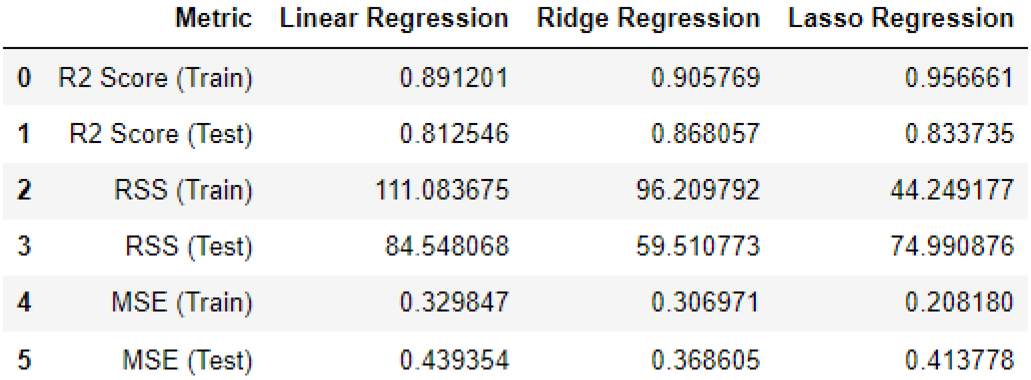
Assignment Part-II Solution

Ishan Anant Karve (ishan.karve@gmail.com)

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?

Finding: The optimal value for alpha for ridge and lasso regression are observed to be 500 for ridge and 0.001 for lasso regression. The comparative model metrics generated are

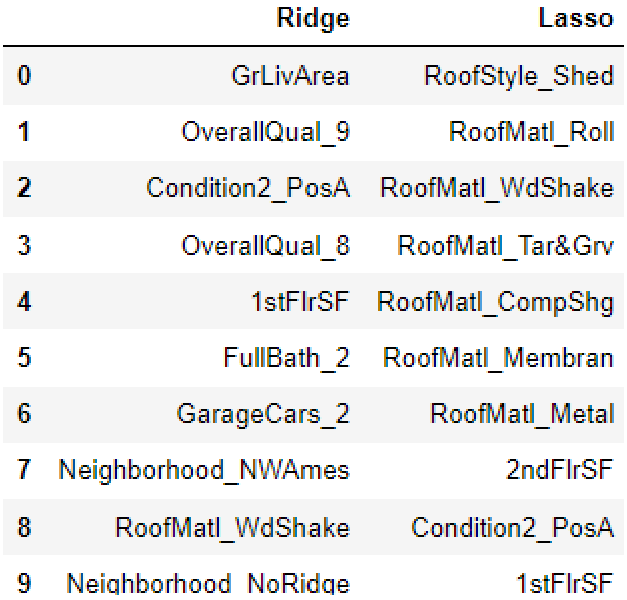
:



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.891201	0.905769	0.956661
1	R2 Score (Test)	0.812546	0.868057	0.833735
2	RSS (Train)	111.083675	96.209792	44.249177
3	RSS (Test)	84.548068	59.510773	74.990876
4	MSE (Train)	0.329847	0.306971	0.208180
5	MSE (Test)	0.439354	0.368605	0.413778

Table 1: Comparison of Regression Metrics

The top features predicted by the model (both positive and negative coefficients) are

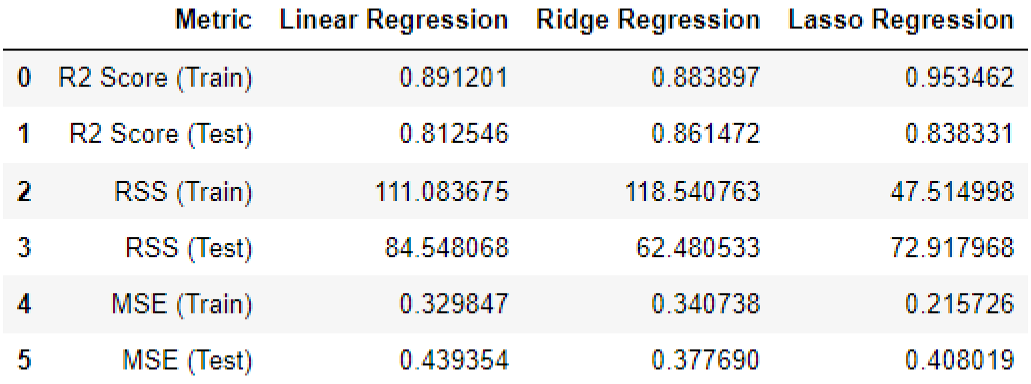


	Ridge	Lasso
0	GrLivArea	RoofStyle_Shed
1	OverallQual_9	RoofMatl_Roll
2	Condition2_PosA	RoofMatl_WdShake
3	OverallQual_8	RoofMatl_Tar&Grv
4	1stFlrSF	RoofMatl_CompShg
5	FullBath_2	RoofMatl_Membran
6	GarageCars_2	RoofMatl_Metal
7	Neighborhood_NWAmes	2ndFlrSF
8	RoofMatl_WdShake	Condition2_PosA
9	Neighborhood_NoRidge	1stFlrSF

