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

Optimal value of alpha 0.0001 for lasso regression

Optima value of alpha 4.0 for ridge regression

After doubling the value of alpha for lasso and ridge is 0.0002 and 8.0

For ridge: coefficient values are increasing as alpha will increase. There is little bit drop in r2 score also of train and test.

R2 score train: 0.9380 to 0.9326
R2 score test: 0.9064 to 0.9044

For lasso: as the value of alpha increases more features were removed from the model. Th ere is little drop in r2 score also of train and
Test.
R2 score train: 0.9388 to 0.9315
R2 score test: 0.9085 to 0.9047

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 model we will choose to apply will depend on the use case.**
- **If we have too many variables and one of our primary goals is feature selection, then we will use lasso.**
- **If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we will use ridge regression.**

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

Top five features: Totalbsmtsf","2ndflrsf","condition2_posn","lotarea","grlivarea"

	Ridge	Lasso	Abs_value_coeff
2ndflrsf	**0.105507**	**0.181091**	**0.181091**
Condition2_posN	**-0.053060**	**-0.148530**	**0.148530**
Totalbsmtsf	**0.105940**	**0.143720**	**0.143720**
Lotarea	**0.038865**	**0.091005**	**0.091005**
Grlivarea	**0.064963**	**0.087756**	**0.087756**

After dropping the top 5 features. These are the next top features.

	Lasso	abs_value_coeff
Bsmtfinsf1	**0.126570**	**0.126570**
Overallqual_9	**0.112662**	**0.112662**
Bsmtfinsf2	**0.108273**	**0.108273**
Bsmtunfsf	**0.100658**	**0.100658**
Totrmsabvgrd	**0.073038**	**0.073038**

Question 4

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?

Answer

The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalizable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

- **A model is robust when any variation in the data does not affect its performance much.**

- **A generalizable model is able to 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 have to take care it doesn't over fit. This is because an over fitting model has very high variance and a 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.**
- **In general, we have to find strike some balance between model accuracy and complexity. This can be achieved by regularization techniques like ridge regression and lasso.**