Subjective Questions

Q-1) What is the optimal value of alpha for ridge & lasso regression? what will be the changes in the model if you choose double the value of alpha for both ridge & lasso? what will be the most important predictor variables after the change is implemented?

Ans) The optimal value of alpha for ridge and lasso is 0.001& 2.0 respectively.

There are no changes in the model for my assignment if we choose to double the value of alpha and has been demonstrated in the .ipynb notebook file as well. (Please refer to .ipynb notebook for the demonstration under 'Q-1' heading)

The most important predictor variable for both the models remain same as well i.e 'LotArea'.

Q-2) You have determined the optimal value of lambda for ridge & lasso regression during the assignment. Now which one will you choose to apply and why?

Ans) Both the models (ridge & lasso) are tuned using GridsearchCV to obtain the optimal value of lambda.

For Lasso model the optimal value of lambda/alpha, are obtained at 11 folds of gridsearchcv with 24 values of lambda/alpha which equals to 264 fits with best alpha coming at 0.001. If try to shift the value of folds & alpha/lambda (higher or lower) then the model is either underfitting or overfitting thus making the model to be highly unstable & vulnerable.

For Ridge model the optimal value of lambda/alpha, are obtained at 3 folds of gridsearchcv with 25 values of lambda/alpha which equals to 75 fits with best alpha coming at 2.0. If we try to shift the value of folds then value of alpha/lambda is also shifting thus making the model to underfit or overfit thus making the model to be highly unstable and vulnerable.

I will choose to use Lasso model as it adds penalty which is equal to the absolute value of the magnitude of the coefficients so as to keep a check on the large coefficient to avoid model to overfit the data.

Q – 3) After building the model, you realized that the five most important predictor variable in the lasso model are not available in the incoming data. You will have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Ans) Please refer to .ipynb notebook file with heading "Question - 3"

Q-4) How can you make sure that a model is robust & generalizable? What are the implications of the same for the accuracy of the model and why?

Ans) A model is said to be robust when the target/dependent variable is accurate despite of change in the one or two independent variable i.e it should have low bias value.

A model is said to be generalizable if it is able to correctly identify the hidden patterns in the unseen data i.e it should have low variance.

We can achieve a robust and generalizable model by performing following methods:

#) There is no defined quantity as to how much data is correct for building a good ml models, it good to have a good amount of data to identify the hidden patterns and make a generalized model and thus avoid underfitting situation.

#) Try to treat missing values and reduce the outliers in the data as it's presence has the capacity to make the model biased.

#) Perform feature transformation, feature creation to make more meaningful columns.

#) Select the feature based on domain knowledge, visualization & statistical parameters like PCA etc.

#) Good to have multiple models built on multiple ml algo so as to identify the best algo in terms of output.

#) Tune the model for the optimal values.