Table 2: Top features for Ridge and Lasso Regression

Post Doubling Alpha

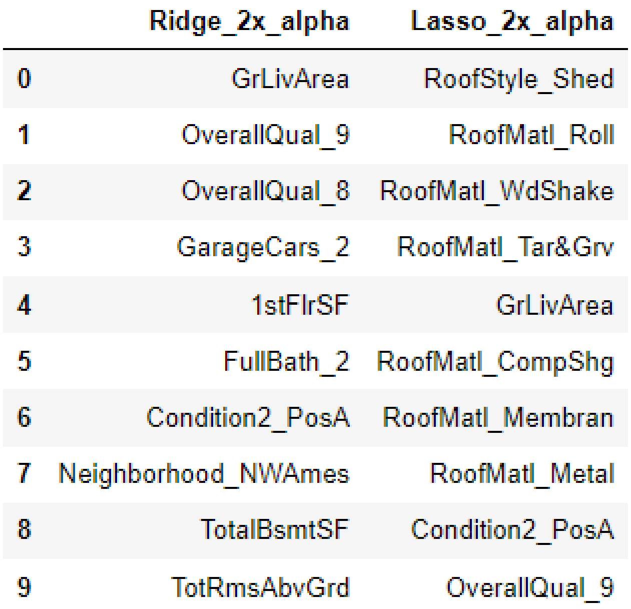
New Metrics are (linear unchanged)



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.891201	0.883897	0.953462
1	R2 Score (Test)	0.812546	0.861472	0.838331
2	RSS (Train)	111.083675	118.540763	47.514998
3	RSS (Test)	84.548068	62.480533	72.917968
4	MSE (Train)	0.329847	0.340738	0.215726
5	MSE (Test)	0.439354	0.377690	0.408019

Table 3: New regression metrics with doubled alpha value

New Parameters Are



	Ridge_2x_alpha	Lasso_2x_alpha
0	GrLivArea	RoofStyle_Shed
1	OverallQual_9	RoofMatl_Roll
2	OverallQual_8	RoofMatl_WdShake
3	GarageCars_2	RoofMatl_Tar&Grv
4	1stFlrSF	GrLivArea
5	FullBath_2	RoofMatl_CompShg
6	Condition2_PosA	RoofMatl_Membran
7	Neighborhood_NWAmes	RoofMatl_Metal
8	TotalBsmtSF	Condition2_PosA
9	TotRmsAbvGrd	OverallQual_9

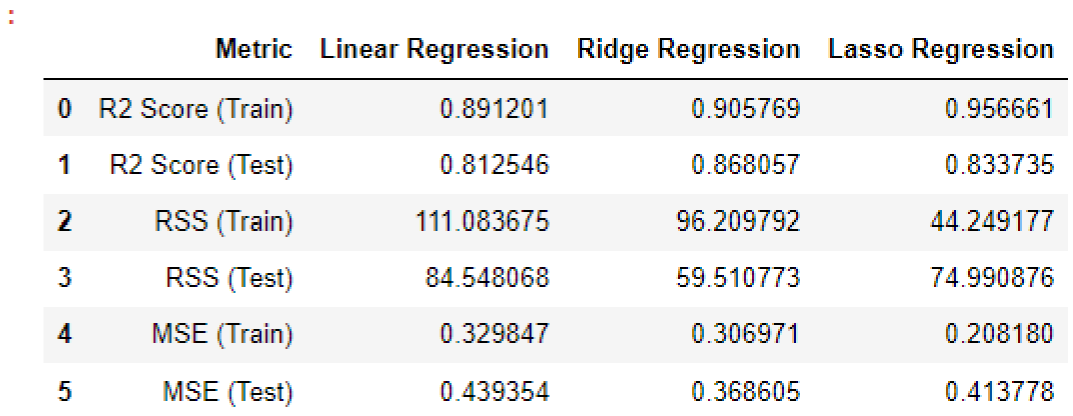
Table 4: Top features with 2x alpha

Observation: *Doubling of alpha for both ridge and lasso decreases the model accuracy, but not by a significant margin. However, significant change is observed in parameters and their relative rankings indicating that the intra-significance also changes. The most important predictor variables (lasso) are RoofStyle_Shed, RoofMatl_Roll, RoofMatl_WdShake, RoofMatl_Tar&Grv, GrLivArea, RoofMatl_CompShg, RoofMatl_Membran, RoofMatl_Metal, Condition2_PosA and OverallQual_9*

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?

Finding: The final data set on which the regression exercise was undertaken comprised of 286 features of which the lasso regression pushed 171 to zero leaving only 115 features for modelling.

The ridge regression on the other hand worked on 286 features like with lasso, however, it could push only 20 features towards zero, leaving 266 variables for modelling.



:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.891201	0.905769	0.956661
1	R2 Score (Test)	0.812546	0.868057	0.833735
2	RSS (Train)	111.083675	96.209792	44.249177
3	RSS (Test)	84.548068	59.510773	74.990876
4	MSE (Train)	0.329847	0.306971	0.208180
5	MSE (Test)	0.439354	0.368605	0.413778

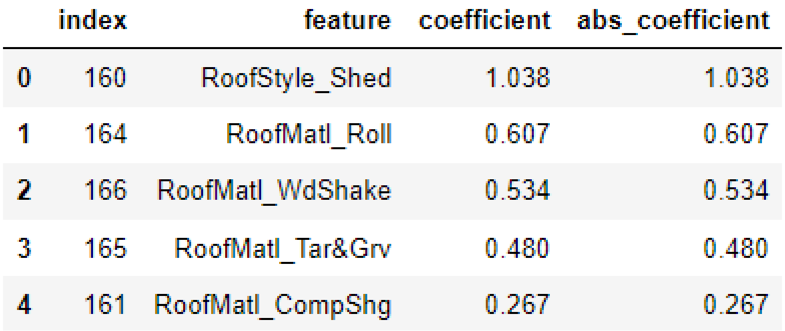
Table 5: Comparative regression metrics

A comparative assessment of the regression metrics, indicates that the Lasso Regression has better predictive power over Ridge and Std Linear Regression (OLS).

Therefore based on the findings at hand, in the extant example, it appears that usage of Lasso regression will yield better results.

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?

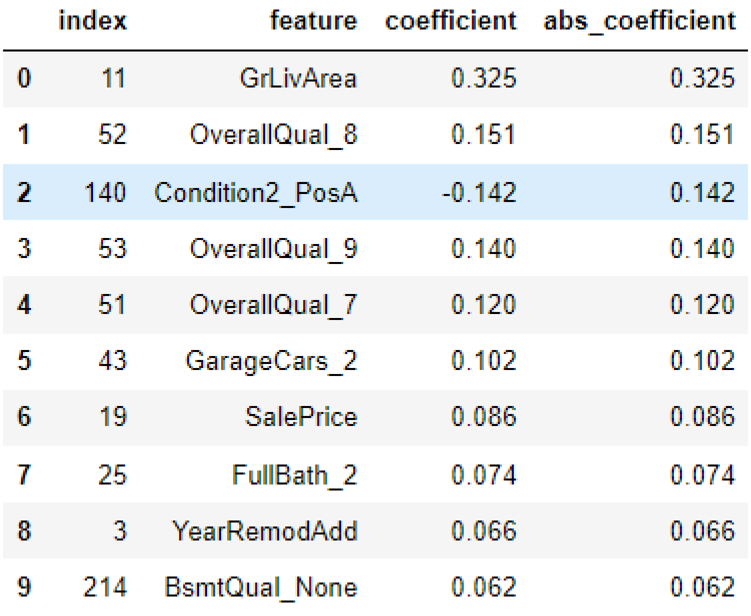
Finding: The five most import predictors deduced in the lasso model are



	index	feature	coefficient	abs_coefficient
0	160	RoofStyle_Shed	1.038	1.038
1	164	RoofMatl_Roll	0.607	0.607
2	166	RoofMatl_WdShake	0.534	0.534
3	165	RoofMatl_Tar&Grv	0.480	0.480
4	161	RoofMatl_CompShg	0.267	0.267

Table 6: Important features for Lasso regression

Post dropping these important predictors, a new model was generated, and the new set of important predictors (five) are *GrLivArea*, *OverallQual_8*, *Condition2_PosA*, *OverallQual_9*, *OverallQual_7*, *GarageCars_2*, *SalePrice*, *FullBath_2*,*YearRemodAdd*, *BsmtQual_None*. Their predictive power is



	index	feature	coefficient	abs_coefficient
0	11	GrLivArea	0.325	0.325
1	52	OverallQual_8	0.151	0.151
2	140	Condition2_PosA	-0.142	0.142
3	53	OverallQual_9	0.140	0.140
4	51	OverallQual_7	0.120	0.120
5	43	GarageCars_2	0.102	0.102
6	19	SalePrice	0.086	0.086
7	25	FullBath_2	0.074	0.074
8	3	YearRemodAdd	0.066	0.066
9	214	BsmtQual_None	0.062	0.062

Table 7: New top features

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?

Finding: A model is considered to be robust and generalisable if it is able to handle data on which it was not trained on (unseen data). At the extremes, both, over-fitting and an under-fitting model will not be a good candidate for generalisation. Therefore, based on the domain knowledge and known constraints a model can be generalised by finding the optimal error threshold limits (up and down). Using K-fold methods we can try to estimate a models efficacy to adapt to unseen data, thereby making it more robust and generalisable. Another common method to generalise model is to use data regularisation. This method penalises complex model thereby making it more general.

Robustness of a model is the ability of the model to be predictable even if its basic assumptions on which its based are altered. Robustness of a model can be increased by better outlier treatment. Also a model can be made robust by making the variable more normalised.

Since robustness and ability of a model to be generalised are a trade-off between over-fitted model and an under-fitted model, there is always an accuracy trade-off. The more generalisable model the less accurate it is, same is the phenomenon with robust. Usually a robust model is less accurate than a non-robust model. This is because in the effort of making the model generalisable and robust, salting, data manipulations and normalisations are undertaken which may affect its accuracy.