Advanced Regression Assignment - 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 :

1. Optimal value of Alpha for Ridge regression = 10.0

   Optimal value of Alpha for Lasso regression = 100.0

Current Metric with optimal value of Alpha is below –

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.946	0.940
R2 Score (Test)	0.891	0.879

After doubling the value of Alpha i.e. 20.0 for Ridge regression & 200.0 for Lasso regression below is metric

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.938	0.926
R2 Score (Test)	0.893	0.889

Top 10 important feature before Alpha is changed –

For Ridge Regression

OverallQual_9	23403.989
GrLivArea	18676.613
OverallQual_8	16702.535
OverallCond_9	15795.155
Neighborhood_Crawfor	14311.412
Functional_Typ	12327.238
Exterior1st_BrkFace	11401.412
TotalBsmtSF	10837.979
Neighborhood_Somerst	9805.271
PoolQC_NA	9650.922

For Lasso Regression

OverallQual_9	44778.938
OverallQual_8	30454.828
GrLivArea	20717.439
Neighborhood_Crawfor	15900.528
Functional_Typ	13983.300
OverallCond_9	12800.600
Exterior1st_BrkFace	11750.661
OverallQual_7	10355.511
Neighborhood_Somerst	10266.145
BsmtExposure_Gd	9197.501

Top 10 important features after changes

For Ridge Regression

OverallQual_9	18330.421
GrLivArea	16910.495
OverallQual_8	15302.973
Neighborhood_Crawfor	11430.141
Functional_Typ	10705.796
OverallCond_9	10523.416
TotalBsmtSF	10215.450
Exterior1st_BrkFace	9011.950
Neighborhood_Somerst	8238.283
Neighborhood_NridgHt	7886.605

For Lasso Regression

OverallQual_9	44778.938
OverallQual_8	30454.828
GrLivArea	20717.439
Neighborhood_Crawfor	15900.528
Functional_Typ	13983.300
OverallCond_9	12800.600
Exterior1st_BrkFace	11750.661
OverallQual_7	10355.511
Neighborhood_Somerst	10266.145
BsmtExposure_Gd	9197.501

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 : The selection between Lasso & Ridge depends on business scenario. Ridge would use all available features, irrespective of number of features.

In case you have limited number of features, and you want to keep all of them in model, then we will use Ridge Regression.

However, Lasso would cause some of co-efficient to Zero, effectively removing that feature from model. In case you need to remove un-necessary features and want to select fewer important features then we will go ahead with Lasso Regression.

In this case, since we have 300+ features (after on hot encoding) , I would go with Lasso Regression, since the there is marginal difference in R2-score of Lasso and Ridge.

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 : We have below top 5 features after earlier top 5 features are removed –

OverallCond_9
2ndFlrSF
Exterior1st_BrkFace
OverallCond_7
OverallCond_8

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

There are few checks of a model's Robustness and generalization –

1. **Less variation -** There is NO much variation in R2 score of a model on test and training data.

For our both models there is NO much difference between R2 score. Hence we can conclude that our models have less variations.

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.946	0.940
R2 Score (Test)	0.891	0.879

2. No over-fitting – Over-fitting occurs when model performs exceptionally well on Training data but performs poorly on test data. Since our model perform almost equally well on training and test data, we can say our models are NOT over-fit.

3. Correct trade-off between Complexity vs Simplicity. Model should NOT be overly complex as well as it should NOT be very simple. Complex modules have less Bias and more Variance, whereas Simple models are more bias and less variance. We need to tune the model with hyper-parameter such that Both, Bias and Variance should be less. We can perform regularization like we did for Lasso and Ridge regression.

4. Good accuracy – Robust models have good accuracy on test and training data.