Advanced Linear Regression Assignment - House Pricing

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: The optimal value of alpha for ridge regression is 7.0 and that for lasso regression is 0.0001.
In general, as alpha increases, the magnitude of coefficients decreases.
For Ridge regression, once the alpha is doubled that is 14.0, we see that the R-squared for both training and test has dropped by approx.1.5%. Further, we see the magnitude of coefficients of the top 10 variables have dropped and the importance/ranking of these variables have changed among them. However, the top 10 variables remain same albeit in different order.

Ridge with alpha = 7.0

	Feature	Coefficient
0	Constant	0.1140
110	OverallQual_VExc	0.0710
9	2ndFlrSF	0.0685
11	GrLivArea	0.0624
75	Neighborhood_NoRidge	0.0592
106	OverallQual_Exc	0.0591
8	1stFlrSF	0.0433
14	FullBath	0.0431
20	GarageCars	0.0409
18	TotRmsAbvGrd	0.0378
171	BsmtQual_Gd	-0.0349

Ridge with alpha = 14

	Feature	Coefficient
0	Constant	0.1377
110	OverallQual_VExc	0.0548
75	Neighborhood_NoRidge	0.0524
9	2ndFlrSF	0.0513
11	GrLivArea	0.0492
106	OverallQual_Exc	0.0478
14	FullBath	0.0392
18	TotRmsAbvGrd	0.0364
8	1stFlrSF	0.0354
20	GarageCars	0.0352
171	BsmtQual_Gd	-0.0318

For Lasso regression, when the alpha is doubled that is 0.0002, we see that the R-squared for both training and test has dropped by approx.1.0%. Further, we see the magnitude of coefficients of the top 10 variables have changed slightly and the importance/ranking of these variables have changed among them. We can see one new feature (FullBath) replacing the older feature (LotArea) in the top 10 features.

Lasso with alpha = 0.0001

	Feature	Coefficient
11	GrLivArea	0.2584
110	OverallQual_VExc	0.1218
106	OverallQual_Exc	0.0981
0	Constant	0.0774
75	Neighborhood_NoRidge	0.0643
9	2ndFlrSF	0.0623
20	GarageCars	0.0562
2	LotArea	0.0482
111	OverallQual_VGood	0.0439
76	Neighborhood_NridgHt	0.0379
66	Neighborhood_Crawfor	0.0377

Lasso with alpha = 0.0002

	Feature	Coefficient
11	GrLivArea	0.2594
110	OverallQual_VExc	0.1253
106	OverallQual_Exc	0.1038
0	Constant	0.0838
75	Neighborhood_NoRidge	0.0634
20	GarageCars	0.0575
111	OverallQual_VGood	0.0464
9	2ndFlrSF	0.0429
66	Neighborhood_Crawfor	0.0344
76	Neighborhood_NridgHt	0.0337
14	FullBath	0.0323

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: We choose lasso with lambda of 0.0001 over ridge with lambda of 7.0, as the RSS and RMSE for lasso regression is less than ridge regression on both train and test sets. Also, lasso regression pushes the near-zero coefficients to zero thus performing feature selection and aiding to build a simpler yet robust model.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score Train	0.859664	0.887003	0.892869
1	R2 Score Test	0.849267	0.850301	0.854468
2	RSS Train	1.726840	1.390430	1.318246
3	MSE Train	0.001691	0.001362	0.001291
4	RMSE Train	0.041126	0.036903	0.035932
5	RSS Test	1.290637	1.281781	1.246099
6	MSE Test	0.002947	0.002926	0.002845
7	RMSE Test	0.054283	0.054097	0.053338

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: The new top 5 features are 1stFlrSF, MasVnrArea, TotRmsAbvGrd, LotArea and GarageCars, apart from the constant term. Below is the screenshot from the code file:

	Feature	Coefficient
8	1stFlrSF	0.1829
0	Constant	0.1510
3	MasVnrArea	0.0821
16	TotRmsAbvGrd	0.0791
2	LotArea	0.0724
18	GarageCars	0.0687

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: In a linear regression model, if we have more features in the dataset, the model complexity increases, which can result in increase in variance and decrease in bias thus overfitting. To overcome this, regularization techniques are used, so that the model can be made more generalizable.
Regularization techniques decrease the model complexity by reducing the magnitude of the coefficients, it does this by adding a penalty/shrinkage term with the cost function of the model. Thus, it reduces the complexity and the variance of the model at the cost of adding a small amount of bias to it.

For Ridge Regression, the penalty term is given by:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2$$

$$= RSS + \lambda \sum_{j=1}^{p}\beta_j^2$$

For Lasso Regression, the penalty term is given by:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

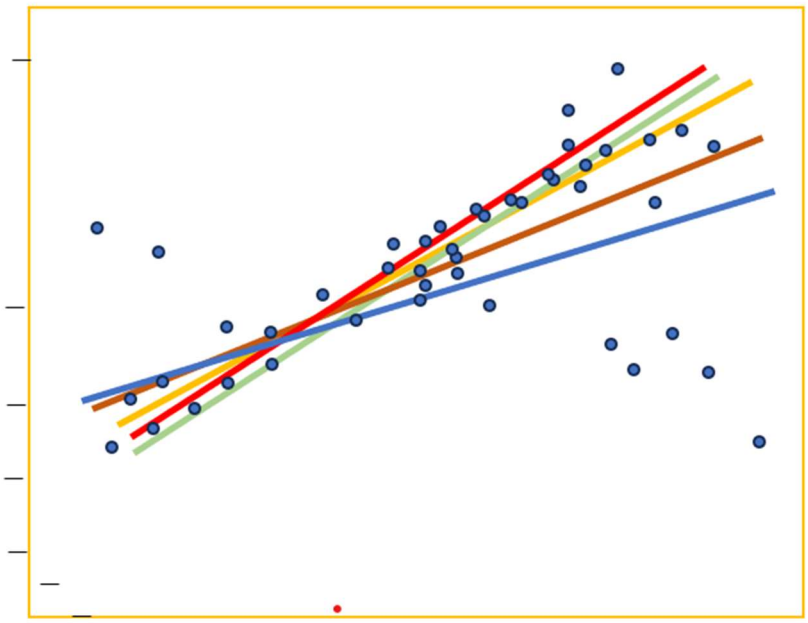
$$= RSS + \lambda \sum_{j=1}^{p}|\beta_j|$$

As there is variance-bias trade-off, we should choose the hyperparameter lambda wisely such that it gives the lowest error.

Also, a robust and generalisable model, would be able to handle the data with skewness and/or outliers in the input or target variables. Hence to ensure that we should train the model with outliers and access the mean absolute error for its performance.

The model building would involve a hyperparameter to handle the effect of outliers. Such model would give a best-fit line that follow the main body of the datapoints (i.e. where the datapoints are concentrated instead of getting affected by the outliers).
Example – In Huber Regression, the smaller values of hyperparameter (epsilon) would consider more of the data outliers, and in turn, make the model more robust to outliers.

Below figure is a simple depiction of best-line fit for various models with respect to the input data points. We see the line-fit in blue is highly affected due to presence of outliers in the dataset. While, the line in orange is much better fit as it follows the main body of the datapoints.



The more robust and generalisable model is the better it is in terms of accuracy on the test dataset.