House Price Prediction Case Study

Using Ridge and Lasso Regression

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: Below are the optimal values of alpha

Ridge Optimal Alpha = 500
Lasso Optimal Alpha = 1000

This seems Optimal Alpha for Lasso is two times of Ridge Regression.
Going with simple Linear Regression and find the Correlation among columns then seems like SalePrice is highly correlated to these independent variables – TotalBsmtSF, 1stFlrSF, GarageArea, GrLivArea. But as assumptions are not ideally met so going with Non-Linear model.

Lasso is better fit comparatively then Ridge so going with Lasso Regression Model.
Most important predictor variables after the change are implemented:

```
GrLivArea               26423.179600
OverallQual_9           12419.596904
OverallQual_10          11454.647506
TotalBsmtSF              8594.310640
OverallQual_8            7863.866889
BsmtFinSF1               7402.068104
GarageCars_3             5226.956847
SaleType_New             4131.093477
```

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: Data Analysis comparison of Ridge and Lasso Regression models

1. Ridge Optimal Alpha = 500
   R2 Score (Train) = .96 and R2 Score (Test) = .84. This seems a good difference so model may not fit accurately as expected on test data.
   MSE (Train) = 1.563905e+04 and MSE (Test) = 3.141967e+04.

2. Lasso Optimal Alpha = 1000
   R2 Score (Train) = .95 and R2 Score (Test) = .87. This seems a good difference so model may not fit accurately as expected on test data. But still better than Ridge
   MSE (Train) = 1.731805e+04 and MSE (Test) = 2.790585e+04. Better than Ridge

Seems Alpha is high in both the regression models so looks like a Underfitting model with lots of noisy predictor variables.
Lasso is better fit then Ridge so going with Lasso Regression Model.

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: With the existing Lasso Regression Model analysis, next five important predictors are –

```
BsmtFinSF1               7402.068104
GarageCars_3             5226.956847
SaleType_New             4131.093477
KitchenAbvGr_1           3975.906291
MasVnrArea_1170.0        3929.810829
```

But we can’t say upfront that the above five are the most important predictors variable in the new Lasso Model because there are many correlated independent variables and if some of them are excluded then model outcome may change altogether.

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?

Answer: Due to the data complexity though it works well for training data but perform poor for test data. This is a classic example of Overfitting model. And either we need remove some of the independent variables that are not much correlated or need to regularize the data. To reduce the complexity and regularize the data by reducing the coefficient to get the best lambda. This way it helps to reduce the impact of high variance in case of any small change in test data and results expected would be same both with train and test data. This way we can say that model is more robust and generalized.

From the model complexity, we should look for a point of intersection of bias and variance. So, there is a trade-off between bias and variance. Ideally We should look for lowest total errors means low bias and low variance. But to manage the model complexity, it should be neither too high nor to less.