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 :

Ridge: 100

Ridge after RFE: 20

Lasso: 0.001

For all the models, the training score has decreased slightly and the testing score has increased slightly. The change is most noticeable for Ridge after RFE. Here, the changes were the largest, so that the gap between train and test data is the smallest.

'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'OverallCond', 'YearBuilt' are the most important predictor variables

Important predictors Lasso: Double Alpha after removing most important predictor variables:

GrLivArea	0.152414
TotalBsmtSF	0.059183
GarageCars	0.043507
YearRemodAdd	0.031122
SaleCondition_Partial	0.030626

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 :

Based on the assignment's findings, I would choose to apply Lasso regression. The R^2 score is marginally higher for Lasso, and the discrepancy between the training and testing scores is slightly smaller. This suggests that Lasso creates a more robust and generalizable model. Additionally, Lasso performs feature selection by reducing the number of features in the model, resulting in a simpler and more interpretable final model. It also achieved the lowest residual sum of squares among all the models tested, further indicating its superior performance.

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 top predictor variables are: 'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'OverallCond', 'YearBuilt'

GrLivArea 0.128692

OverallQual 0.065629

TotalBsmtSF 0.048505

OverallCond 0.043846

YearBuilt 0.043804

After removing them, the predictor variables are:

GrLivArea 0.146928

MSZoning_RL 0.083526

MSZoning_RM 0.065606

TotalBsmtSF 0.056324

GarageCars 0.043311

The predictor variables remain the same, with a slightly different order for Lasso applied to the RFE created variables.

GrLivArea 0.136432

MSZoning_RL 0.083509

TotalBsmtSF 0.073250

MSZoning_RM 0.069320

GarageCars 0.041585

It is noticeable that the top predictor variables are rather in line with the variables that I would intuitively consider important for prediction of Sales Price. Without them, the intuition I had does not carry to the new variables.

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 :

Generalization is crucial, and test accuracy should be higher than the training score; however, the difference shouldn't be excessively large. A model should generalize well during training. If the training score is very high compared to the testing score, it indicates that the model is overfitting and memorizing the data rather than learning from it. Ideally, there should not be significant discrepancies between the training and testing results. Low test scores can occur if the dataset is split too early in preprocessing, causing some steps to be skipped for the test data.

Robustness of a model isn't determined solely by high test scores; it also relies on the assumption that the training scores are higher than the testing scores. Both scores need to be sufficiently high to meet the specific business requirements and expectations for the model. Additionally, it’s essential to consider the values obtained for both training and testing to ensure the model performs well on unseen data. Retaining some outliers in the data can help improve predictions. As demonstrated in the assignment, model accuracy varies depending on data processing and feature selection methods. There may not be a perfect model, but different steps can be taken to ensure the model is suitable for its specific context and business needs.

This aligns with Occam’s razor, suggesting that the chosen model should be no more complex than necessary.