Advanced Linear Regression

Part II – Subjective Questions

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?

If we double the optimal value of alpha in both Ridge and Lasso regression, the coefficients of the model will be penalized more heavily, and the model will become more biased i.e., the coefficients will be pushed closer to zero.

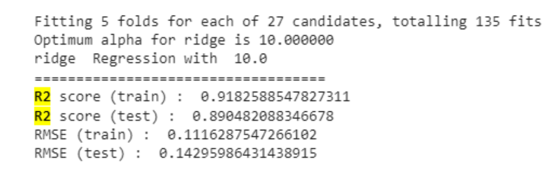
In Ridge regression, doubling the value of alpha will make the coefficients be pushed closer to zero, resulting in a simpler model with reduced variance.

In Lasso regression, doubling the value of alpha will make more coefficients to be set to exactly zero, resulting in a simpler model with reduced variance.

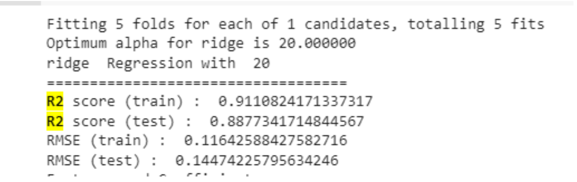
In both cases, it may also lead to increased bias, as the model may miss out on some relevant features of the data. Overall, doubling the value of alpha in Ridge and Lasso regression will increase the bias of the model and make it simpler, but it will also reduce its variance. The exact impact on the performance of the model will depend on the specific data and the chosen value of alpha.

The optimal alpha value is :

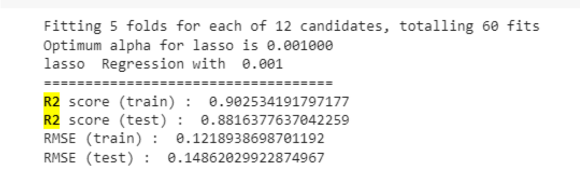
- Ridge = 10
- Lasso = 0.001



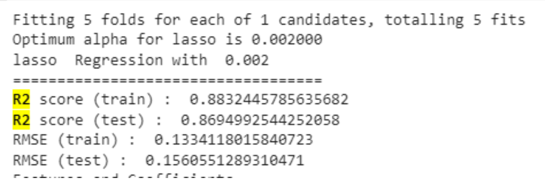
```
Fitting 5 folds for each of 27 candidates, totalling 135 fits
Optimum alpha for ridge is 10.000000
ridge  Regression with  10.0
===================================
R2 score (train) :  0.9182588547827311
R2 score (test) :  0.890482088346678
RMSE (train) :  0.1116287547266102
RMSE (test) :  0.14295986431438915
```



```
Fitting 5 folds for each of 1 candidates, totalling 5 fits
Optimum alpha for ridge is 20.000000
ridge  Regression with  20
===================================
R2 score (train) :  0.9110824171337317
R2 score (test) :  0.8877341714844567
RMSE (train) :  0.11642588427582716
RMSE (test) :  0.14474225795634246
```

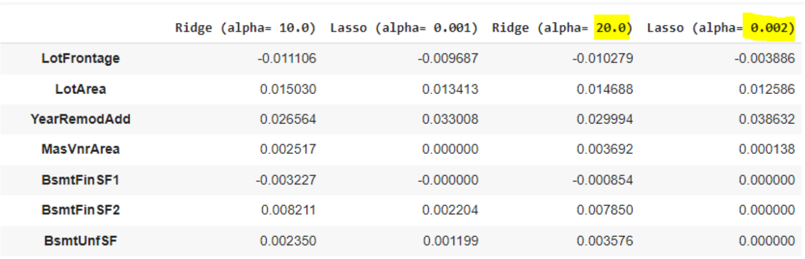


```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
Optimum alpha for lasso is 0.001000
lasso  Regression with  0.001
===================================
R2 score (train) :  0.902534191797177
R2 score (test) :  0.8816377637042259
RMSE (train) :  0.1218938698701192
RMSE (test) :  0.14862029922874967
```



```
Fitting 5 folds for each of 1 candidates, totalling 5 fits
Optimum alpha for lasso is 0.002000
lasso  Regression with  0.002
===================================
R2 score (train) :  0.8832445785635682
R2 score (test) :  0.8694992544252058
RMSE (train) :  0.1334118015840723
RMSE (test) :  0.1560551289310471
```

