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.

Ridge Co-Efficient		Ridge Double Co-Efficient	
OverallQual	0.094649	OverallQual	0.078834
GrLivArea	0.064797	TotRmsAbvGrd	0.057099
TotRmsAbvGrd	0.063054	GrLivArea	0.053331
GarageCars	0.059155	GarageCars	0.051570
OverallCond	0.055437	FullBath	0.047486
2ndFlrSF	0.053943	2ndFlrSF	0.047117
1stFlrSF	0.050296	OverallCond	0.044136
FullBath	0.050295	1stFlrSF	0.040557
Neighborhood_StoneBr	0.047304	Neighborhood_StoneBr	0.039636
Exterior1st_BrkFace	0.035895	Exterior1st_BrkFace	0.033150
BedroomAbvGr	0.035157	BedroomAbvGr	0.031752
BsmtFullBath	0.033024	Neighborhood_Crawfor	0.029405
Neighborhood_Crawfor	0.031114	GarageArea	0.029316
ScreenPorch	0.030845	Neighborhood_NoRidge	0.028547
Neighborhood_NoRidge	0.030661	BsmtQual_Ex	0.028386
SaleCondition_Alloca	0.029795	BsmtFullBath	0.028345
HalfBath	0.028912	HalfBath	0.026517
GarageArea	0.027495	ScreenPorch	0.026008
BsmtQual_Ex	0.027396	WoodDeckSF	0.023738
WoodDeckSF	0.026970	KitchenQual_Ex	0.022326

	Lasso Co-Efficient			Lasso Double Co-Efficient
GrLivArea	0.219540		GrLivArea	0.203525
OverallQual	0.168557		OverallQual	0.181104
GarageCars	0.084304		GarageCars	0.092888
TotRmsAbvGrd	0.070263		TotRmsAbvGrd	0.063752
OverallCond	0.061062		OverallCond	0.050163
Neighborhood_StoneBr	0.045281		FullBath	0.041380
FullBath	0.044275		BsmtFullBath	0.034365
BsmtFullBath	0.036880		BsmtQual_Ex	0.031867
Neighborhood_Crawfor	0.036292		Neighborhood_StoneBr	0.030302
Exterior1st_BrkFace	0.035828		Exterior1st_BrkFace	0.029935
ScreenPorch	0.028925		Neighborhood_Crawfor	0.029263
BsmtQual_Ex	0.028267		BsmtExposure_Gd	0.024624
Neighborhood_NoRidge	0.026369		KitchenQual_Ex	0.020055
Neighborhood_NridgHt	0.025665		ScreenPorch	0.018153
BsmtExposure_Gd	0.024100		Neighborhood_NoRidge	0.017406
WoodDeckSF	0.023474		HalfBath	0.017372
HalfBath	0.022844		Condition1_Norm	0.016494
KitchenQual_Ex	0.020696		WoodDeckSF	0.016352
Alley_Pave	0.019361		Neighborhood_NridgHt	0.015434
Neighborhood_ClearCr	0.019285		BldgType_1Fam	0.015371

2. 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.

3. Rigid:

```
R2 Score (Train) :  0.9165570334480639
R2 Score (Test)  :  0.8758742962131395
RMSE (Train)     :  0.0371732439217904
RMSE (Test)      :  0.04740827073445509
```
4.

5. Lasso:

```
R2 Score (Train) :  0.9000728041551366
R2 Score (Test)  :  0.8852616945291791
RMSE (Train)     :  0.040679672118713345
RMSE (Test)      :  0.0455803303141375
```
6.

7. 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:

   ```
   Optimal value for Ridge is  4.0  and Lasso is 0.0002
   ```

2. The RMSEs are as follows :

   Rigid:
   ```
   R2 Score (Train) :  0.9165570334480639
   R2 Score (Test)  :  0.8758742962131395
   RMSE (Train)     :  0.0371732439217904
   RMSE (Test)      :  0.04740827073445509
   ```

   Lasso:
   ```
   R2 Score (Train) :  0.9000728041551366
   R2 Score (Test)  :  0.8852616945291791
   RMSE (Train)     :  0.040679672118713345
   RMSE (Test)      :  0.0455803303141375
   ```

3. From above we observe that Mean square error are almost same. So, both have almost same accuracy.

4. 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.

```
R2 Score of the model on the test dataset is 0.8806410447108167
MSE of the model on the test dataset is 0.002161232618386302
The most important predictor variables are as follows:
```

	Lasso Co-Efficient
1stFlrSF	0.243433
2ndFlrSF	0.125495
OverallCond	0.086183
TotRmsAbvGrd	0.071939
Neighborhood_StoneBr	0.067348

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.
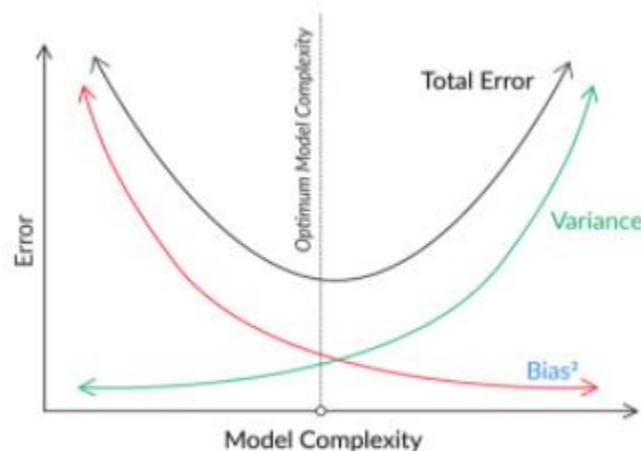
Therefore, achieving a robust and generalizable model involves seeking simplicity without oversimplification. Regularization techniques aid in this pursuit by maintaining a delicate equilibrium—keeping the model simple while ensuring it retains the necessary

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.