

Surprise Housing Price Case Study

Vasudha Srinivasaiah

IIIT-B Machine Learning & AI Program -Dec-2023

4/24/24

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

Q1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to 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 The optimal value of alpha for ridge and lasso regression

Ridge Alpha: 9.739119536819391

Lasso Alpha : 0.018301662744060254

Before Changing the model evaluation metrics are

```
final_metric
```

	Metric	Linear Regression	Lasso Regression	Ridge Regression
0	R2 Score (Train)	0.832305	8.247994e-01	0.838761
1	R2 Score (Test)	0.841391	8.481174e-01	0.837302
2	RSS (Train)	161.652803	1.733252e+06	155.428951
3	RSS (Test)	78.651865	3.638304e+05	80.680024
4	MSE (Train)	0.397710	4.065124e-01	0.389978
5	MSE (Test)	0.423758	4.146753e-01	0.429186

Post Changes to the Model by changing the alpha to double the previous values:

Ridge Alpha: 19.47823907363878

Lasso Alpha : 0.0366033254881204

```
final_metric
```

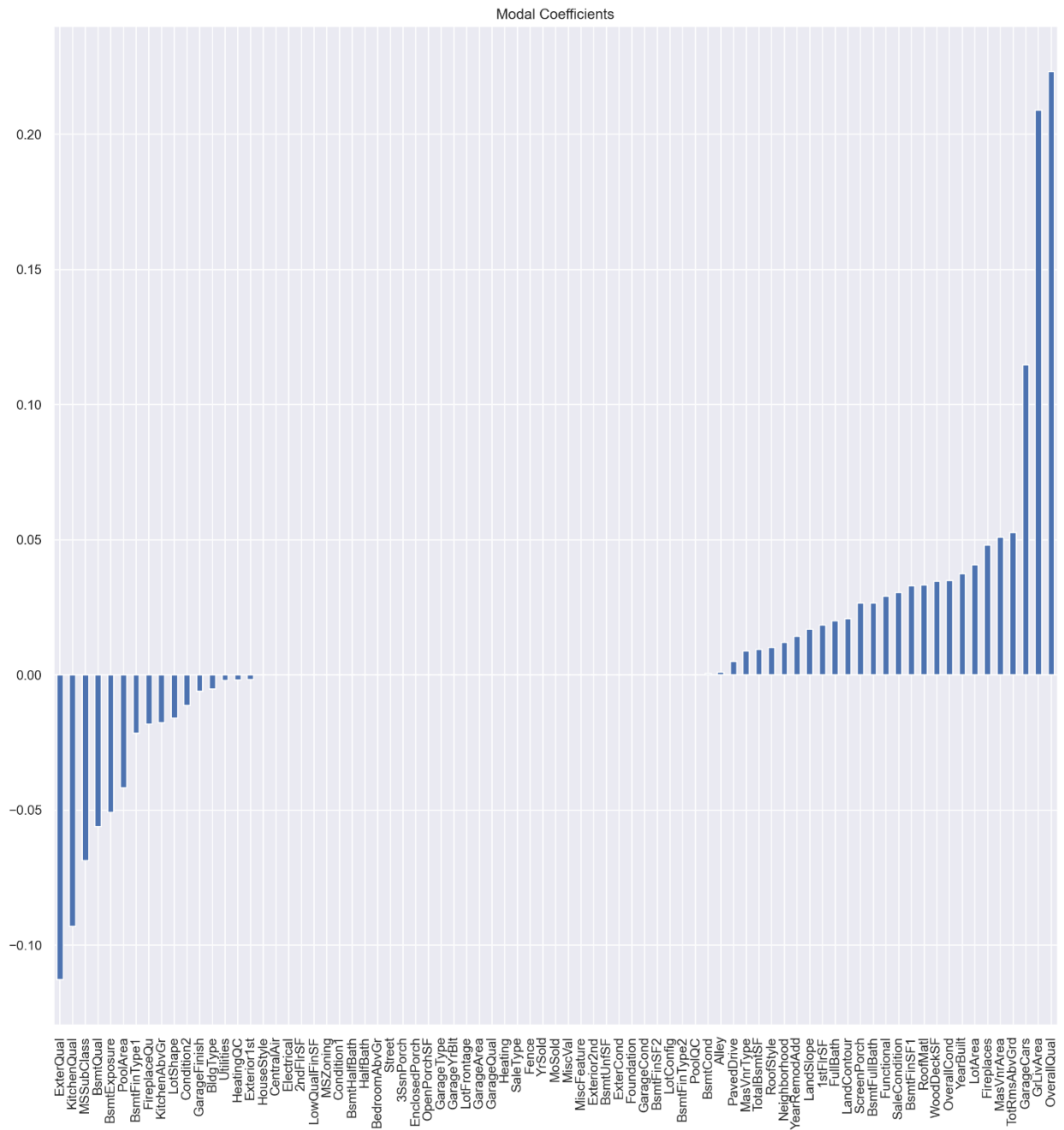
:

