Problem Statement - Part II

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 in the case of ridge regression is 20. Whereas, the optimal value of alpha for the lasso regression is 0.001.

If we double the value of alpha for ridge and lasso, the regularisation will impose a higher penalty and will impact the coefficients of the features adversely. This might cause underfitting of the model. In the case of lasso regression, in addition to underfitting the model, the model would choose a lesser number of significant features.

Impact on Ridge model on doubling the alpha value:

The optimal value of alpha is 20. Doubling it will make the alpha value as 40.

- The R2 value on the train data varies marginally from 0.8843 to 0.8698
- The R2 value on the test data varies marginally from 0.8656 to 0.8608

The important predictor variables with alpha values 20 and 40 are as listed below. The variables are ordered in a decreasing order of significance in the table.

Significant predictor variables in the model (Ridge Regression)	
With Alpha value 20	**With Alpha Value 40**
Neighborhood	OverallQual
OverallQual	Neighborhood
Exterior1st	BsmtQual
BsmtExposure	BsmtExposure
ExterCond	ExterCond

Impact on Lasso model on doubling the alpha value:
The optimal value of alpha is 0.001. Doubling it will make the alpha value as 0.002.

- The R2 value on the train data varies marginally from 0.9031 to 0.8898
- The R2 value on the test data varies marginally from 0.8797 to 0.8720

The important predictor variables with alpha values 0.001 and 0.002 are as listed below. The variables are ordered in a decreasing order of significance in the table.

Significant predictor variables in the model (Lasso Regression)	
With Alpha value 0.001	**With Alpha Value 0.002**
Condition2	Exterior1st
Exterior1st	Neighborhood
Neighborhood	OverallQual
OverallCond	Condition2
OverallQual	OverallCond

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 value of lambda is chosen based on a GridSearch experimenting with different lambda values on the training dataset. The scoring is done based on “negative mean absolute error”.

I will choose the best lambda value found by the GridSearch algorithm because it compares how the model performs on different lambda values and provides the best lambda value for the model.

The Ridge regression had an optimal lambda value of 20 and the lasso regression had an optimal lambda value of 0.001. The R2 score was found for each of the models and we found the score to be slightly higher in the lasso model.

Apart from the fact that the lasso model had a higher R2 score for the test model, the lasso regression model is a simpler model. It has fewer number of features because the Lasso regression makes some of the feature's coefficients zero nullifying the effect of insignificant features. In the model built on the housing price prediction dataset, the number of features reduced from 297 to just 81.

Since the lasso regression model is a simpler model as per Occam's razor principle, we could use this model with the optimal lambda value of 0.001.

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 lasso model (with alpha=0.001) that was built had the following predictors as the five most important predictors:

1. Condition2
2. Exterior1st
3. Neighborhood
4. OverallCond
5. OverallQual

When removing these variables in the model and creating another model with lasso regression, we found the following characteristics (with train-test split's random seed set as 1):

- The optimal value was found as 0.01
- Number of features chosen by lasso regression was only 24.
- The test R2 score was found as 0.839
- The five most important predictor variables are as follows:
    - BsmtQual
    - BsmtFullBath
    - ExterCond
    - Functional
    - 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:

Some of the ways that we can make sure the model is robust and generalisable are as follows:

- Choose the model which is the most simple one. Follow the Occum's razor principle while choosing the model.
- A simple model can be defined as the one which has the least number of predictor features without compromising much on the model's accuracy.
- The model must not be underfit or overfit. When a model is overfit, it performs well on the training dataset but not on the test dataset. When a model is underfit, the model has not learnt anything and does not perform well on both training and testing dataset.
- To avoid overfitting, one should use regularisation. This comes in handy when the number of predictor variables are very high in number.
- There are different regularisation techniques. Some of the popular regularisation techniques are Lasso and Ridge regression techniques.
- Always find a balance between variance and bias that could give the best performance from the model.

When a model is robust and generalisable, the model can perform well on any unseen data. This is because the model would have been neither underfit nor overfit. Also, the model would be having a balance between the variance and bias.

When using regularisation techniques such as lasso regression, the model would even reduce the number of predictor variables without compromising much on the model's accuracy. This makes the model more efficient and uses less memory to run the model.