# ASSIGNMENT - ADVANCED REGRESSION

## PROBLEM STATEMENT - PART II - SUBJECTIVE QUESTIONS & ANSWERS

### UPGRAD IIITB EPGP in ML & AI - MLC39

**By:**

- PADMAVATHI D

## 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 for ridge & lasso regression are:

Ridge – 4

Lasso – 0.0003

When we double the value of alpha for our ridge regression, the model will apply more penalty on the curve & try to make the model more generalised that is making model more simpler & thinking to fit every data of the dataset & we can observe more error for both test & train.

Similarly when we double the value for our lasso regression, we try to penalize more our model & more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases.

Hence the most important features after the changes are

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- GarageCars
- MSZoning_RL

Predictors are same but the coefficient of these predictor has changed.

## 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**

### Root Mean Square Error

- The RMSE value in case of Ridge & Lasso is as below
  - Ridge - 0.1164209999665545
  - Lasso - 0.11577837113309898

### Mean Squared Error

- The Mean Squared error in case of Ridge and Lasso are:
  - Ridge - 0.013553849233212482
  - Lasso - 0.013404631222233608
- The Mean Squared Error of Lasso is slightly lower than that of Ridge

### R2 score

- The R2 score for Ridge regression
  - y_train_pred - 0.9158977425870722
  - y_test_pred - 0.9025764131981605
- The R2 score for Lasso regression
  - y_train_pred - 0.9158868899977957
  - y_test_pred - 0.9036489759509891
- Though the model performance by Ridge Regression was better in terms of R2 values of Train and Test,it is better to use Lasso, since it brings and assigns a zero value to insignificant features, enabling us to choosethe predictive variables.It is always advisable to use simple yet robust model.Equation can be formulated using the features and coefficients obtained by Lasso.
- Log (Y) = Constant + 0.110941(MSZoning_RL) + 0.103085(GrLivArea) + 0.079024(MSZoning_RM) + 0.075035(OverallQual) + 0.072357(MSZoning_FV) + 0.050725(TotalBsmtSF) + 0.043987(OverallCond) + 0.040826(Foundation_PConc) + 0.040208(GarageCars) + 0.038377(BsmtFinSF1) + Error term (RSS + alpha * (sum of absolute value of coefficients)
- Suggestions for Surprise Housing is to keep a check on these predictors affecting the price of the house.The higher values of positive coeeficients suggest a high sale value.Some of those features are:-
  - GrLivArea
  - OverallQual
  - OverallCond
  - TotalBsmtSF
  - GarageCars

    - MSZoning_RL
- The higher values of negative coeeficients suggest a decrease in sale value.Some of those features are:-
    - MSSubClass
- When the market value of the property is lower than the Predicted Sale Price, its the time to buy.

Since Lasso helps in feature reduction (as the coefficient value of one of the feature became 0), Lasso has a better edge over Ridge.

Hence based on Lasso, the factors that generally affect the price are the Zoning classification, Living area square feet, Overall quality and condition of the house, Foundation type of the house, Number of cars that can be accomodated in the garage, Total basement area in square feet and the Basement finished square feet area

Therefore, the variables predicted by Lasso in the above bar chart as significant variables for predicting the price of a house.

## 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 five most important predictor variables that will be excluded are:

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. MSZoning_RL

## 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**

The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. 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 generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.

The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much importance should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis.