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?

Answer1:

Optimal value of alpha for ridge: **10**

Optimal value of alpha for lasso: **0.01**

Note:- The Rsquared values for both the Ridge and Lasso don’t seem to be too great, however, a balance between train and test has been achieved and upon using the above values of alpha, we are not seeing significant drop in the Rsquared value on test set. But notably, the model is performing better on test set than the train set in both cases for the selected values of alpha.

After make the double alpha for ridge and lasso i.e. **20 and 0.02**

For Ridge:

- Coefficient values are increasing slightly.
- Rsquared on train set is decreasing from 0.501 to 0.493
- Rsquared on test set is decreasing from 0.516 to 0.510

For Lasso:

- More feature coefficients are reduced to 0. But this is not significant as alpha is not increasing significantly.
- Rsquared on train set is decreasing from 0.450 to 0.393
- Rsquared on test set is decreasing from 0.465 to 0.385

Top 5 Features:-

For Ridge-

For alpha=10	***For alpha=20***
MSSubClass	MSSubClass
BsmtFinSF2	BsmtFinSF2
CentralAir	Street
Street	OverallCond
BsmtUnfSF	BsmtUnfSF

For Lasso:-

For alpha=0.01	**For alpha=0.02**
TotalBsmtSF	TotalBsmtSF
MSZoning_RM	MSZoning_RM
GarageType_No_Garage	
SaleType_New	

Lasso converted all other feature coefficients to 0.

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 2:

We will choose **Ridge** as it is performing significantly better on both the train and test set. Moreover, I think Lasso works blindly as it doesn't have the domain knowledge.

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 3:

After dropping the only four coefficients with non zero value, the Lasso regression became absurd and started giving unacceptable Rsquared values of below 10 percent for both train and test set for alpha=0.01. Nonetheless, the new main predictors is “RoofMatl_WdShngl”, as all other feature coefficients have been converted to 0.

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 4:

To make model robust and generalisable 3 features are required:

1. **Stable performance.** The model is performing similarly for both train and test data.
2. **P-value** of all the features is < 0.05
3. **VIF** of all the features are < 5

Thus we are sure that model is robust and generalisable.