	Metric	Linear Regression	Lasso Regression	Ridge Regression	Lasso Regression-2*Alpha	Ridge Regression-2*Alpha
0	R2 Score (Train)	0.832305	8.247994e-01	0.838763	8.070822e-01	0.838665
1	R2 Score (Test)	0.841391	8.481174e-01	0.837300	8.447187e-01	0.837354
2	RSS (Train)	161.652803	1.733252e+06	155.427143	1.675269e+06	155.522072
3	RSS (Test)	78.651865	3.638304e+05	80.680765	3.538901e+05	80.654057
4	MSE (Train)	0.158173	1.652524e-01	0.152081	1.819635e-01	0.152174
5	MSE (Test)	0.179570	1.719556e-01	0.184203	1.758035e-01	0.429117
6	RMSE (Train)	0.397710	4.065124e-01	0.389976	4.265718e-01	0.390095
7	RMSE (Test)	0.423758	4.146753e-01	0.429188	4.192893e-01	0.429117

It is evident from the metrics that there is a marginal increase in the Train MSE and R2 score but a reduction in the Test data MSE and R2 Score.

The top 10 important variables after the change is implemented are:

OverallQual 0.223239
 GrLivArea 0.208957
 GarageCars 0.114760
 ExterQual 0.112768
 KitchenQual 0.093113
 MSSubClass 0.068854
 BsmtQual 0.056108
 TotRmsAbvGrd 0.052717
 MasVnrArea 0.050867
 BsmtExposure 0.050797



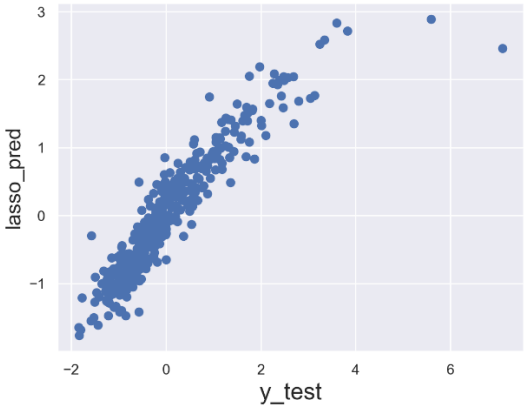

2. You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:

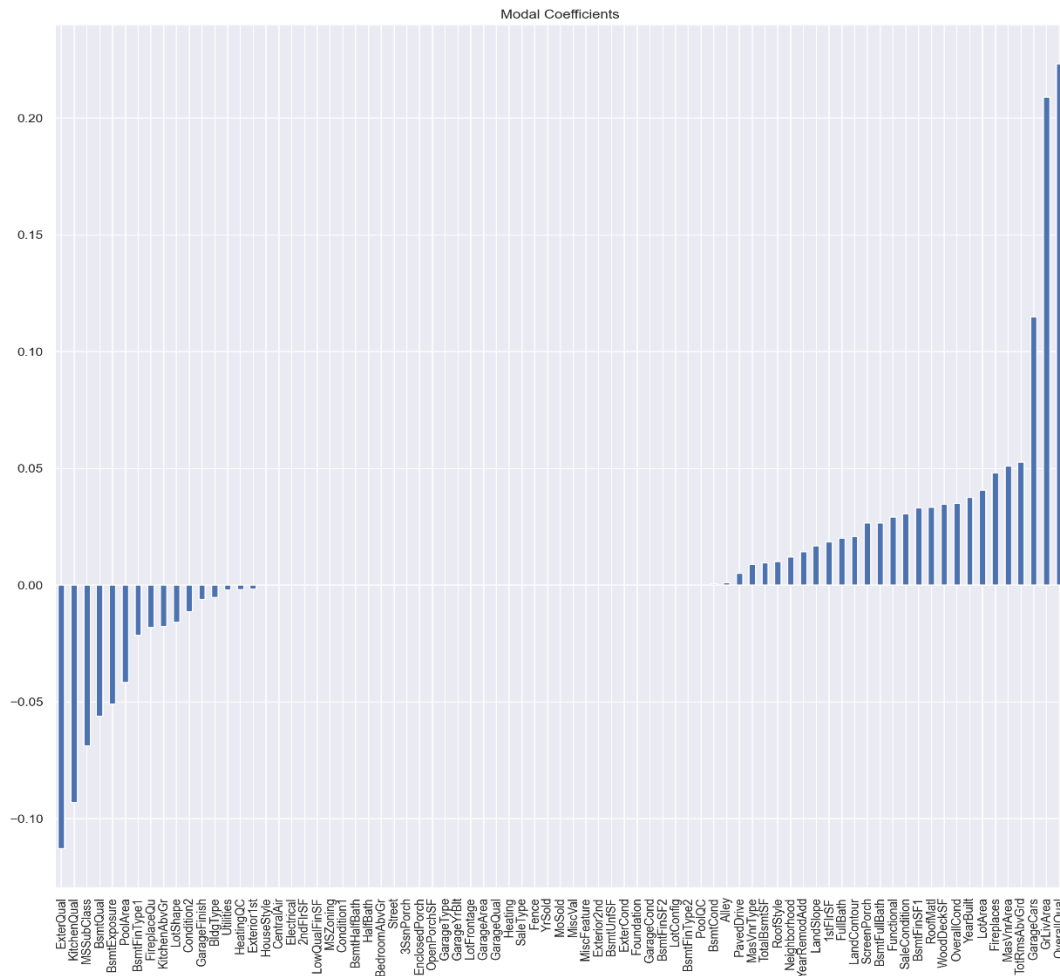
Lasso Regression	Ridge Regression
<ul style="list-style-type: none"> Lasso regression performs both variable selection and regularization by adding an L1 penalty term to the loss function. It tends to shrink the coefficients of less important features to exactly zero, effectively removing them from the model. Lasso can be useful when you suspect that many of your features are irrelevant or redundant, as it can automatically select a subset of features that are most relevant for prediction. 	<ul style="list-style-type: none"> Ridge regression adds an L2 penalty term to the loss function, which penalizes the coefficients based on their magnitude. Unlike Lasso, Ridge regression generally does not set coefficients exactly to zero but rather shrinks them towards zero. Ridge regression can be useful when you have many moderately important features and want to prevent overfitting by penalizing large coefficients.

