Problem Statement - Part II

Advanced Regression | House Price Prediction

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:

The optimal value of alpha is the one that strikes the right balance between fitting the data well and preventing overfitting. Essentially, it helps in finding a model that generalizes well to new, unseen data.

If we double the value of alpha for both ridge and lasso regression, it significantly increases the penalty on the coefficients. For ridge regression, this means the model will tend to shrink the coefficients more aggressively.

After implementing changes, the top 5 important variables for Ridge regression are:

- MSZoning_FV
- MSZoning_RL
- Neighborhood_Crawfor
- MSZoning_RH
- MSZoning_RM

For Lasso regression, the top 5 important variables are:

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- BsmtFinSF1

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:

Lasso regression seems to be the preferred over Ridge regression for the following reasons:

- Better Prediction Performance: Lasso regression slightly outperforms Ridge regression in predicting outcomes for new data, indicated by a higher R2 test score.
- Simpler Model: Lasso automatically selects the most important features by reducing the coefficients of irrelevant or redundant features to zero. This makes the model easier to understand and interpret.
- Less Risk of Overfitting: Lasso has a lower optimal lambda value, which means it's less likely to overfit the data and more robust to noise.

Both Ridge and Lasso regression help improve prediction accuracy and reduce overfitting, but Lasso's ability to perform feature selection by setting coefficients to zero makes it particularly advantageous when dealing with large datasets with many 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:

The five most crucial predictor variables that will be left out are:

- Above grade (ground) living area square feet (GrLivArea)
- Rates the overall material and finish of the house (OverallQual)
- Rates the overall condition of the house (OverallCond)
- Total square feet of basement area (TotalBsmtSF)
- Size of garage in square feet (GarageArea)

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:

Creating a model is an ongoing process where we adjust various parameters and settings to make our model strong. Some things to think about include:

- Handling outliers carefully.

- Using a reliable way to measure errors.
- Trying different transformations.
- Experimenting with various models to find the best one.
- Simplifying a model involves a trade-off between bias and variance:

A complex model can be very sensitive to changes in the data, making it unstable.

On the other hand, a simpler model might not change much even if the dataset grows or shrinks.

Bias tells us how accurate a model is likely to be on new data. A complex model can be accurate if it has enough training data. But if a model is too basic, like always giving the same answer no matter the input, it has high bias because it's consistently wrong.

Variance measures how much a model changes when the training data changes.

So, finding a balance between bias and variance helps maintain accuracy and minimizes errors in the model.

***