Assignment-based 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

In the case of ridge regression, the train error shows an increasing trend as the value of alpha grows, and the error term decreases as the value of alpha increases from 0. This is evident when we plot the curve between negative mean absolute error and alpha. We chose to use a value of alpha equal to two for our ridge regression since the test error is at its lowest when alpha is at 2.

For lasso regression, I have chosen to maintain a very low value of 0.01; as alpha increases, the model attempts to punish more and reduce the majority of the coefficient values to zero. It was initially reported as having an alpha and negative mean absolute error of 0.4.

For our ridge regression, doubling the alpha value will result in taking the value of alpha equal to 10. At this point, the model will attempt to make itself more generalized, which will make it simpler and eliminate the need to fit every piece of data in the data set. It is evident from the graph that higher error is encountered in both training and testing with alpha values of 10.

Similar to this, when we increase the lasso's alpha value, we attempt to penalize our model more, resulting in more variable coefficients being reduced to zero. Consequently, our r2 square value similarly declines as we increase the alpha value.

The most important variable after the changes has been implemented for ridge regression are as follows:

- MSZoning_FV
- MSZoning_RL
- Neighborhood_Crawfor
- MSZoning_RH
- MSZoning_RM
- SaleCondition_Partial
- Neighborhood_StoneBr
- GrLivArea
- SaleCondition_Normal
- Exterior1st_BrkFace

The most important variable after the changes has been implemented for lasso regression are as follows:

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- BsmtFinSF1

- GarageArea
- Fireplaces
- LotArea
- LotArea
- LotFrontage

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

Regularizing coefficients, increasing prediction accuracy together with a reduction in variance, and ensuring that the model is comprehensible are crucial.

As the penalty is the square of the magnitude of the coefficients, which is determined via cross validation, ridge regression employs a tuning parameter called lambda. Use of the penalty will result in a minor residual sum or squares. Higher-valued coefficients are punished because the penalty is equal to lambda times the sum of squares of the coefficients. The model's variance decreases and its bias stays constant when we increase the value of lambda. Ridge regression, in contrast to Lasso regression, incorporates all variables into the final model.

By using cross-validation to identify the absolute value of the magnitude of the coefficients as the penalty, Lasso regression employs a tuning parameter known as lambda. Lasso makes the variables exactly equal to zero by shrinking the coefficient towards zero as the lambda value increases. Additionally, Lasso selects variables. The model does simple linear regression when the lambda value is small. As the lambda value rises, shrinkage occurs and variables with a value of 0 are ignored.

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

Those 5 most important predictor variables that will be excluded are:

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- GarageArea

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

The model should be as straightforward as feasible; this will increase its robustness and generalizability at the expense of decreased accuracy. The trade-off between bias and variance can also be used to understand it. More bias but lower variance and more generalizability are seen in simpler models. It implies that a reliable and generalizable model will function similarly on training and test data, meaning that accuracy will not significantly vary between the two sets of data.

Bias is an error in a model that occurs when it is unable to draw conclusions from the data. High bias indicates that the model cannot pick up on subtleties in the data. On training and testing data, the model performs poorly.

Variance is the result of a model's attempt to learn too much from the data. High variance indicates that the model performs remarkably well on training data since it has been trained on this type of data, but it performs terribly on testing data because the model has not seen this type of data.

To prevent both overfitting and underfitting of the data, bias and variance must be balanced.