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?

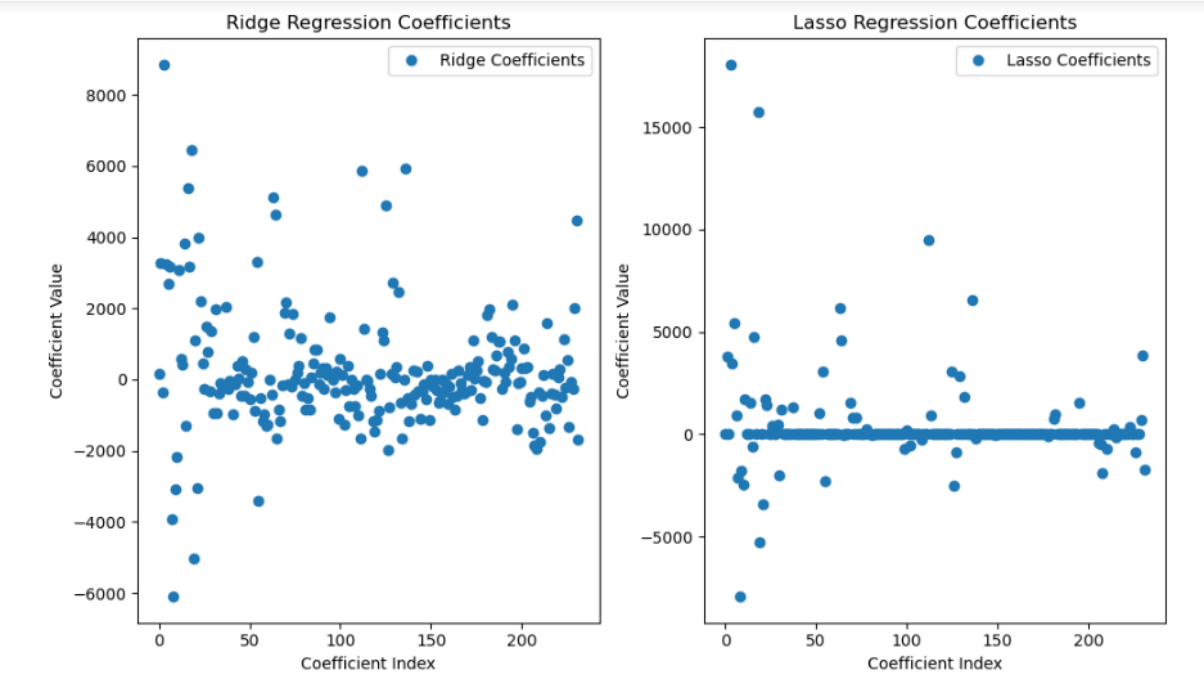
Answer

Optimal value of alpha for ridge and lasso regression

Ridge : 500

Lasso : 1500

	Ridge	Lasso
r2_train	0.88	0.87
r2_test	0.85	0.85
rss_val_y_train	7.47241E+11	8.13475E+11
rss_val_y_test	4.24132E+11	4.2889E+11
mse_val_y_train	731872050.2	796743840.8
mse_val_y_test	968338669.9	979202021.3

