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 1

Ridge regression: Ridge Regression is a technique for analysing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. For ridge regression, the optimal value of alpha is 20.

Lasso Regression: Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). In this case of Lasso regression, the optimal value for alpha is 1.

If we choose to double the value of alpha for both ridge and lasso regression, model complexity will have a greater contribution to the cost. Because the minimum cost hypothesis is selected, this means that higher λ will bias the selection toward models with lower complexity.

After the second model is built, we compare the r square value of new model with the old one. The model which is having high r square of test and train dataset, we will select the features/variables from that model. And the variable is selected based on the high coefficient value.

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 2

We found the optimal values of lambda for lasso and ridge to be 0.001 and 0.9 respectively.

TOP 5 FEAUTURES FROM ABOVE MODELS ARE:

- YearBuilt_Old

- Neighborhood_Gilbert
- HalfBath
- MiscVal
- Neighborhood_IDOTRR

1 For the same values of alpha, the coefficients of lasso regression are much smaller as compared to that of ridge regression.

2. For the same alpha, lasso has higher RSS (poorer fit) as compared to ridge regression

3. Many of the coefficients are zero even for very small values of alpha.

Typical Use Cases

Ridge: It is majorly used to prevent overfitting. Since it includes all the features, it is not very useful in case of exorbitantly high #features, say in millions, as it will pose computational challenges.

Lasso: Since it provides sparse solutions, it is generally the model of choice (or some variant of this concept) for modelling cases where the #features are in millions or more. In such a case, getting a sparse solution is of great computational advantage as the features with zero coefficients can simply be ignored. It's not hard to see why the stepwise selection techniques become practically very cumbersome to implement in high dimensional cases. Thus, lasso provides a significant advantage. We will select lasso regression with this dataset.
Lasso regression would be a better option as it would help in feature elimination and the model will be more robust. Because, in the Ridge, the coefficients of the linear transformation are normally distributed and, in the Lasso, they are distributed in Laplace form. In the Lasso, this makes it easier for the coefficients to be zero and therefore easier to eliminate some of your input variable as not contributing to the output.
Ridge regression can't zero out coefficients; thus, you either end up including all the coefficients in the model, or none of them. In contrast, the LASSO does both parameter shrinkage and variable selection automatically.
Lasso regression can produce many solutions to the same problem.
Ridge regression can only produce one solution to one problem.

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 3

Statistical measures can show the relative importance of the different predictor variables.

However, these measures can't determine whether the variables are important in a practical sense. To determine practical importance, one needs to use his own subject area knowledge. How one collects and measures his sample can bias the apparent importance of the variables in his sample compared to their true importance in the population. If one randomly samples his observations, the variability of the predictor values in his sample likely reflects the variability in the population. In this case, the standardized coefficients and the change in R-squared values are likely to reflect their population values. However, if one selects a restricted range of predictor values for his sample, both statistics tend to underestimate the importance of that predictor. Conversely, if the sample variability for a predictor is greater than the variability in the population, the statistics tend to overestimate the importance of that predictor. Also, consider the accuracy and precision of the measurements for the predictors because this can affect their apparent importance. For example, lower-quality measurements can cause a variable to appear less predictive than it truly is. How one defines the "most important" often depends on one's goals and subject area. While statistics can help one identify the most important variables in a regression model, applying subject area expertise to all aspects of statistical analysis is crucial. Real world issues are likely to influence which variable someone identifies as the most important in a regression model.

For example, if my goal is to change predictor values in order to change the response, I will use my expertise to determine which variables are the most feasible to change. There may be variables that are harder, or more expensive, to change. Some variables may be impossible to change. Sometimes a large change in one variable may be more practical than a small change in another variable. "Most important" is a subjective, context sensitive characteristic. One can use statistics to help identify candidates for the most important variable in a regression model, but he will likely need to use his subject area expertise as well.

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 4

- A model is robust when any variation in the data does not affect its performance much.
- A generalizable model can adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.
- To make sure a model is robust and generalizable, we must take care so that, it doesn't overfit. This is because an overfitting model has very high variance and the smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data but fail to pick up the patterns in unseen test data.
- In other words, the model should not be too complex in order to be robust and generalizable.

- If we look at it from the perspective of Accuracy, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. The addition of bias means that accuracy will decrease.
- In general, we have to make some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge Regression and Lasso.