Assignment – 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 for ridge regression is alpha = 4 and for lasso regression is alpha = 100.

When we to double the value of alpha for regression, it tries to make the model more generalized, i.e. making the model simpler with low variance and high bias, due to which the model leads to underfit. Also, model coefficients are pushed more towards zero.

The most important variable after changes are implemented are following:

- For ridge = 1stFlrSF
- For lasso = GrLivArea

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:

	r2_train	r2_test	Alpha
Ridge	0.960	0.875	4
Lasso	0.948	0.894	100

I would choose **Lasso** regression. Although the R2 score values are really close for both methods. But for Lasso, the test score is more, and closer to the train score, meaning, it will give better results for unseen data. Also, lasso provides feature selection, which makes it easier to interpret models.

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:

Below are the five most important predictor variables after removing the previous important ones:

- 1stFlrSF
- 2ndFlrSF
- BsmtFinSF1
- BsmtUnfSF
- Neighborhood_StoneBr

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:

It is important to find a balance between bias and variance to avoid overfitting and underfitting of data.

There are several ways to make sure that a model is robust and generalizable.

- Use cross-validation if the dataset is small. It helps as it has been trained and evaluated on different subsets of the data
- Use regularisation if there's an overfitting to make sure it reduces the complexity of a model
- Splitting the data into train-test dataset to make sure the same trained data is not used for testing

A model which is robust and generalisable will have a better accuracy for new data, whereas, if a model is not generalised and is overfitted on training data, then the accuracy will be really poor for the upcoming data because it has never learned to generalised new situations.