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:

Based on 5 folds CV the optimum parameters for alpha is calculated as follows

Ridge : 0.14532

Lasso : 0.0000099

These hyperparameter values were arrived at by performing coarse and fine tuning.

Changes in model performance due to doubling the optimum Alpha are as follows

Ridge (optimum alpha)		
Metric	Train	Test
r2	0.867147	0.886267
mse	0.001608	0.00141

Lasso (optimum alpha)		
Metric	Train	Test
r2	0.880232	0.889237
mse	0.00145	0.001373

Ridge (twice alpha)		
Metric	Train	Test
r2	0.859186	0.883935
mse	0.001704	0.001439

Lasso (twice alpha)		
Metric	Train	Test
r2	0.877942	0.888612
mse	0.001477	0.001381

Changes in variable importance due to doubling the optimum Alpha are as follows

Ridge (optimum alpha)	Ridge	Ridge_abs
RoofMatl_WdShngl	0.4521	0.4521
RoofMatl_CompShg	0.3652	0.3652
RoofMatl_Membran	0.3537	0.3537
RoofMatl_Tar&Grv	0.3482	0.3482
RoofMatl_WdShake	0.3410	0.3410

Lasso (optimum alpha)	Lasso	Lasso_abs
RoofMatl_WdShngl	0.8397	0.8397
RoofMatl_Membran	0.7873	0.7873
RoofMatl_CompShg	0.7541	0.7541
RoofMatl_Metal	0.7521	0.7521
RoofMatl_Tar&Grv	0.7423	0.7423

Ridge (twice alpha)	Ridge	Ridge_abs
RoofMatl_WdShngl	0.3222	0.3222
1stFlrSF	0.2681	0.2681
RoofMatl_CompShg	0.2358	0.2358
RoofMatl_Tar&Grv	0.2177	0.2177
RoofMatl_Membran	0.2124	0.2124

Lasso (twice alpha)	Lasso	Lasso_abs
RoofMatl_WdShngl	0.7235	0.7235
RoofMatl_Membran	0.6572	0.6572
RoofMatl_CompShg	0.6360	0.6360
RoofMatl_Tar&Grv	0.6229	0.6229
RoofMatl_Metal	0.6212	0.6212

It can be seen that for both Ridge and Lasso regression there is slight deterioration in model performance metrics (R Square and MSE). The coefficient magnitudes and order of variable importance also changes slightly.

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:

Based on 5 folds CV the optimum parameters for alpha is calculated as follows

Ridge : 0.14532

Lasso : 0.0000099

The following model performance parameters can be seen on Train and Test set

Ridge (optimum alpha)				Lasso (optimum alpha)		
Metric	Train	Test		Metric	Train	Test
RSquare	0.8671	0.8863		RSquare	0.8802	0.8892
MSE	0.0016	0.0014		MSE	0.0015	0.0014

It can be seen that model performance is same across Train and Test indicating no underfitting.

The model performance for Lasso regression is slightly better than Ridge Regression across train and test and for both R Square and MSE. As a result will choose lasso regression parameters in this case given the slightly better model performance.

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 5 most important variables in the original equation are as follows

1. RoofMatl: Roof material
2. 1stFlrSF: First Floor square feet
3. TotalBsmtSF: Total square feet of basement area
4. 2ndFlrSF: Second floor square feet
5. BsmtFinSF1: Type 1 finished square feet

If we exclude the top 5 variables and redo the whole iterative variable selection process the top 5 remaining variables are found to be as follows.

1. GrLivArea: Above grade (ground) living area square feet
2. LotArea: Lot size in square feet
3. OverallQual: Rates the overall material and finish of the house
4. Bedroom: Bedrooms above grade (does NOT include basement bedrooms)
5. Neighborhood: Physical locations within Ames city limits

It can be seen that the resultant model is still able to capture around 83% of the variance in SalesPrice. There was only 5% deterioration in R Square due to dropping the top 5 variables.

Model performance of the new model is as shown below.

Ridge (drop top 5 var)				Lasso (drop top 5 var)		
Metric	Train	Test		Metric	Train	Test
r2	0.8296	0.8618		r2	0.8296	0.8618
mse	0.0021	0.0017		mse	0.0021	0.0017

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

To make a model more robust and generalisable, overfitting needs to be avoided. An overfitted model will have very good performance/good accuracy on training data however it has just memorized the training data without correctly modelling the underlying relationships. As a result, the performance/accuracy on unseen data will be subpar as the estimated model is not robust and hence not generalisable. Generally, models with low bias (more complex models) tend to overfit compared to models with high bias (low complexity).

While overfitting needs to be avoided, underfitting also needs to be avoided. In case of underfitting the model performs worse on both seen and unseen data as it has not optimally captured the underlying patterns. This may happen because the chosen model methodology is not complex enough to model the underlying patterns or due to some issues with parameter estimation like lack of sufficient data, etc.

As a result, we need to manage the model complexity such that it is neither too high nor too low so as to avoid both overfitting and underfitting.

Regularization helps with managing model complexity by increasing the bias slightly but at the same time significantly reducing the variance and hence an overall reduction in total error. By controlling the hyperparameter lambda the magnitude of the reduction in model complexity can be controlled. The hyperparameter Lambda is tuned on a validation data so as to arrive at an optimum value which results in the resulting model neither being too complex nor too simple. This results in a robust and generalisable model which performs equally well on both seen and unseen data. The regularization is done by adding a penalty term to the overall cost function. Depending on the penalty term we can have Ridge or Lasso Regularization. Lasso regularization can result in parameter being reduced to zero for certain variables and hence can help in variable reduction.