Co efficients for the optimal and double alpha values are -



	Ridge (alpha= 10.0)	Lasso (alpha= 0.001)	Ridge (alpha= 20.0)	Lasso (alpha= 0.002)
LotFrontage	-0.011106	-0.009687	-0.010279	-0.003886
LotArea	0.015030	0.013413	0.014688	0.012586
YearRemodAdd	0.026564	0.033008	0.029994	0.038632
MasVnrArea	0.002517	0.000000	0.003692	0.000138
BsmtFinSF1	-0.003227	-0.000000	-0.000854	0.000000
BsmtFinSF2	0.008211	0.002204	0.007850	0.000000
BsmtUnfSF	0.002350	0.001199	0.003576	0.000000

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 to and why?

The house dataset contains many correlated variables, Ridge regression may be a better choice. If the dataset contains many irrelevant or redundant features, Lasso regression may be more appropriate.

If there is multi collinearity in the dataset, lasso tends to make them 0 arbitrarily to resolve it.

If the primary goal is to identify variables impacting price, ridge might be better choice as Lasso might make a few coefficients 0.

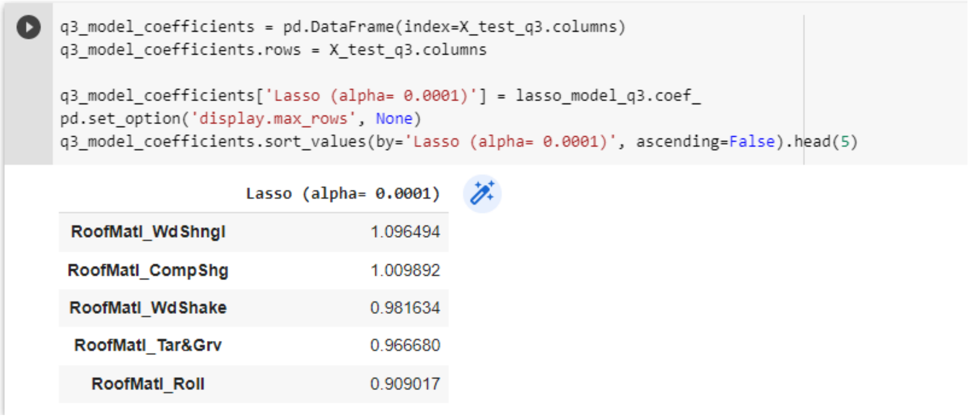
It is important to note that the choice between Ridge and Lasso depends on the specific goals of the analysis and the trade-off between bias and variance that is desired.

Here are some factors that can help in deciding between the two:

- If we expect that only a small number of predictors are important, Lasso regression is preferred because it can set the coefficients of irrelevant predictors to zero. In contrast, Ridge regression shrinks all the coefficients towards zero, but it does not set any of them exactly to zero.
- If we want to interpret the model coefficients, Ridge regression may be preferred because it does not eliminate any of the predictors entirely. In contrast, Lasso regression can eliminate predictors, making it harder to interpret the model.
- If there are highly correlated predictors in the data, Ridge regression is preferred because it will shrink the coefficients of all the predictors equally. In contrast, Lasso regression may arbitrarily select one of the correlated predictors and eliminate the others.

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?

The top 5 important predictors were removed, and model was run, it predicted the below variables in lasso model



```
q3_model_coefficients = pd.DataFrame(index=X_test_q3.columns)
q3_model_coefficients.rows = X_test_q3.columns

q3_model_coefficients['Lasso (alpha= 0.0001)'] = lasso_model_q3.coef_
pd.set_option('display.max_rows', None)
q3_model_coefficients.sort_values(by='Lasso (alpha= 0.0001)', ascending=False).head(5)
```

	Lasso (alpha= 0.0001)
RoofMatl_WdShngl	1.096494
RoofMatl_CompShg	1.009892
RoofMatl_WdShake	0.981634
RoofMatl_Tar&Grv	0.966680
RoofMatl_Roll	0.909017

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?

- A model is robust when any variation in the data does not affect its performance much.
- A generalizable model is able to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.
- To make sure a model is robust and generalizable, we have to take care it doesn't overfit. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data.
- In other words, the model should not be too complex in order to be robust and generalizable.
- If we look at it from the prespective of Accuracy, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.
- In general, we have to find strike some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge Regression and Lasso.