

Surprise housing assignment

SUBJECTIVE QUESTIONS by PRERNA SHUKLA



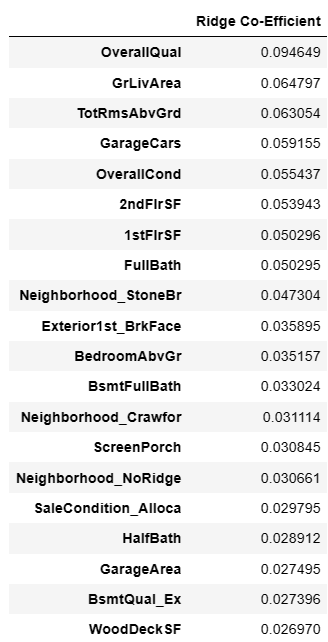
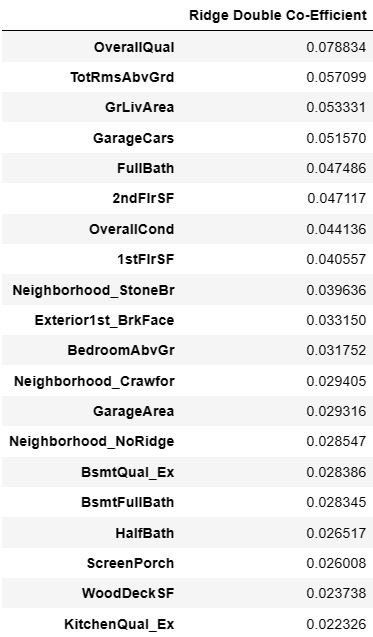
**Question 1**

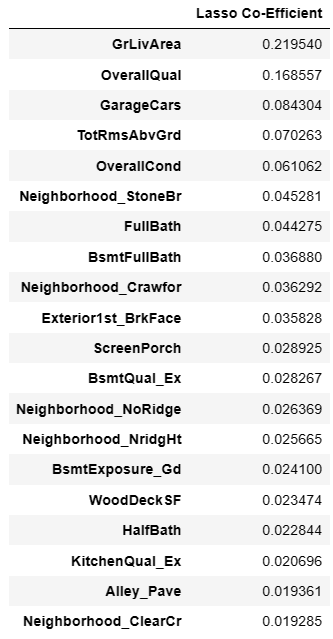
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

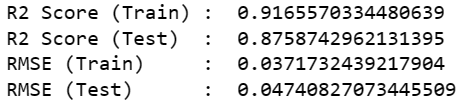
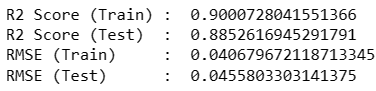
**Answer:**

The optimal value for the Ridge model is determined to be 4.0 and Lasso is 0.0002. Upon doubling the alpha value to 8 and 0.0004 respectively, the following observations are noted:

1. A marginal decrease in R2 score and a slight increase in RMSE values are observed in the train data.

1. R2 score Negligible alterations are noticed in the coefficients of the features. For instance, there's a marginal change in the coefficient values of features.
2. Rigid:
3. 
4. Lasso:
5. 
6. As differences are very small, we do not see major change in models after doubling the value of alpha.

**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?

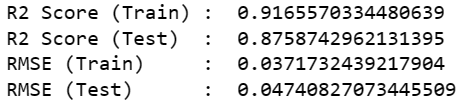
**Answer:**

1. The optimum lambda value in case of Ridge and Lasso is as follows:

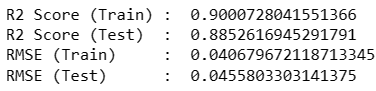


1. The RMSEs are as follows :

Rigid:



Lasso:



1. From above we observe that Mean square error are almost same. So, both have almost same accuracy.
2. As Lasso helps in feature reduction, Lasso has a better edge over Ridge and should be used as the final model. We can choose and apply Lasso regression.

**Question 3**

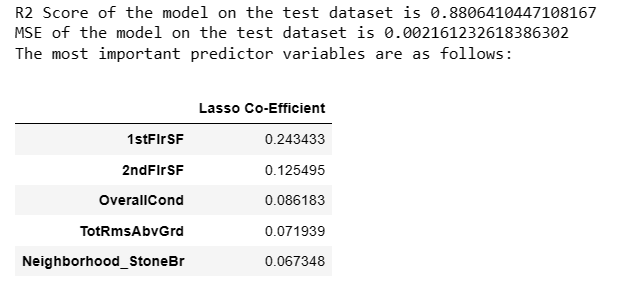
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:

The top five predictors in the existing model are :

* OverallQual,
* GrLivArea,
* GarageCars,
* GarageArea,
* TotalBsmtSF

After dropping these columns and re-building model again with Lasso regression, R2 score dropped to 88.06% and MSE is 0.002 only.



We got new top 5 columns as :

* 1stFlrSF
* 2ndFlrSF
* OverallCond
* TotRmsAbvGrd
* Neighborhood\_StoneBr

**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:

Occam's Razor, a principle in model selection, suggests that among models demonstrating similar performance in limited training or test data, the simpler one is usually preferred on the test data. This preference for simplicity carries several advantages:

- **Applicability**: Simpler models tend to be more versatile and widely applicable across various scenarios.

- **Efficiency**: They often require fewer training samples, making them easier and more efficient to train.

- **Robustness:** Simpler models exhibit more robust behavior compared to complex ones. Complex models can drastically change with variations in the training data, while simpler models have lower variance and higher bias.

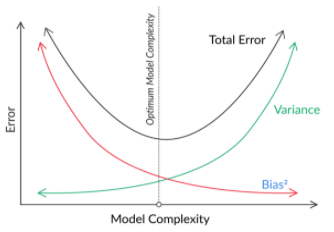
The balance between bias and variance further illustrates the trade-off between complexity and stability in models:

- **Complexity Impact**: Complex models, highly sensitive to changes in the dataset, can be unstable, requiring constant adjustments. Conversely, simpler models, capturing overarching patterns, tend to remain more consistent despite alterations in the dataset.

- **Bias and Variance:** Bias reflects a model's accuracy on test data. While a complex model can predict accurately with sufficient training data, excessively simplistic models (e.g., those providing the same answer for all inputs) possess high bias and yield poor predictions.

- **Variance Variation:** Variance signifies the extent of model changes concerning variations in the training data.

Balancing bias and variance is pivotal for maintaining model accuracy and minimizing overall error. This balance can be depicted graphically, highlighting the optimal trade-off that enhances a model's performance.



Therefore, achieving a robust and generalizable model involves seeking simplicity without oversimplification. Regularization techniques aid in this pursuit by maintaining a delicate equilibrium—keeping the model simple while ensuring it retains the necessary complexity to be practically useful. In regression, regularization incorporates additional terms into the cost function, penalizing extreme parameter values to foster a balanced model.

By comprehending the interplay between simplicity, complexity, bias, and variance, model builders can steer their models towards a sweet spot. This sweet spot optimizes accuracy while guarding against overfitting or oversimplification, ultimately fostering a more reliable and adaptable model for diverse real-world applications.