

## Problem Statement - Part II

### Subjective Questions

**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?

Lasso Regression when alpha = 0.001		
1	R2 score of train dataset	0.9163
	R2 score of test dataset	0.8851
	Mean Square Error	0.0162
	Lasso Regression when alpha = 0.002	
	R2 score of train dataset	0.9062
	R2 score of test dataset	0.8849
	Mean Square Error	0.0162
Ridge Regression when alpha 7.0		
2	R2 score of train dataset	0.9320
	R2 score of test dataset	0.8838
	RMSE value of ridge model	0.1638
	Ridge Regression when alpha 14	
	R2 score of train dataset	0.9275
	R2 score of test dataset	0.8849
	RMSE value of ridge model	0.1622

#### Changes in Ridge Regression metrics:

- R2 score of train dataset decreased from 0.9163 to 0.9062
- R2 score of test dataset decreased from 0.8851 to 0.8849

#### Changes in Lasso metrics:

- R2 score of train dataset decreased from 0.9320 to 0.9275
- R2 score of test dataset decreased from 0.8838 to 0.8849

**Note:** Detailed Steps and solution are mentioned in the jupyter notebook

**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?

**Solution:**

- In order to determine which model to apply, we will need to consider the use case of the system.
- The Lasso technique is a good one to use if we have a lot of variables and one of the primary goals is feature selection.
- The Ridge Regression is a good choice if we do not want to get too large coefficients, and reducing the magnitude of the coefficients is one of our priorities.

**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?

**Solution:**

**The five most important predictor variables are**

- SaleType\_New
- 2ndFlrSF
- OverallQual
- 1stFlrSF
- SaleCondition\_Normal

**Note: Detailed Steps are mentioned in the jupyter notebook**

**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?

**Solution:**

- The performance of a model is robust when any variation in the data does not significantly affect the performance of the model.
- It is important to know that generalizable models are able to adapt appropriately to new information drawn from the same distribution as when the model was created, that had previously not been observed before.
- It is very important to ensure that a model is robust and generalizable by ensuring that it is not overfit. This is due to the fact that an overfitted model has a very high variance which means even a small change in the data can have a profound effect on the model prediction. This type of model is able to identify all the patterns in the training data, but is unable to pick up any patterns in unseen test data.

- Therefore, it is necessary for the model to be robust and generalizable so that it should not be too complex.