Assignment Part II Questions

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?

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?

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?

Question 4

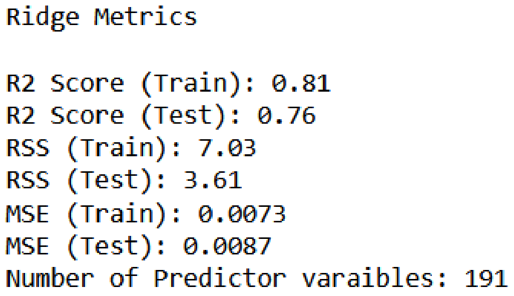
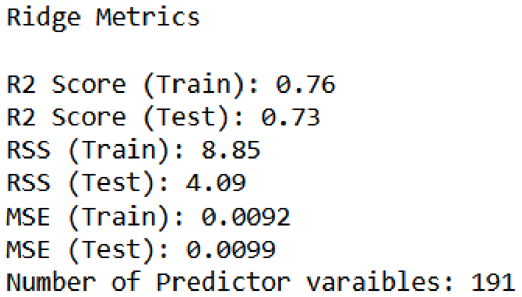
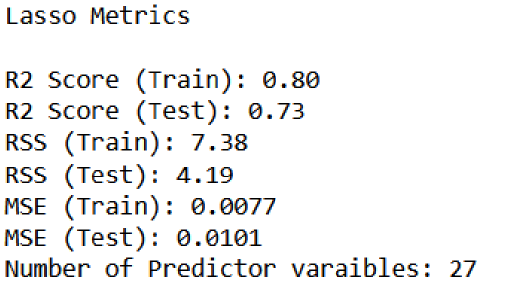
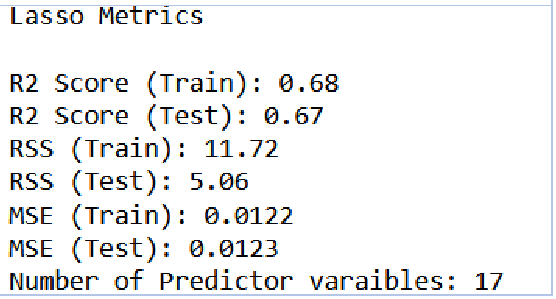
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

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?

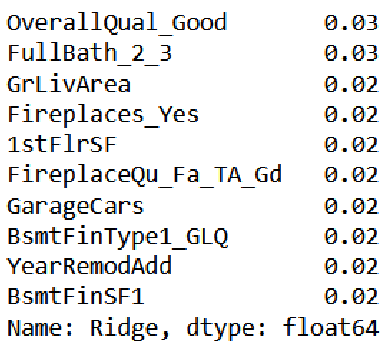
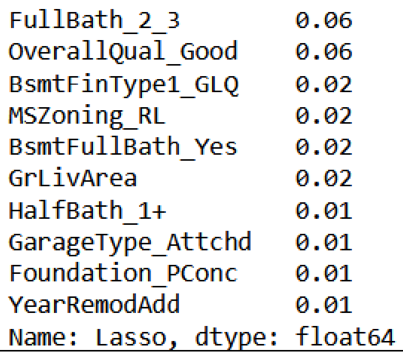
Answer:

At Optimal Value and Doubled Value:

	Optimal Ridge Alpha: 150 **Optimal Lasso Alpha: 0.004**	**Doubled Ridge Alpha: 300** **Doubled Lasso Alpha: 0.008**
Ridge	 Ridge Metrics R2 Score (Train): 0.81 R2 Score (Test): 0.76 RSS (Train): 7.03 RSS (Test): 3.61 MSE (Train): 0.0073 MSE (Test): 0.0087 Number of Predictor varaibles: 191	 Ridge Metrics R2 Score (Train): 0.76 R2 Score (Test): 0.73 RSS (Train): 8.85 RSS (Test): 4.09 MSE (Train): 0.0092 MSE (Test): 0.0099 Number of Predictor varaibles: 191
Lasso	 Lasso Metrics R2 Score (Train): 0.80 R2 Score (Test): 0.73 RSS (Train): 7.38 RSS (Test): 4.19 MSE (Train): 0.0077 MSE (Test): 0.0101 Number of Predictor varaibles: 27	 Lasso Metrics R2 Score (Train): 0.68 R2 Score (Test): 0.67 RSS (Train): 11.72 RSS (Test): 5.06 MSE (Train): 0.0122 MSE (Test): 0.0123 Number of Predictor varaibles: 17

Inference: The R2 score on the test data decreases on doubling the Alpha. And the RSS, MSE are also increased.

Top 10 Predictors after Doubled Alpha Value:

Top 10 Ridge predictors and their coef	Top 10 Lasso predictors and their coef
 OverallQual_Good 0.03 FullBath_2_3 0.03 GrLivArea 0.02 Fireplaces_Yes 0.02 1stFlrSF 0.02 FireplaceQu_Fa_TA_Gd 0.02 GarageCars 0.02 BsmtFinType1_GLQ 0.02 YearRemodAdd 0.02 BsmtFinSF1 0.02 Name: Ridge, dtype: float64	 FullBath_2_3 0.06 OverallQual_Good 0.06 BsmtFinType1_GLQ 0.02 MSZoning_RL 0.02 BsmtFullBath_Yes 0.02 GrLivArea 0.02 HalfBath_1+ 0.01 GarageType_Attchd 0.01 Foundation_PConc 0.01 YearRemodAdd 0.01 Name: Lasso, dtype: float64

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?

