Advanced Regression

Regularization

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?

Answer:

Optimal value of alpha for ridge and lasso regression:

- Ridge Regression Alpha: **0.8**
- Lasso Regression Alpha: **0.01**

Please find below the metrics after choosing double the value of alpha for both ridge and lasso.

Metrics	Ridge		Lasso	
	Alpha 0.8	Alpha 1.6	Alpha 0.01	Alpha 0.02
r2_score train	0.927241385	0.926631701	0.927529163	0.927488616
r2_score test	0.926467116	0.926782447	0.925882754	0.926129899
rss train	1619914.672	1633488.802	1613507.519	1614410.269
rss test	865611.376	861899.3784	872490.3401	869581.0167
mse train	1814.01419	1829.214783	1806.839327	1807.850245
mse test	1967.298582	1958.862224	1982.932591	1976.320493

Coefficient values are slightly changing in the train data and test data after doubling the values of alpha for both ridge and lasso regression.

**** Coding solution is given in the notebook. ****

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 Lasso regression model fits the data slightly better than the Ridge regression model, as indicated by r2_scores.

- The optimal alpha value for Lasso regression is close to 0, which indicates that the model is performing a standard linear regression with minimal penalty on the coefficients.

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:

In Lasso regression model, the initial top five most important predictor variables are **OverallQual, GrLivArea, OverallCond**, **YearBuilt , 2ndFlrSF** with r2_score of 92% in train data and 91 % in test data. However, after dropping these variables, we attained the top five important variables are **FullBath, TotRmsAbvGrd, Street, 1stFlrSF, HalfBath** with r2_score of 89 % in train data and 89% in test data.

**** Coding solution is given in the notebook. ****

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:

To create a robust model, it's important to use cross-validation with multiple k-folds on the training data. This helps identify the best model and optimal number of folds. After finding the best model, evaluate it on unseen data. To ensure the model generalizes well, handle outliers properly, use various encoding methods, and apply techniques like regularization, random forest, and decision tree. These steps help prevent overfitting and multicollinearity.

A weak model might memorize the training data instead of learning and generalizing patterns in the dataset.