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 1

For both Ridge and Lasso, when I double the alpha, the R^2 slightly decreases from 0.94 to 0.93 on train datasets, while it remains the same on test datasets.

In term of features, double alpha in Ridge (in this case) doesn't affect number of features, which is of course correct. In contrast, double alpha in Lasso results in a significant decreases in number of features (from 113 to 82 features).

In fact, even I increase alpha 10 times, the R^2 for both Ridge and Lasso only decrease by 4%, while number of features in Lasso decrease dramatically (from 113 features to 38 features). In my point of view, predictive model should use less as less variable as posible while we still maintain R-squared and MSE.

Therefore, in this assignment, I decide to use alpha is ten times higher than proposed by Lasso. The final alpha for Lasso is 0.001 (the proposed is 0.0001).

When I increase alpha in Lasso, following features are always appear top

- YearBuilt
- OverallQual
- MasVnrArea
- BsmtFinSF1
- TotalBsmtSF

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

As explained above, Lasso is chosen over Ridge because of following reason

- In term of R^2, RSS, MSE both model give similar result
- Lasso reduce numbers of predictors, which in my point of view is good. Less predictors make model more simpler, more generalised and easier to explain to business.

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

In my model, 5 most important variables are in my current model.

- YearBuilt
- OverallQual
- MasVnrArea
- BsmtFinSF1
- TotalBsmtSF

Because this is machine learning assignment, I let the algorithm do feature selection itself. However, I think we should combine both machine learning approach and expertise approach. Here are steps:

- Box plots should be used for all categorical variables with y axis was SalePrice.
- According to plots, next I pick out varibles that had sale price vary among categories.

I think it may be useful as it's nonsense to pass variables that may not impact on dependent variable.

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

Here are following things I do in this assignment to ensure model is robust and generalisable:

- Evaluating model on test dataset.
- Comparing the R^2 on test dataset with train dataset, to ensure the test accuracy is not lesser than the training score.
- Slightly increase the optimum alpha proposed by Lasso, to see the traded-off between bias and variance. The simpler the model, the more bias, the less variance and more generalisable.
- When I reach certain alpha, the accuracy decrease significant, it implies that the model is now more generalisable but less robust and accuracy.