Advanced Regression Assignment – Surprise Housing Case Study

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: The optimal value of alpha for ridge regression is 9 & for lasso regression is 0.001.

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.940015	0.920280
R2 Score (Test)	0.927004	0.925048
RSS (Train)	8.539168	11.348530
RSS (Test)	2.847327	2.923659
MSE (Train)	0.007311	0.009716
MSE (Test)	0.009751	0.010013
RMSE (Train)	0.085504	0.098571
RMSE (Test)	0.098748	0.100063

Figure 1: with alpha 9 & 0.001

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.934606	0.904667
R2 Score (Test)	0.927917	0.911313
RSS (Train)	9.309120	13.571188
RSS (Test)	2.811735	3.459390
MSE (Train)	0.007970	0.011619
MSE (Test)	0.009629	0.011847
RMSE (Train)	0.089276	0.107792
RMSE (Test)	0.098129	0.108845

Figure 2: with double values of alpha

Changes in Ridge Regression metrics:

- R2 score of train set decreased from 0.94 to 0.93
- R2 score of test set remained same at 0.92

Changes in Lasso metrics:

- R2 score of train set decreased from 0.92 to 0.90
- R2 score of test set decreased from 0.92 to 0.91.

Further, the most important predictor variables after we double the alpha values are:-

- GrLivArea
- OverallQual_8
- OverallQual_9
- Neighborhood_Crawfor
- Functional_Typ
- Exterior1st_BrkFace
- OverallCond_9
- TotalBsmtSF
- CentralAir_

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: The model we will choose to apply will depend on the use case. If we have too many variables and one of our primary goals is feature selection, then we will use Lasso. If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we will use Ridge Regression.

As we add more predictors to the model, the training r2 score increases whereas the test r2 score does not increase. This is because of overfitting of the model i.e. model has memorized the training data. To overcome this issue, we use regularized regression. Regularized regression penalizes the model for its complexity.

Both ridge and lasso regression does feature selection. Ridge uses L2 regularization technique regression model whereas Lasso uses L1 regularization technique.

Optimal value of alpha for our models are as below:

- Ridge regression – 9
- Lasso regression – 0.001

For optimal value of alpha in ridge and lasso regression, r2 score for both training and test data was calculated.

In the current scenario, we shall go with Ridge regression has this gives us much consistent values with both train and test data.

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: Here, we will drop the top 5 features in Lasso model and build the model again.

Top 5 Lasso predictors were:

OverallQual_9, GrLivArea, OverallQual_8, Neighborhood_Crawfor, & Functional_Typ

After dropping our top 5 lasso predictors, we get the following new top 5 predictors:-

2ndFlrSF, CentralAir_Y, 1stFlrSF, MSSubClass_70 & Neighborhood_Somerst.

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:

Per Occam's razor – model should be as simple as necessary. Advantages of simple model are as below:

- Generalizability
- Robustness
- Making few assumptions
- Less data is required for learning

When a model is resilient, any fluctuation in the data has little effect on its performance. A generalizable model may adapt to additional, previously unknown data obtained from the same distribution as the one used to generate the model.

To ensure that a model is resilient and generalizable, we must ensure that it does not over-fit. This is because an overfitting model has a very high variance, and even the smallest change in data has a large impact on model prediction.

In general, we must establish a happy medium between model correctness and complexity. Regularisation techniques such as Ridge Regression and Lasso can help with this.