Assignment based subjective questions

Submitted by: Sumanth

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: Increasing the value of alpha in both Ridge and Lasso regression intensifies the regularization strength. This amplifies the penalty imposed on large coefficients, causing them to shrink further towards zero.

In Ridge Regression, doubling alpha strengthens the penalty on large coefficients, resulting in increased shrinkage towards zero compared to the original alpha value. This typically reduces model complexity, potentially enhancing its ability to generalize to unseen data.

Similarly, in Lasso Regression, doubling alpha heightens the penalty on large coefficients, leading to sparser solutions where more coefficients are driven precisely to zero. This can result in more aggressive feature selection, potentially reducing the number of predictor variables deemed important by the model.

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: The selection between Ridge and Lasso regression depends on the unique characteristics of your dataset and the objectives of your analysis. Here are several factors to consider:

1. Feature Importance: If you believe that only a subset of features significantly influences prediction and prefer a concise model that selects only those features, Lasso regression (with L1 regularization) is preferable. Conversely, if you believe that all features are pertinent but some may introduce noise or redundancy, Ridge regression (with L2 regularization) might be more suitable. Ridge regression reduces the coefficients of less important features towards zero without entirely eliminating them.

2. Model Interpretability: Lasso regression often yields sparse models by setting some coefficients precisely to zero. This feature enhances the interpretability of the model by highlighting the most influential features.

3. Computational Considerations: Lasso regression can be computationally intensive, especially with a large number of features or a sizable dataset. In contrast, Ridge regression typically has a simpler computational implementation.

4. Bias-Variance Tradeoff: Ridge regression generally handles multicollinearity more effectively than Lasso regression because it doesn't force coefficients to be exactly zero. Therefore, if multicollinearity is a concern, Ridge regression might provide better predictions with less variance.

5. Cross-Validation Performance: Ultimately, the decision may be based on the performance of each model on cross-validation or holdout datasets. It's essential to assess both Ridge and Lasso regression using cross-validation and choose the one that offers the best predictive performance.

Question 3.

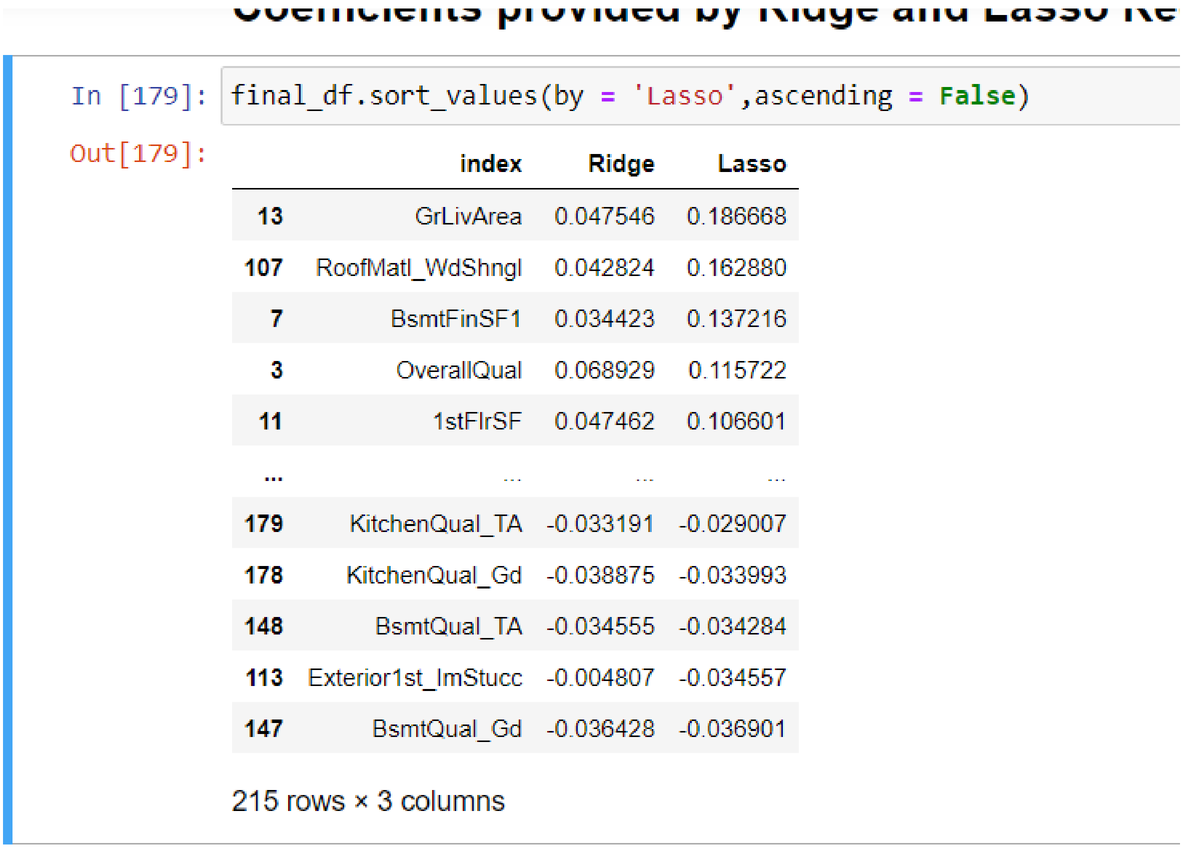
After building the model, you realised that the five most important predictor variables

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?

1st Iteration of Lasso before removing the five most important predictor variables



Coefficients provided by Ridge and Lasso Re

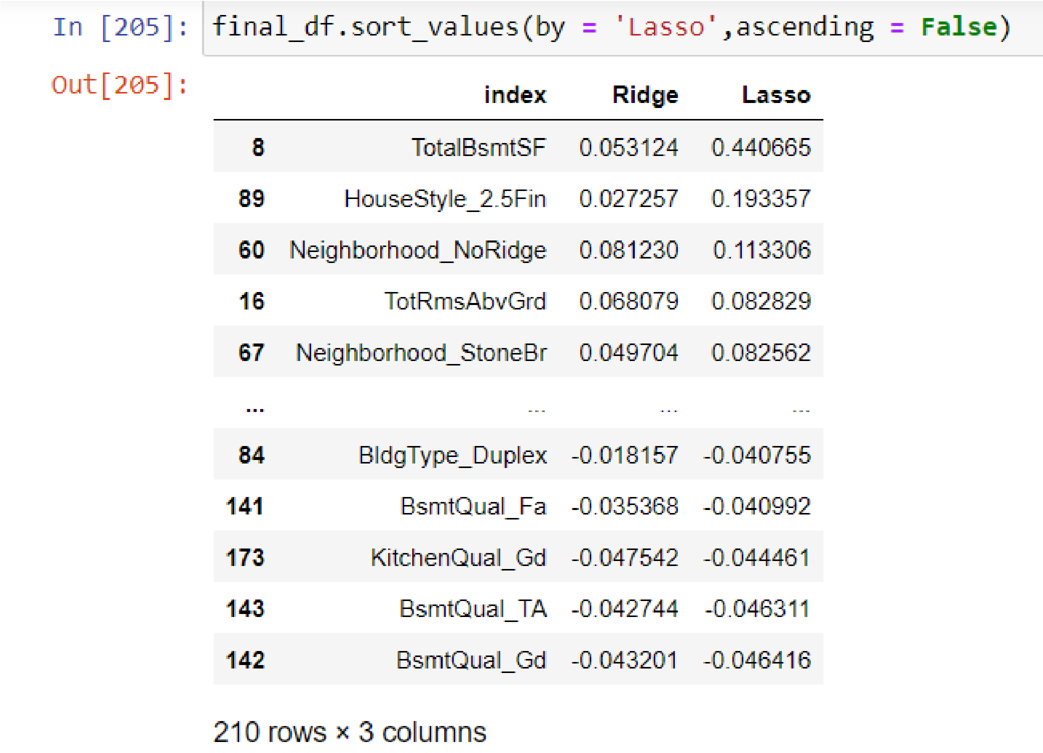
```
In [179]: final_df.sort_values(by = 'Lasso',ascending = False)
```

Out[179]:

	index	Ridge	Lasso
13	GrLivArea	0.047546	0.186668
107	RoofMatl_WdShngl	0.042824	0.162880
7	BsmtFinSF1	0.034423	0.137216
3	OverallQual	0.068929	0.115722
11	1stFlrSF	0.047462	0.106601
...	...	...	...
179	KitchenQual_TA	-0.033191	-0.029007
178	KitchenQual_Gd	-0.038875	-0.033993
148	BsmtQual_TA	-0.034555	-0.034284
113	Exterior1st_ImStucc	-0.004807	-0.034557
147	BsmtQual_Gd	-0.036428	-0.036901

215 rows × 3 columns

2nd iteration of lasso after before removing the five most important predictor variables



```
In [205]: final_df.sort_values(by = 'Lasso',ascending = False)
```

Out[205]:

	index	Ridge	Lasso
8	TotalBsmtSF	0.053124	0.440665
89	HouseStyle_2.5Fin	0.027257	0.193357
60	Neighborhood_NoRidge	0.081230	0.113306
16	TotRmsAbvGrd	0.068079	0.082829
67	Neighborhood_StoneBr	0.049704	0.082562
...	...	...	...
84	BldgType_Duplex	-0.018157	-0.040755
141	BsmtQual_Fa	-0.035368	-0.040992
173	KitchenQual_Gd	-0.047542	-0.044461
143	BsmtQual_TA	-0.042744	-0.046311
142	BsmtQual_Gd	-0.043201	-0.046416

210 rows × 3 columns

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 enhance the robustness and generalizability of a model, particularly when employing ridge and lasso regularization, consider implementing the following strategies:

1. Cross-validation: Employ techniques such as k-fold cross-validation to evaluate the model's performance across multiple subsets of the data. This aids in assessing how effectively the model generalizes to unseen data and mitigates the risk of overfitting.

2. Regularization Strength Tuning: Conduct a grid search or similar methods to fine-tune the regularization parameter (alpha) for ridge and lasso regression. This process assists in determining the optimal balance between bias and variance, resulting in a more resilient model.

3. Evaluation on Holdout Data: Set aside a portion of the dataset as a holdout set and assess the model's performance on this unseen data. This approach provides a more realistic estimation of the model's performance on entirely new data.

4. Feature Scaling: Ensure that all features are appropriately scaled, particularly when utilizing regularization techniques like ridge and lasso. This practice helps prevent features with larger scales from dominating the regularization process, leading to more balanced regularization effects.

5. Feature Selection: In lasso regularization, feature selection occurs naturally as some coefficients are driven to zero. However, it's crucial to validate the importance of these selected features through techniques such as cross-validation or domain knowledge validation. This ensures that the retained features are genuinely informative for the model.