Almost 79 predictors are available in the data set. From Lasso 44 top predictors were enlisted for the model. Hence Lasso can be used for the feature selection. The number of predictors influencing the Sales Price prediction is high, and the model's metrics indicate that The R2 score is approximately the same for train and test data for Ridge. However, the MSE on test data is comparatively low on the test data under Lasso. Hence would consider the Lasso model for Prediction

The model evaluation resulted below

Lasso Regression	Ridge Regression
<p>y_test vs lasso_pred</p> 	<p>y_test vs ridge_pred</p> 





Q3. 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?

Answer:

Before change -Significant predictors top 5 predictors are :

Top Features selected by Lasso along with their Coefficients:

OverallQual	0.223239
GrLivArea	0.208957
GarageCars	0.114760
ExterQual	0.112768
KitchenQual	0.093113

New model

The Optimal The optimal value of alpha for ridge and lasso regression

Ridge Alpha: 9.707582653824577

Lasso Alpha: 0.01912548934600116

After Change - Significant predictors The top 5 predictors are :

The top new Features selected with their Coefficients are as given herewith:

```
2ndFlrSF      0.159079
1stFlrSF      0.154323
BsmtQual      0.131549
GarageArea    0.127840
TotRmsAbvGrd  0.121038
```

New metrics are given below

	Metric	Lasso Regression3	Ridge Regression-3
0	R2 Score (Train)	7.879023e-01	0.805254
1	R2 Score (Test)	8.203990e-01	0.816979
2	RSS (Train)	1.683996e+06	187.728863
3	RSS (Test)	3.531138e+05	90.757614
4	MSE (Train)	2.000544e-01	0.183688
5	MSE (Test)	2.033373e-01	0.455202
6	RMSE (Train)	4.472744e-01	0.428588
7	RMSE (Test)	4.509294e-01	0.455202

It is observed that the R2 Score has decreased in both cases and the MSE has increased.

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:

Various methods can be used to make sure a model is reliable and generalizable. Some of them are listed below

- **Cross-Validation:** The model's performance may be assessed by splitting the data set into training and validation sets and using cross-validation methods like k-fold cross-validation. This gives an estimate of the robustness of the model and aids in evaluating how well it generalizes to new data.
- **Hyperparameter tuning:** The regularization parameter (λ) can be varied to determine the best value that strikes a compromise between the model's bias and variance. Techniques like grid search and randomized search with cross-validation can be used for this.
- **Feature Scaling:** By normalizing or standardizing the predictor variables, feature scaling allows us to ensure that the regularization penalty is applied uniformly to all variables.
- **Feature Selection:** Lasso regression is a useful tool for feature selection because it can efficiently exclude less significant variables from the model by shrinking coefficients to exactly zero. As a result, the model may become more frugal, less prone to overfitting, and more capable of generalizing to new data.

- **Composite Methods:** Utilizing strategies like model stacking or averaging, combine several models trained with Lasso or Ridge regression. The shortcomings of individual models can be lessened with the use of collective approaches, which can enhance performance and generalization.

The implications of ensuring model robustness and generalizability for the accuracy of the model are listed below

- By ensuring that the model is robust and generalizable, we hope to increase its performance on previously unseen data. A robust and generalizable model is less likely to overfit the training data and can capture the underlying patterns in the data, resulting in superior predictive performance on new, previously unknown data.
- Regularization strategies like Lasso and Ridge regression assist in achieving a balance between the two in the model. Regularization strategies lower the likelihood of overfitting (high variance) while preserving the model's ability to identify significant patterns in the data (low bias) by penalizing the model's complexity. Better generalization performance and, eventually, greater accuracy on unobserved data result from this trade-off.