Assignment Part -II - 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?

Answer1

The optimal values of alpha obtained for Ridge and Lasso regression are 2.0 and 0.0001, respectively. If we were to double the alpha value for both Ridge and Lasso, there would be some changes in the model. In the case of Ridge, there is a small increase in mean squared error, but the r2 values for both train and test remain almost the same. However, in the case of Lasso, there is a slight increase in mean squared error, a slight decrease in r2 value for train, and a huge fall in r2 value for test, making the model and prediction worse. Additionally, this further penalizes the model and causes a larger number of coefficients of a variable to shrink towards zero.

After the changes are implemented, the most important predictor variables for Ridge and Lasso are as follows:

For Ridge:

- Total_sqr_footage
- OverallQual
- GrLivArea
- Neighborhood_StoneBr
- OverallCond
- TotalBsmtSF
- LotArea
- YearBuilt
- Neighborhood_Crawfor
- Fireplaces

For Lasso:

- Total_sqr_footage

- OverallQual
- YearBuilt
- GrLivArea
- Neighborhood_StoneBr
- OverallCond
- LotArea
- Neighborhood_Crawfor
- Neighborhood_NridgHt
- GarageCars

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 optimal alpha values obtained for Ridge and Lasso regression are 2.0 and 0.0001, respectively. **The r2 values for Ridge and Lasso regression are Train = 0.930, Test = 0.896 and Train = 0.927, Test = 0.902, respectively.**

The difference in **r2 values for Ridge and Lasso is 0.046 and 0.025, respectively.** The Mean Squared Error for Ridge and Lasso is 0.00297 and 0.00280, respectively, and we can observe that the **Mean Squared Error for Lasso is slightly lower than that of Ridge.**

Moreover, the difference in r2 between train and test is less in Lasso compared to Ridge. Lasso is beneficial for feature reduction as it shrinks the coefficient value of one of the Lasso's features towards zero, and it increases model interpretation by taking the magnitude of the coefficients. Therefore, **Lasso has an advantage over Ridge.**

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

After dropping the top 5 most important predictor variables in the Lasso model, we created a new model, and the following five predictor variables became the most important:

- TotalBsmtSF
- TotRmsAbvGrd
- OverallCond
- Total_Bathrooms
- LotArea

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

To ensure that a model is robust and generalizable, we need to balance between the model's complexity and its performance on both the training and testing data. Here are some ways to achieve that:

- **Regularization**: By using regularization techniques such as Ridge, Lasso, or Elastic Net regression, we can manage the model's complexity and reduce the risk of overfitting. These techniques work by adding a penalty term to the loss function that shrinks the model's coefficients towards zero, making it simpler.
- **Cross-validation**: By using cross-validation, we can evaluate the model's performance on different subsets of the data and check its generalizability. This technique helps in estimating the model's performance on unseen data.
- **Feature Selection**: By selecting only relevant features that have a significant impact on the model's performance, we can simplify the model and make it more robust. This technique reduces the model's complexity and avoids overfitting.

The implications of having a robust and generalizable model are that it can make accurate predictions on new data, which is not seen during model training.

A model that is overfitted to the training data may have a high accuracy on the training data but may perform poorly on new data, whereas a model that is too simple may not have enough information to make accurate predictions. Therefore, it is essential to balance the model's complexity and accuracy to achieve a robust and generalizable model.