Subjective Questions

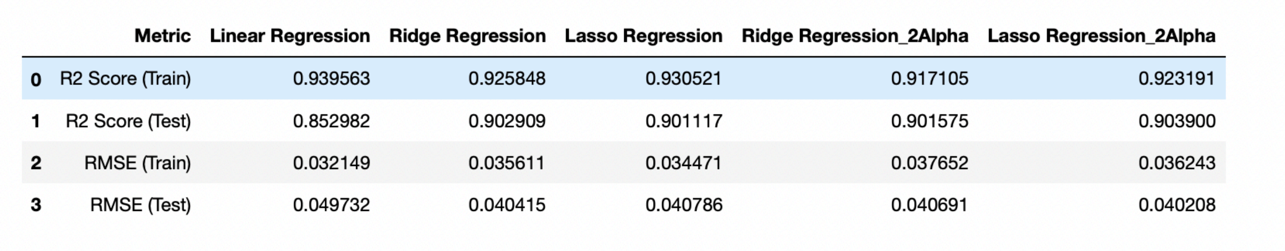
Question 1

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?

Answer:

The optimal value of alpha for ridge regression is 5.0 and the optimal value of alpha for lasso regression is 0.0001

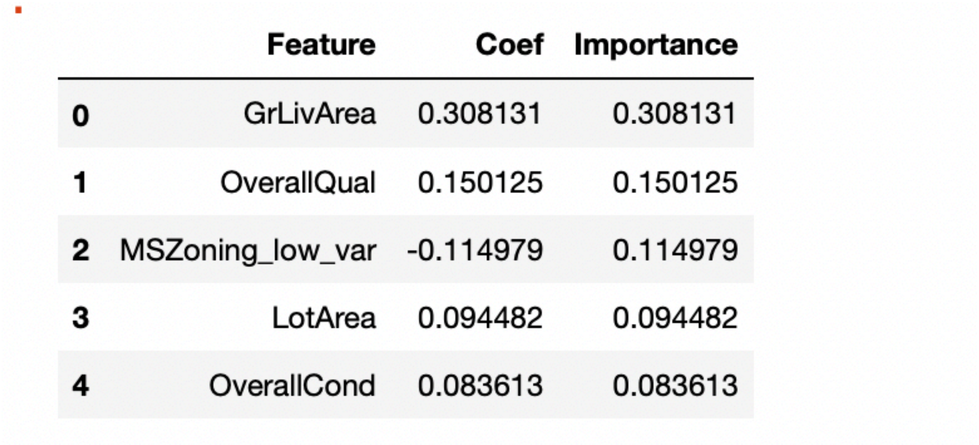
Doubling the value of alpha doesn’t have much of an effect on our model performance as we can see that the R-squared as well as the RMSE value are almost the same. Also, the top 5 features for predicting house sale price also remain the same even though there is some change in the model coefficients.



	Metric	Linear Regression	Ridge Regression	Lasso Regression	Ridge Regression_2Alpha	Lasso Regression_2Alpha
0	R2 Score (Train)	0.939563	0.925848	0.930521	0.917105	0.923191
1	R2 Score (Test)	0.852982	0.902909	0.901117	0.901575	0.903900
2	RMSE (Train)	0.032149	0.035611	0.034471	0.037652	0.036243
3	RMSE (Test)	0.049732	0.040415	0.040786	0.040691	0.040208

However, the lasso model that we build after doubling the alpha values uses less number of variables in the model compared to the earlier model. After doubling the alpha value, the lasso regularisation selects 93 features for predicting the house sale prices. On the other hand, earlier lasso regularisation was using 117 features for predictions.

The most important predictor variables after implementing the change are:



	Feature	Coef	Importance
0	GrLivArea	0.308131	0.308131
1	OverallQual	0.150125	0.150125
2	MSZoning_low_var	-0.114979	0.114979
3	LotArea	0.094482	0.094482
4	OverallCond	0.083613	0.083613

As the above grade living area or Lot size increases, the sale price of the house also increases. Also, the sale price increases as the overall quality and overall condition of the house improves.
On the other hand, if the zoning classification of the sale is Commercial, the price of the house decreases.

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:

Since the model performance for both ridge and lasso regression is similar with similar r-squared values on train and test data as well as similar RMSE values for train and test data, I will choose to apply lasso regression on this data to predict house prices.

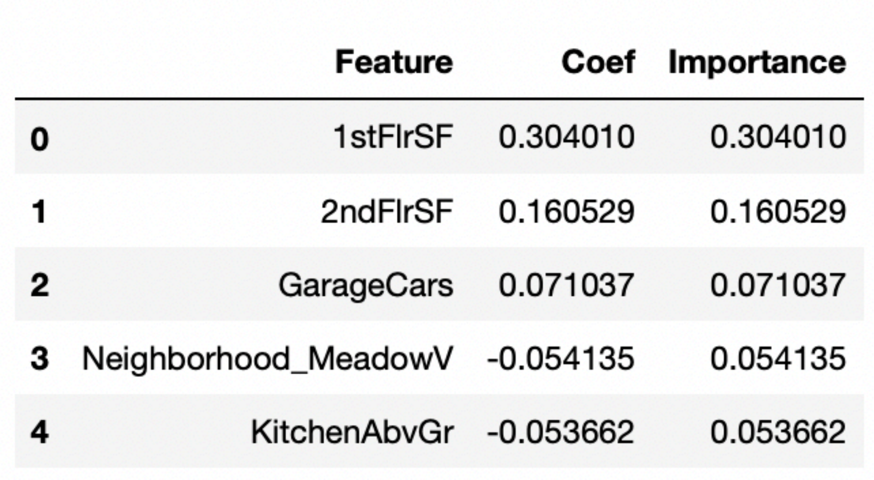
Reason for choosing Lasso regression is because it helps us in feature selection by shrinking coefficients of less important features to zero without compromising on the model performance. We will be able to predict the house prices with similar accuracy and a simpler model compared to ridge regression.

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 five most important predictor variables in the new model are:



	Feature	Coef	Importance
0	1stFlrSF	0.304010	0.304010
1	2ndFlrSF	0.160529	0.160529
2	GarageCars	0.071037	0.071037
3	Neighborhood_MeadowV	-0.054135	0.054135
4	KitchenAbvGr	-0.053662	0.053662

As the first floor or second floor square feet increases, or the size of garage in car capacity increases, sale price of the house increases.
On the other hand, if the house is located in Meadow Village or the number of kitchens above ground increases, the sale price of the house decreases.

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:

A model is robust and generalisable if it is able to perform reasonably well on unseen but similar data compared to the data it is trained on. This basically means that the model should have a test error rate similar to train error rate and not overfit. The model is able to strike a balance between the bias-variance tradeoff.

We can make sure that a model doesn’t overfit by using regularisation. Regularisation adds a penalty term to the loss function. This in turn penalises a model more as its complexity increases. We can use either ridge or lasso regularisation for making our models robust and generalisable.

Ridge regression adds “squared magnitude” of coefficient as penalty term to the loss function

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

Lasso regression adds “absolute value of magnitude” of coefficient as penalty term to the loss function.

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

The implications for making a model robust and generalisable is that it will be complex enough to capture the underlying pattern in the training data but not so complex that it captures all the noise in the train data too. It will have low enough variance such that it performs well on unseen data as well.