

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:

- Alpha values:

Ridge – 100

Lasso – 0.01

- After doubling the alpha value,

For Lasso model, training r2 increased and test r2 decreased , making the difference to jump from 3% to 5%.

For Ridge mode, training r2 decreased but the test r2 remained unaffected. Also the r2 difference is not significantly chaged!

- Top predictor variables for Ridge: (will give top 10)

```
[  
  "OverallQual",  
  "GrLivArea",  
  "RoofMatl_CompShg",  
  "1stFlrSF",  
  "OverallCond",  
  "Condition2_PosN",  
  "2ndFlrSF",  
  "YearBuilt",  
  "GarageCars",  
  "RoofMatl_WdShngl"  
]
```

- Top predictor variables for Lasso: (will give top 15)

```
[  
  "GrLivArea",  
  "OverallQual",  
  "GarageCars",  
  "YearBuilt",  
  "Condition2_PosN",  
  "YearRemodAdd",  
  "OverallCond",  
  "BsmtFullBath",  
  "Fireplaces",  
  "TotalBsmtSF"  
]
```

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:

Will go with lasso as it performs auto feature selection by making some of the coefficients zero and the r^2 difference is only 3% which is better than ridge. The test r^2 seems to be almost same which indicates generalization performance is same on both models and hence it would be better to choose lasso for better efficiency and 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:

Below are the top 5 predictor variables contributed to Lasso model:

- "OverallQual"
- "GrLivArea"
- "RoofMatl_CompShg"
- "1stFlrSF"
- "OverallCond"

After removing and building the model, below are the top 5 predictor variables:

- "1stFlrSF"
- "2ndFlrSF"
- "GarageArea"
- "YearRemodAdd"
- "Fireplaces"

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:

We need to perform cross validation to determine best suited alpha for reliable estimates.

Use regularization techniques like ridge/lasso to avoid overfitting the model. This improves the generalization on unseen data.

Perform feature selection/elimination to simplify the model as much as possible.

In our case Lasso model shows good robustness.