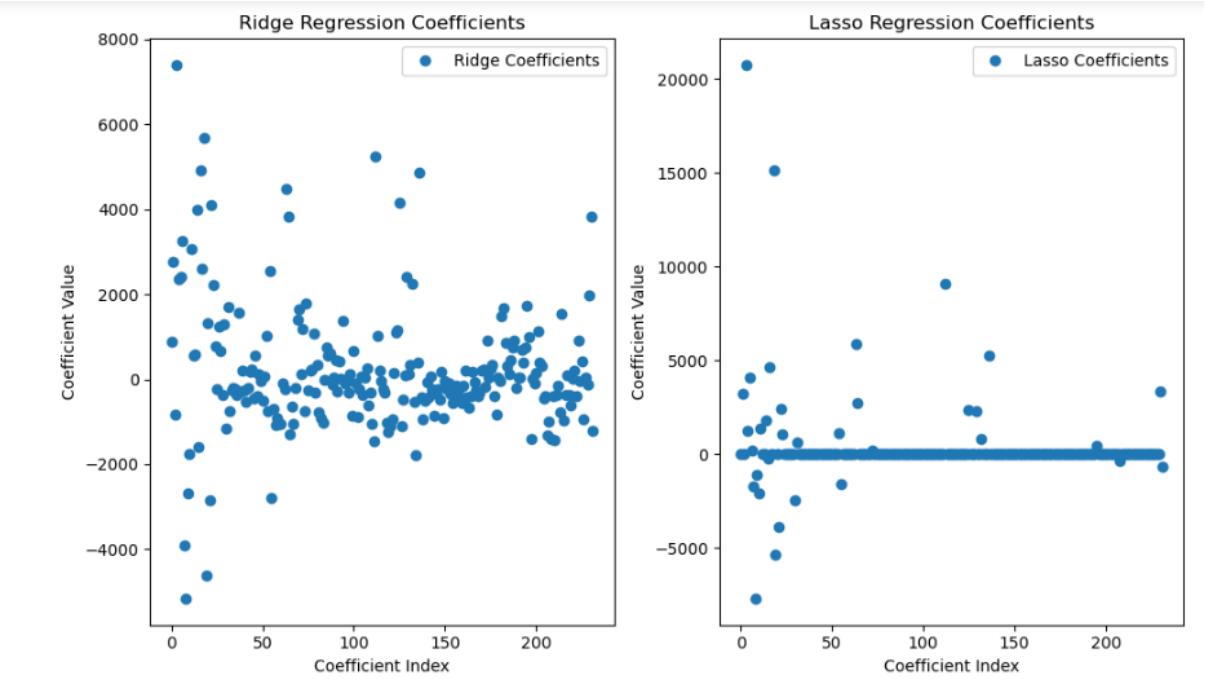


After doubling the value of alpha for both ridge and lasso

Ridge : 1000

Lasso : 3000

	Ridge	Lasso
r2_train	0.86	0.85
r2_test	0.84	0.84
rss_val_y_train	8.600878e+11	9.678308e+11
rss_val_y_test	4.501958e+11	4.538654e+11
mse_val_y_train	2.902408e+04	3.078838e+04
mse_val_y_test	3.206001e+04	3.219041e+04



Important predictor variables after the change is implemented

- GrLivArea
- OverallQual
- GarageCars_3
- FullBath_3l
- Neighborhood_NoRidge (Northridge)
- YearBuilt
- BsmtQual (-ve coeff)

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

Lasso will be chosen here in this House Pricing Prediction model regularization.

Reason**:**

Number of features are very high after spreading out the nominal categorical variables in to dummies. Its around 235 features. Lasso reduced most of the features by setting few of the features coefficients as 0 after penalizing the model as part of regularization.

As number of feature are more and feature selection is more important here, Lasso is preferred.

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

OverallCond

1stFlrSF

Neighborhood_Crawfor (CrawFord)

Neighborhood_NridgHt (Northridge Heights)

Fireplaces_2 (Fire places 2)

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

Model is robust and generalized after applying below strategies

1.Cross Validation:

Applied GridSerach Cross Validation with 5 folds. This ensures Model is trained on various

split data sets and come out as more generalized.

2.Applied Regularization techniques:

This avoids overfitting in the model. so, model can be more robust and reliable in predicting

unseen data.

	Ridge	**Lasso**
r2_train	0.88	0.87
r2_test	0.85	0.85

3.Hyperparameter Tuning:

Grid Search hyperparameter tuning is applied to find the suitable value for model generation.

4.Feature Selection:

Lasso ensured the feature selection besides preventing the overfitting.

More than 50% of the features became Zero.

These strategies ensured accuracy as well.

1.R2 value is consistent in both train and test data.

2.Coefficients plot graphs are shown close to zero after regularizaton.

3.Hyperparaamter tuning given best alpha value that prevents overfitting in the model.

4.Feature selection happened through Lasso and coefficients close to zero in Ridge, this can ensure that no multicollinearity in the independent variables.