Surprise Housing : Linear Regression for Housing Data of Australian Market

Assignment Part II

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?

Ans:

Optimal Values for alpha captured in the regression:
- Ridge Regression : 100
- Lasso Regression : 500

When the alpha was doubled,
- The coefficients moved more towards 0.
- The MSE for train dataset increased slightly.
- The R2 score for train and test datasets did not change much.
- The MSE for test dataset did not change much.
- The number of predictor variables dropped from 41 to 28 in case of Lasso regression

The most important 5 predictor variables after the change were:
- GrLivArea :- Above grade (ground) living area square feet
- OverallQual_8 :- The overall material and finish of the house (Very Good)
- OverallQual_9 :- The overall material and finish of the house (Excellent)
- YearBuilt :- Original construction date
- BsmtFinSF1 :- Type 1 finished square feet

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?

Ans:

Optimal Values for alpha captured in the regression:
- Ridge Regression : 100
- Lasso Regression : 500

We applied the respective optimal values of alpha for Ridge and Lasso and got the below results.

Ridge Regression¶
- R2 Score Train : 0.897
- R2 Score Test : 0.876

Lasso Regression¶
- R2 Score Train : 0.897
- R2 Score Test : 0.876

The R2 score is almost same. But with Lasso regression, since the feature selection also happens, the model can be explained better. We will go with Lasso.

Question 3

After building the model, you realized 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 the five most important predictor variables, which are:
- GrLivArea :- Above grade (ground) living area square feet
- OverallQual_8 :- The overall material and finish of the house (Very Good)
- OverallQual_9 :- The overall material and finish of the house (Excellent)
- YearBuilt :- Original construction date
- BsmtFinSF1 :- Type 1 finished square feet

We reran Lasso regression for the same alpha (500), the new most significant five predictors now are:

1. OverallQual_5 :- The overall material and finish of the house (Average)
2. OverallQual_6 :- The overall material and finish of the house (Above Average)
3. 2ndFlrSF :- Second floor square feet
4. TotalBsmtSF :- Total square feet of basement area
5. OverallQual_4 :- The overall material and finish of the house (Below Average)

Question 4

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

Ans:

Robust => The model is not impacted by the outliers in the training data.
Generalizable => The test accuracy is not much different than the training accuracy.

The higher weightage (coefficient values) in the model would result in higher deviations in the test results. Especially when there are outliers. Outlier analysis needs to be done carefully and all unnecessary outliers should be removed from the training data before the model is built.

Since the Lasso regression tries to keep the coefficients to the minimum, it is more robust than any other models.