Problem Statement - Part II

Please limit your answers to less than 500 words per question.

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 (i):

- Optimal value of lambda for Ridge Regression = 10

- Optimal value of lambda for Lasso = 0.001

Answer (ii):

Changes in Ridge Regression metrics:

- R2 score of train set decreased from 0.94 to 0.93

- R2 score of test set remained same at 0.93

Changes in Lasso metrics:

- R2 score of train set decreased from 0.92 to 0.91

- R2 score of test set decreased from 0.93 to 0.91

Answer (iii):

The most important predictor variables after we double the alpha values are:-

- GrLivArea

- OverallQual_8

- OverallQual_9

- Functional_Typ

- Neighborhood_Crawfor

- Exterior1st_BrkFace

- TotalBsmtSF

- CentralAir_Y

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:

- The model we will choose to apply will depend on the use case.

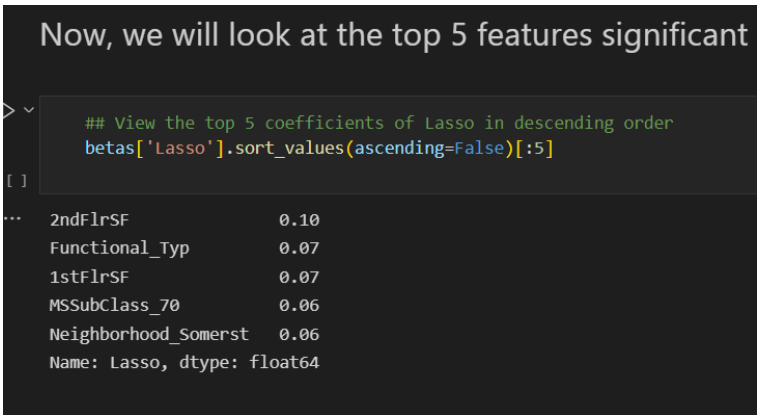
- If we have too many variables and one of our primary goals is feature selection, then we will use Lasso.

- If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we will use Ridge Regression.

Question 3

After building the model, you realized 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:



After dropping our top 5 lasso predictors, we get the following new top 5 predictors: -

- 2ndFlrSF

- Functional_Typ

- 1stFlrSF

- MSSubClass_70

- Neighborhood_Somerst

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 (i):

To ensure a model is robust and generalizable, the following steps can be taken:

-Split the data into training and validation sets, and use cross-validation techniques during training.

-Use regularization techniques such as L1/L2 regularization or dropout to prevent overfitting on the training data.

-Ensure that the model has seen enough diverse and representative examples during training to handle unseen data.

-Monitor performance on the validation set during training and use early stopping to prevent overfitting.

-Finally, evaluate the model on a held-out test set, that it has not seen during training, to obtain an estimate of its generalization performance.

Additionally, using data augmentation techniques to artificially increase the size and diversity of the training set can also help improve the robustness and generalization of the model.

Answer (ii):

Accuracy is a measure of how well a model correctly predicts the target values. Ensuring that a model is robust and generalizable has implications for its accuracy because:

-Overfitting: If a model overfits on the training data, it will perform well on the training set but poorly on unseen data, leading to a low accuracy on the validation set and test set.

-Bias-Variance tradeoff: A model that is too complex for the data can lead to overfitting, whereas a model that is too simple can lead to underfitting and a high bias. Regularization techniques help to balance this tradeoff, leading to improved accuracy.

-Representativeness of the data: If the training data is not representative of the real-world data the model will be used on, the model may have low accuracy when applied to new, unseen data. Using a diverse and representative training set helps to improve the accuracy of the model.

-Model capacity: If the model is too complex or has too many parameters, it may be overfitting the data. On the other hand, if it is too simple, it may not capture the underlying patterns in the data, leading to underfitting. Choosing a model with the appropriate capacity for the data helps to improve accuracy.

In summary, ensuring that a model is robust and generalizable is important for improving its accuracy on unseen data.