Assignment Part-II

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?

Ans: The optimal parameter for Ridge regression, denoted as alpha, is determined to be 2, while for Lasso regression, the optimal value is 0.001. With these chosen alpha values, the R2 score of the model is approximately 0.83.

Upon doubling the alpha values in both Ridge and Lasso, the predictive accuracy remains consistent at around 0.82. However, there is a subtle alteration in the coefficient values. The updated model is showcased in the Jupyter notebook, highlighting the modifications in the coefficients.

Ridge Regression

	Ridge Co-Efficient
Total_sqr_footage	0.169122
GarageArea	0.101585
TotRmsAbvGrd	0.067348
OverallCond	0.047652
LotArea	0.043941
CentralAir_Y	0.032034
LotFrontage	0.031772
Total_porch_sf	0.031639
Neighborhood_StoneBr	0.029093
Alley_Pave	0.024270
OpenPorchSF	0.023148
MSSubClass_70	0.022995
RoofMatl_WdShngl	0.022586
Neighborhood_Veenker	0.022410
SaleType_Con	0.022293
HouseStyle_2.5Unf	0.021873
PavedDrive_P	0.020160
KitchenQual_Ex	0.019378
LandContour_HLS	0.018595
SaleType_Oth	0.018123

	Ridge Doubled Alpha Co-Efficient
Total_sqr_footage	0.149028
GarageArea	0.091803
TotRmsAbvGrd	0.068283
OverallCond	0.043303
LotArea	0.038824
Total_porch_sf	0.033870
CentralAir_Y	0.031832
LotFrontage	0.027526
Neighborhood_StoneBr	0.026581
OpenPorchSF	0.022713
MSSubClass_70	0.022189
Alley_Pave	0.021672
Neighborhood_Veenker	0.020098
BsmtQual_Ex	0.019949
KitchenQual_Ex	0.019787
HouseStyle_2.5Unf	0.018952
MasVnrType_Stone	0.018388
PavedDrive_P	0.017973
RoofMatl_WdShngl	0.017856
PavedDrive_Y	0.016840

Assignment Part-II

Lasso Regression

	Lasso Co-Efficient
Total_sqr_footage	0.202244
GarageArea	0.110863
TotRmsAbvGrd	0.063161
OverallCond	0.046686
LotArea	0.044597
CentralAir_Y	0.033294
Total_porch_sf	0.028923
Neighborhood_StoneBr	0.023370
Alley_Pave	0.020848
OpenPorchSF	0.020776
MSSubClass_70	0.018898
LandContour_HLS	0.017279
KitchenQual_Ex	0.016795
BsmtQual_Ex	0.016710
Condition1_Norm	0.015551
Neighborhood_Veenker	0.014707
MasVnrType_Stone	0.014389
PavedDrive_P	0.013578
LotFrontage	0.013377
PavedDrive_Y	0.012363

	Lasso Doubled Alpha Co-Efficient
Total_sqr_footage	0.204642
GarageArea	0.103822
TotRmsAbvGrd	0.064902
OverallCond	0.042168
CentralAir_Y	0.033113
Total_porch_sf	0.030659
LotArea	0.025909
BsmtQual_Ex	0.018128
Neighborhood_StoneBr	0.017152
Alley_Pave	0.016628
OpenPorchSF	0.016490
KitchenQual_Ex	0.016359
LandContour_HLS	0.014793
MSSubClass_70	0.014495
MasVnrType_Stone	0.013292
Condition1_Norm	0.012674
BsmtCond_TA	0.011677
SaleCondition_Partial	0.011236
LotConfig_CulDSac	0.008776
PavedDrive_Y	0.008685

Overall since the alpha values are small, we do not see a huge change in the model after doubling the 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?

- The optimal lambda values for Ridge and Lasso are determined as follows:
  - Ridge – 2
  - Lasso – 0.0001
- The Mean Squared Errors for Ridge and Lasso are as follows:
  - Ridge - 0.0018396090787924262
  - Lasso - 0.0018634152629407766

▪ Both models exhibit nearly identical Mean Squared Errors. Considering Lasso's capability for feature reduction, where the coefficient values of some features become zero, Lasso holds a distinct advantage over Ridge. As a result, Lasso is recommended as the preferred final model.

Question 3
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?

The five most crucial predictor variables in the existing Lasso model are:

- Total_sqr_footage
- GarageArea
- TotRmsAbvGrd
- OverallCond
- LotArea

After excluding these attributes from the dataset, we constructed a new Lasso model in the Jupyter notebook. The R2 of the updated model, without the top 5 predictors, decreases to 0.73. Simultaneously, the Mean Squared Error sees an increase to 0.0028575670906482538.

The top 5 features are

	Lasso Co-Efficient
LotFrontage	0.146535
Total_porch_sf	0.072445
HouseStyle_2.5Unf	0.062900
HouseStyle_2.5Fin	0.050487
Neighborhood_Veenker	0.042532

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?

Assignment Part-II

According to Occam's Razor, when confronted with two models demonstrating similar performance on finite training or test data, it is advisable to choose the one that makes fewer assumptions. This preference for simplicity is grounded in several key reasons:

- Simpler models tend to be more generic and widely applicable.
- They require fewer training samples for effective training, making them easier to train.
- Simpler models often exhibit greater robustness compared to complex ones.

