ADVANCED REGRESSION ASSIGNMENT- SUBJECTIVE

Manjiri Gajmal (C57)

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 -

optimal_alpha_ridge = 2.0

optimal_alpha_lasso = 0.0003

For Ridge Regression Model (Original Model, alpha=2.0): *************************************** For Train Set: R2 score: 0.9168892364761739 MSE score: 0.08311076352382606 MAE score: 0.20895656282126895 RMSE score: 0.28828937462873316 For Test Set: R2 score: 0.8887449639711062 MSE score: 0.10785434533565937 MAE score: 0.2249407929854833 RMSE score: 0.3284118532204028 ***************************************	For Ridge Regression Model (Doubled alpha model, alpha=2*2=4.0): *************************************** For Train Set: R2 score: 0.9159352080712294 MSE score: 0.08406479192877059 MAE score: 0.20951676079559758 RMSE score: 0.2899392900742681 For Test Set: R2 score: 0.8890196847658441 MSE score: 0.10758802182776123 MAE score: 0.22471375184890868 RMSE score: 0.32800613077770546 ***************************************
For Lasso Regression Model (Original Model: alpha=0.0003): *************************************** For Train Set: R2 score: 0.9174421607935511 MSE score: 0.08255783920644885 MAE score: 0.20871547634234877 RMSE score: 0.28732879982077825 For Test Set: R2 score: 0.8877962610337671 MSE score: 0.10877404962839955 MAE score: 0.22613273845053983 RMSE score: 0.32980911089355847 ***************************************	For Lasso Regression Model: (Doubled alpha model: al pha:0.0003*2 = 0.0006) *************************************** For Train Set: R2 score: 0.9166619132505517 MSE score: 0.08333808674944836 MAE score: 0.20939123416427446 RMSE score: 0.2886833676356301 For Test Set: R2 score: 0.8880856587419308 MSE score: 0.10849349783074852 MAE score: 0.2261099797897605 RMSE score: 0.32938351177730274 ***************************************

After doubling the value for alpha, there is no major change in model; as alpha values are small.

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?

Ans-

Optimal values for lambda in case of Ridge and Lasso regression are as follows-

1) Ridge – 2.0
2) Lasso – 0.0003

Mean Squared Error (MSE) in case of Ridge and Lasso regression are as follows-

1) Ridge - 0.10785434533565937
2) Lasso - 0.10877404962839955

MSE in both regression is almost same
Lasso regularization promotes sparsity in the model by driving the coefficients of less significant features precisely to zero. This inherent sparsity characteristic renders Lasso particularly effective for automatic feature selection, as it tends to eliminate irrelevant features from the model. Hence, Lasso regression will be choice of regression for final model.

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?

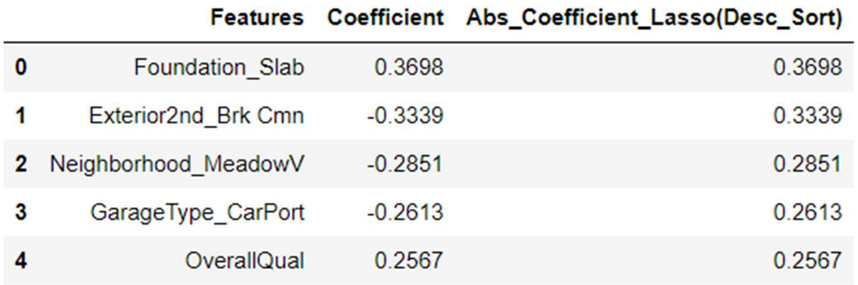
Ans-

In final lasso model following are top 5 features -

1) 'Exterior1st_BrkComm',
2) 'Foundation_Wood',
3) 'Neighborhood_IDOTRR',
4) 'MSZoning_FV',
5) 'Neighborhood_StoneBr'

After rebuilding Lasso model excluding above features are –

1) 'Foundation_Slab',
2) 'Exterior2nd_Brk Cmn',
3) 'Neighborhood_MeadowV',
4) 'GarageType_CarPort',
5) 'OverallQual'



	Features	Coefficient	Abs_Coefficient_Lasso(Desc_Sort)
0	Foundation_Slab	0.3698	0.3698
1	Exterior2nd_Brk Cmn	-0.3339	0.3339
2	Neighborhood_MeadowV	-0.2851	0.2851
3	GarageType_CarPort	-0.2613	0.2613
4	OverallQual	0.2567	0.2567

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?

To ensure that a model is robust and capable of generalizing well to new, unseen data, several strategies can be employed:

1. **Cross-Validation Techniques:**

Utilize methods like training-validation splits and cross-validation to assess how well the model performs on both the training data and unseen validation data. This helps gauge the model's generalization ability.

2. **Effective Feature Management:**

Carefully curate and engineer features, avoiding the inclusion of irrelevant or redundant ones. This aids in preventing overfitting and promotes a more generalized model.

3. **Regularization Methods:**

Apply regularization techniques, such as L1 (Lasso) and L2 (Ridge), to control the model's complexity and prevent it from memorizing noise in the training data, thus enhancing its ability to generalize.

4. **Hyperparameter Optimization:**

Tune hyperparameters systematically to find optimal values. This process ensures that the model adapts well to different datasets and does not rely excessively on specific parameter configurations.

5. **Ensemble Learning:**

Harness the power of ensemble methods, like bagging and boosting, to combine insights from multiple models. This often leads to improved generalization by reducing the impact of individual model biases.

6. **Diverse Evaluation Metrics:**

Evaluate the model's performance using a variety of metrics, offering a more nuanced understanding of its strengths and weaknesses. This comprehensive assessment helps ensure that the model's accuracy is not skewed by focusing on a single metric.

7. **Outlier Identification and Handling:**

Detect and appropriately handle outliers in the data, as they can distort the model's predictions and hinder its generalization capabilities.

8. **External Validation:**

Validate the model on entirely different datasets whenever possible. External validation provides additional assurance of the model's generalizability beyond the specific training data.

Implications for Accuracy:

Prioritize a balance between accuracy on the training set and the ability to generalize to new data.

Acknowledge that models with high accuracy on the training set but poor generalization may not perform well in real-world scenarios.

Understand that sacrificing a small amount of training accuracy can lead to significantly improved performance when faced with previously unseen data.