Advanced Regression

Q1: 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?

Ans:

In the models built for "House Price Predictions", the optimal values of alpha for lasso and ridge model are as below.

1. Ridge: Optimal value of alpha for ridge model is 10.
2. Lasso: Optimal value of alpha for lasso model is 0.0001.

In general, as alpha value is increased, values of coefficients for the features tends to 0 value. In this case, when alpha value of the model is doubled coefficients of the features are reduced.

Considering the Optimal value for Ridge and Lasso below are the top 5 features (considering absolute values of the coefficients)

Ridge:

1. PoolQC_Gd [2.538906]
2. Condition2_PosN [1.391778]
3. OverallQual_2 [0.338979]
4. Functional_Sev [0.211065]
5. BsmtFullBath_3 [0.198986]
6. OverallQual_9 [0.197138]

Lasso:

1. PoolQC_Gd [2.538906]
2. Condition2_PosN [1.391778]
3. OverallQual_2 [0.338979]
4. Functional_Sev [0.211065]
5. BsmtFullBath_3 [0.198986]
6. OverallQual_9 [0.197138]

After doubling the values for alpha, below are the top 6 features for Ridge and Lasso models along with their coefficient values.

Ridge:

1. OverallQual_9 [1.002570e-01]
2. Neighborhood_Crawfor [8.228277e-02]
3. OverallCond_3 [7.093949e-02]
4. OverallQual_8 [6.936498e-02]
5. Neighborhood_IDOTRR [6.916424e-02]
6. GrLivArea [6.781082e-02]

Lasso:

1. PoolQC_Gd [2.300422]
2. Condition2_PosN [1.245013]
3. OverallQual_2 [0.272672]
4. OverallQual_9 [0.197632]
5. Neighborhood_MeadowV [0.176939]
6. OverallCond_3 [0.163772]

By observing above features and coefficients we can say that the top features are changed. Also values for some common features is reduced, which says that as alpha increases value of coefficients moves towards zero.

Q2: 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?

Ans:

When opting between Ridge and Lasso Regression models below points must be taken into consideration:

1. If you have lots of variables and want to reduce the complexity then use Lasso model.
2. If the variables is important in business point of view and you need it in your model then use Ridge model.
3. When dataset is large, it's advisable to use Lasso model.

After considering above points, we will choose **"Lasso"** model

As our model has

1. Around 300 features (some of which are obsolete and needs to be removed)
2. 1400 + records (Not large dataset but results will be better)
3. R2 value of Lasso model is 95% while R2 value of Ridge model is 92%.

Q3: 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?

Ans:

The top 5 features those positively affect the house prices are as below:

(Note: We have taken those features only those have positive affect on prediction)

1. OverallQual_9 [0.197632]
2. OverallCond_9 [0.139074]
3. OverallQual_10 [0.126821]
4. Neighborhood_Crawfor [0.123890]
5. SaleType_ConLD [0.110448]

As per the situation, we are not getting these 5 features, hence after rebuilding the model, the top 5 features are:

1. Neighborhood_Crawfor [0 0.132274]
2. Condition2_PosA [0.129827]
3. Neighborhood_StoneBr [0.099100]
4. Street_Pave [0.097288]
5. KitchenAbvGr_1 [0.092950]

Q4: 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?

Ans:

Robust Model:

A system which performs effectively while its variables or assumptions are altered. A robust model will produce positive results in any conditions.

Also a robust model is unaffected by outliers present in the dataset.

Whenever a model is being built, first it is trained on a training dataset. While training the model, techniques such as mean () or standard deviations are used.

To make sure that a model is robust, model must be built using Median or Median Absolute Deviation which are unaffected by outliers in the data. And will not be affected while predicting on unseen or altering dataset.

Generalized Model:

It is models ability to fit on unseen instance or unseen data.

To make sure model is generalizable:

1. If the distribution change, the model needs to change or in other words, it needs to be retrained in order to fit the new distribution of the population.
2. Need to make sure that training data is representative of the entire population.
3. In practice, the distribution of population changes over time. Thus, any ML model needs frequent re-training rounds in order to keep up with the change of the population.

Robust and Generalizable model makes sure that accuracy of the model does not deviate much from the original model.