Complex models, with low bias and high variance, can behave erratically with changes in the training dataset. On the other hand, simpler models, characterized by high bias and low variance, may make more errors in the training set but are less prone to overfitting and more adaptable to new data.

To strike a balance between simplicity and utility, regularization is employed. Regularization involves introducing a regularization term to the cost function, which penalizes the absolute values or squares of the model parameters. This process ensures that the model remains simple without becoming overly naive.

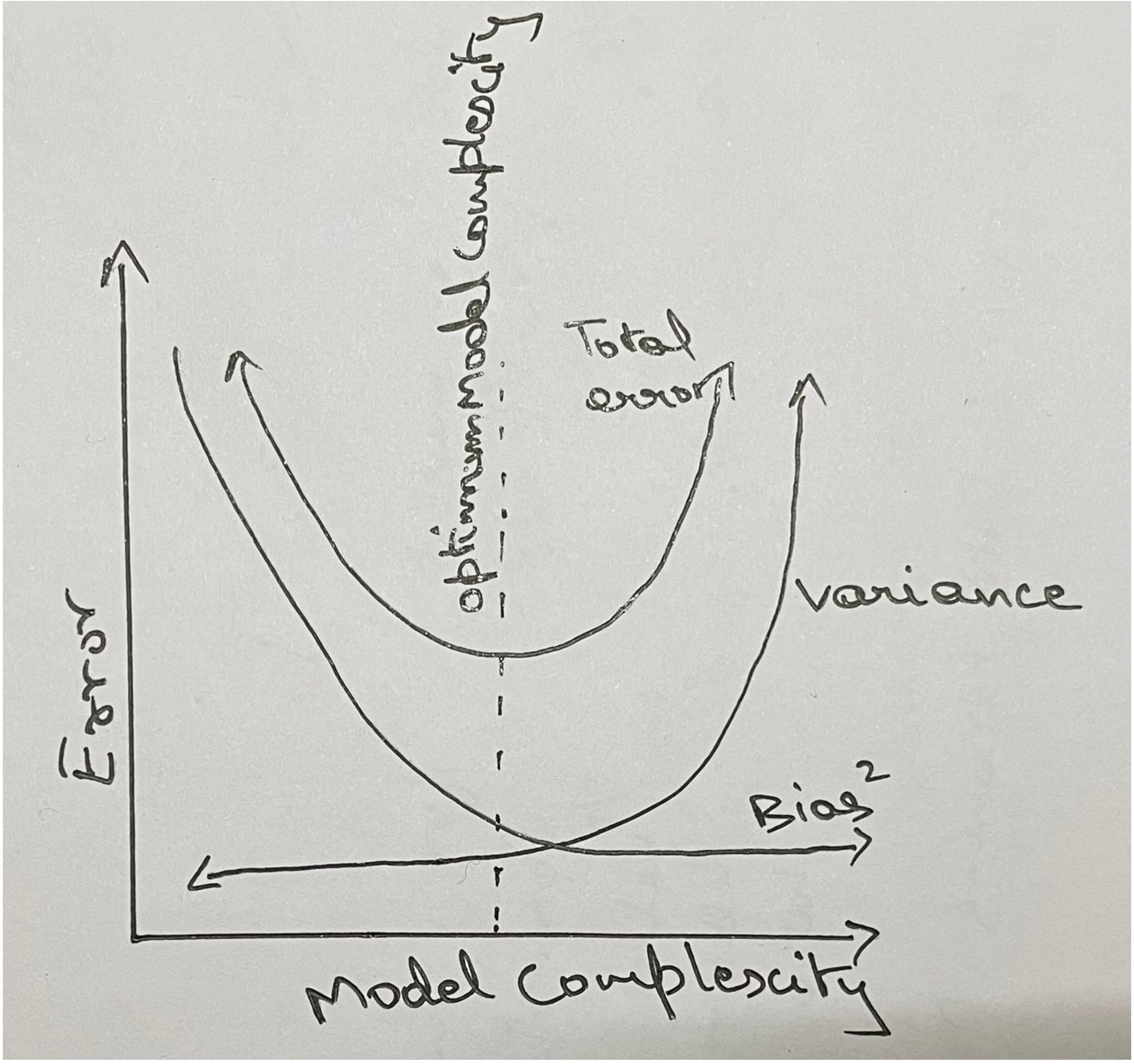
Furthermore, embracing model simplicity contributes to the Bias-Variance Trade-off:

- Complex models are highly sensitive to changes in the dataset, leading to instability.
- Simpler models, abstracting essential patterns from data, are less likely to undergo drastic changes with additions or removals of data points.

Bias quantifies the model's likely accuracy on test data, and a complex model can be accurate with sufficient training data. However, overly naive models exhibit a high bias, as they fail to discriminate among test inputs effectively.

Variance refers to the model's susceptibility to changes in the training data. Striking a balance between bias and variance is essential for maintaining model accuracy, as illustrated in the accompanying graph, which minimizes the total error. This balance ensures that the model is both accurate and adaptable across different datasets.

Assignment Part-II



Submitted by Vignesh V