

After the exclusion of the above features, the following features are obtained as the top 5 most predictor variables using Lasso Regression.

S.No	Predictor Variables	Coefficient
1	Heating_Grav	-0.5754
2	Neighborhood_NridgHt	0.4684
3	Exterior1st_BrkComm	-0.3572
4	MSZoning_RM	-0.3241
5	SaleCondition_Partial	0.2758

Question 4

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

Answer

A model is said to be robust and generalizable when the accuracy (r2 score) is significantly consistent on unseen data. That is the changes in the dataset should not affect the model's performance and accuracy.

For the given House Price Prediction, robustness and generalization are verified by the train and test scores of the Lasso and Ridge models.

If the model is not robust and generalizable, then the model would not perform well on unseen data. That is the model will be an underfitting model.