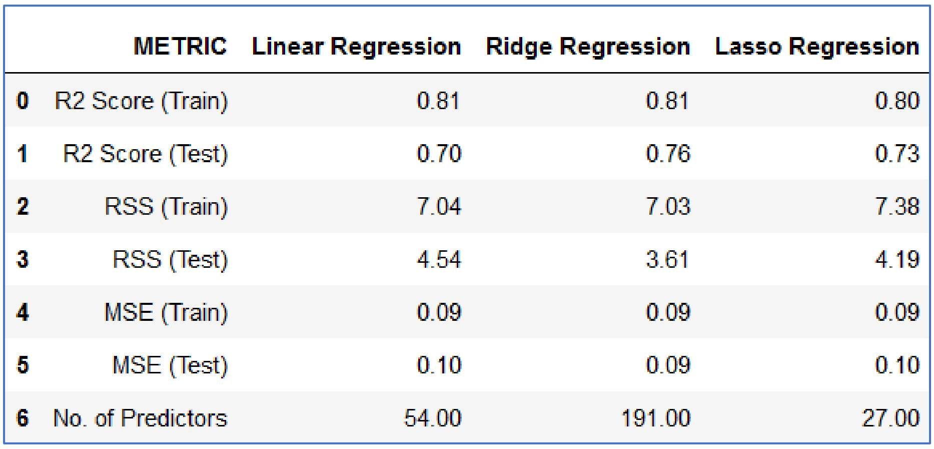
Answer:

The optimal lambda/alpha value will be used, otherwise the model will end up underfitting or will not handle overfitting. Between Lasso and Ridge, **Lasso** will be chosen to apply.

Reason for choosing and applying Lasso:

1. The Number of Predictor variables used by Lasso is significantly lesser than Ridge
2. The R2, RSS & MSE are relatively comparable with the Linear and Ridge with lot lesser Predictor variables
3. With Lasso the Interpretability of the Model increases due to a smaller number of Predictors

Metrics comparison of Linear, Ridge and Lasso:



	METRIC	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.81	0.81	0.80
1	R2 Score (Test)	0.70	0.76	0.73
2	RSS (Train)	7.04	7.03	7.38
3	RSS (Test)	4.54	3.61	4.19
4	MSE (Train)	0.09	0.09	0.09
5	MSE (Test)	0.10	0.09	0.10
6	No. of Predictors	54.00	191.00	27.00

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?

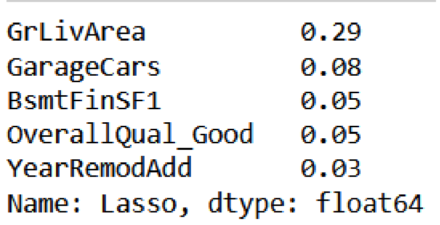
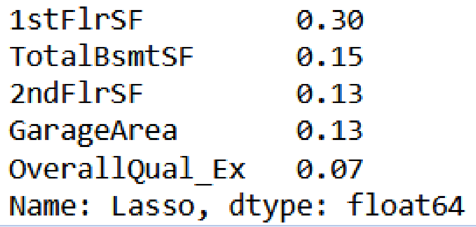
Answer:

Before:

Top 5 predictor variables.	Data Description
GrLivArea	Above grade (ground) living area square feet
GarageCars	Size of garage in car capacity
BsmtFinSF1	Type 1 Basement finished square feet
OverallQual_Good	Good overall material and finish of the house
YearRemodAdd	Construction date or the Remodel date

After:

Top 5 predictor variables.	Data Description
1stFlrSF	First Floor square feet
TotalBsmtSF	Total square feet of basement area
2ndFlrSF	Second floor square feet
GarageArea	Size of garage in square feet
OverallQual_Ex	Excellent overall material and finish of the house

Predictor Variables Before & their coef	**After dropping the 5 most important predictors dropped & tuning of Alpha**
 GrLivArea 0.29 GarageCars 0.08 BsmtFinSF1 0.05 OverallQual_Good 0.05 YearRemodAdd 0.03 Name: Lasso, dtype: float64	 1stFlrSF 0.30 TotalBsmtSF 0.15 2ndFlrSF 0.13 GarageArea 0.13 OverallQual_Ex 0.07 Name: Lasso, dtype: float64

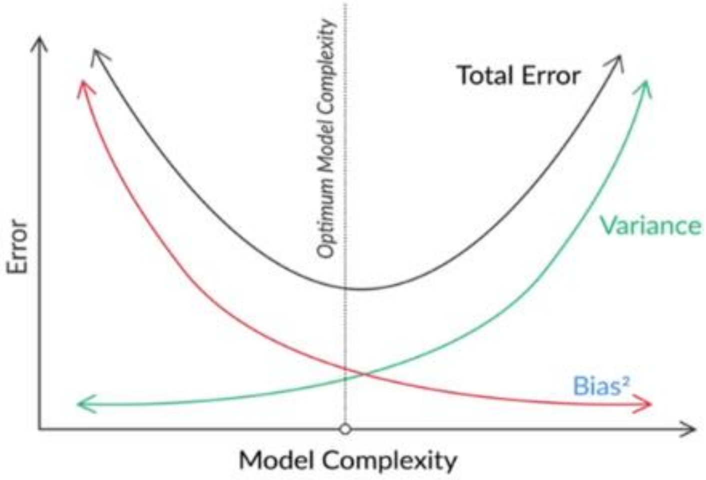
Question 4

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:

Optimal value of lambda or Alpha will help in choosing the optimal complexity thus making the model robust and generalisable.

- If the value of lambda or Alpha is High, it will lead to underfitting
- If the value of lambda or Alpha is Low, it will not handle the overfitting


Optimum Model Complexity
Total Error
Variance
Error
Bias²
Model Complexity