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:

Optimal value of alpha for

Ridge = 10

Lasso =0.001

R-Squared (Test) of both models are almost same with best alpha.

After doubling alpha values, highlighted the important predictor variables in below table.

	Best Alpha	**Double value of Best Alpha**
Ridge Regression	**Alpha =10** R-Squared (Train) = 0.95 R-Squared (Test) = 0.90 RSS (Train) = 6.32 RSS (Test) = 5.57 MSE (Train) = 0.01 MSE (Test) = 0.01 RMSE (Train) = 0.08 RMSE (Test) = 0.1	**Alpha =20** R-Squared (Train) = 0.94 R-Squared (Test) = 0.90 RSS (Train) = 7.16 RSS (Test) = 5.55 MSE (Train) = 0.01 MSE (Test) = 0.01 RMSE (Train) = 0.08 RMSE (Test) = 0.11
	GrLivArea 1.093435 OverallQual_9 1.089846 Neighborhood_Crawfor 1.084737 OverallQual_8 1.082082 OverallCond_9 1.078183 TotalBsmtSF 1.062831 Functional_Typ 1.060670 Exterior1st_BrkFace 1.054403 Neighborhood_Somerst 1.050898 MSSubClass_70 1.046778	GrLivArea 1.085171 OverallQual_8 1.073845 OverallQual_9 1.068418 Neighborhood_Crawfor 1.067120 TotalBsmtSF 1.058635 Functional_Typ 1.052955 OverallCond_9 1.051047 Neighborhood_Somerst 1.044468 Exterior1st_BrkFace 1.042851 MSSubClass_70 1.042299
Lasso Regression	**Alpha=0.001** R-Squared (Train) = 0.93 R-Squared (Test) = 0.89 RSS (Train) = 8.65 RSS (Test) = 5.77 MSE (Train) = 0.01	**Alpha=0.002** R-Squared (Train) = 0.92 R-Squared (Test) = 0.88 RSS (Train) = 10.63 RSS (Test) = 6.64 MSE (Train) = 0.01

	MSE (Test) = 0.01 RMSE (Train) = 0.09 RMSE (Test) = 0.11	MSE (Test) = 0.02 RMSE (Train) = 0.10 RMSE (Test) = 0.12
	OverallQual_9 1.160192 OverallQual_8 1.127971 GrLivArea 1.110877 Neighborhood_Crawfor 1.100521 Functional_Typ 1.064080 OverallQual_7 1.061799 Neighborhood_Somerst 1.059394 TotalBsmtSF 1.055890 Exterior1st_BrkFace 1.047260 OverallCond_9 1.046992	GrLivArea 1.113664 OverallQual_8 1.109393 OverallQual_9 1.106611 Neighborhood_Crawfor 1.073535 Functional_Typ 1.057430 TotalBsmtSF 1.056062 OverallQual_7 1.053188 YearRemodAdd 1.041376 Condition1_Norm 1.034122 GarageArea 1.030037

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:

For below reasons, will go with Lasso regression model:

- R-Squared (Test) of both models are almost same with best alpha.
- Lasso regression have eliminated unimportant features.

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:

Most important predictors after removing top5 predictors variables of Lasso model are:

```
Condition2_PosA    0.280342
OverallCond_9      0.179527
SaleType_ConLD     0.156266
OverallCond_8      0.095422
2ndFlrSF           0.094494
```

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:

A model is considered to be robust if its output and forecasts are consistently accurate even if one or more of the input variables or assumptions are changed due to unforeseen circumstances.

Understand the Problem and the Data: It is crucial to have a thorough understanding of the problem to solve and the data given. Exploratory data analysis, understanding relationships between variables, identify potential outliers, and understanding of overall structure of your data. This initial exploration helps to make informed decisions about which machine learning techniques might be most appropriate for your specific problem.

Data Preprocessing: Data preprocessing is a critical step to build robust model. This may include tasks such as handling missing values, encoding categorical variables, normalizing numerical variables, or dealing with class imbalance in your target variable. The goal of preprocessing is to make data compatible with the machine learning algorithm we are using and to improve the algorithm's ability to uncover meaningful patterns.

Feature Engineering: Feature engineering involves creating new features from existing ones through domain knowledge. It can have a significant impact on the performance of a machine learning model.

Model Selection: The choice of model should depend on the problem and the nature of the data, and the requirement of the task. It's often a good idea to start with simpler models and move to more complex ones if necessary. Simpler models are easier to interpret and less likely to overfit the data.

Cross-Validation: Cross-validation is a powerful technique that can help prevent overfitting, a common problem in machine learning where a model learns the training data too well and performs poorly on unseen data. By dividing your data into training and validation sets multiple times, you can ensure that your model generalizes well to unseen data.

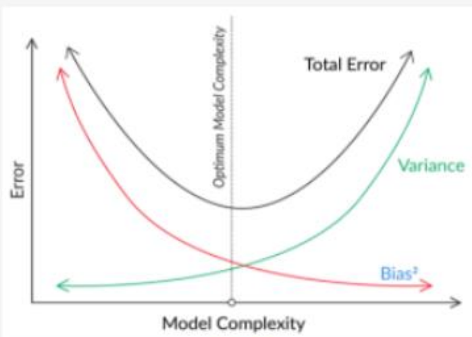
Hyperparameter Tuning: Hyperparameters are the parameters of the learning algorithm itself, and they can significantly affect model performance. Techniques such as grid search or random search can be used to find optimal hyperparameters for your model. However, remember that over-tuning the hyperparameters on your validation set may lead to overfitting.

Regularization: Regularization is a technique used to prevent overfitting by adding a penalty term to the loss function.

Model Evaluation: Always evaluate model using appropriate metrics. The choice of metrics should align with the business objective.

Implications:

An important implication is that too generalized or too robustness will have impact on accuracy of the model.



Image