Problem Statement - Part II

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?

Ans. Optimum value of alpha for ridge regression is 2 and lasso regression is 50.

The changes observed in the model if we choose double the value of alpha for both ridge and lasso are the important predictor variables are same but the coefficient of these predictors are modified.

In ridge regression, the coefficients that have greater values gets penalized by increasing the value of alpha the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

In lasso regression, As the lambda value increases Lasso shrinks some of the coefficients towards zero and it make the variables exactly equal to 0. Lasso also does variable selection. When alphavalue is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

Example: in the above project the variable "BsmtFinSF2" coefficient value is   6437.363663 when alpha value is 2. When alpha value increased to 4 coefficient values reduced to 5267.755329. Whereas, in lasso regression "BsmtFinSF2" coefficient value is 1596.049551 when alpha value is 50 and reduced to zero when alpha value is doubled.

	Ridge	Ridge4	Lasso	Lasso100
BsmtFinSF2	6437.363663	5267.755329	1596.049551	0.000000

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?

Ans. we prefer lasso regerssion eventhoug R2 score values are same in both the cases, because some of the variable's coefficient values are reduced to exactly zero values in lasso so we can remove those variables in the final model selection for example

	ridge	lasso
BsmtUnfSF	5498.192276	0.000000
2ndFlrSF	38238.143335	0.000000
MSZoning_RH	8065.364150	0.000000
MSZoning_RM	4941.629702	-0.000000
Utilities_NoSeWa	-17059.471142	-0.000000

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?

Ans. After removing top five variables 'MSSubClass', 'LotArea', 'LotFrontage', 'OverallQual', 'OverallCond','YearBuilt'
Now, five most important predictor variables are

	Lasso50
MasVnrArea	15785.986292
BsmtFinSF1	51258.207916
BsmtFinSF2	451.829317
TotalBsmtSF	56047.692653
1stFlrSF	31637.167702

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?

Here are some changes we can make to our data:

1. Data collection: Always getting more data helps in better model building, always attempt to get as much data as possible

2. Fix missing values and outliers: If the data has missing values and outliers can lead to inaccurate model. Outliers can affect the mean, median that we are imputing to continuous variables

3. Transform your data. If your data has a very pronounced right tail, try a log transformation. User derived variables and drop the given variable if this adds more value to the data

4. Remove the outliers. This works if there are very few of them and its certain they're anomalies and not worth predicting

5. Feature Selection: Domain knowledge plays an important role in feature selection,additional techniques like data visualization also helps the selecting the features.Statistical parameters like p-Values, VIF can give us significant variables.

6. Algorithm selection: Choosing the right machine learning algorithm is very important to get accurate model.

7. Cross validation: To reduce overfitting user cross validation i.e. leave a sample on which you do not train the model & test the model on this sample before got to the final model