Assignment

Advanced Regression – Housing Price Prediction

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:

Optimal value is when we have the highest R2 value (0.83)

Optimal Value for Ridge = 2

Optimal value for Lasso = 0.0001.

If we double the value of Ridge to 4 then R2 score of the model is still around 0.826.

The MSE of the model on the test dataset for doubled alpha is 0.00186

Out[133]:

	Ridge Doubled Alpha Co-Efficient
Total_sqr_footage	0.149028
GarageArea	0.091803
TotRmsAbvGrd	0.068283
OverallCond	0.043303
LotArea	0.038824
Total_porch_sf	0.033870
CentralAir_Y	0.031832
LotFrontage	0.027526
Neighborhood_StoneBr	0.026581
OpenPorchSF	0.022713
MSSubClass_70	0.022189
Alley_Pave	0.021672
Neighborhood_Veenker	0.020098
BsmtQual_Ex	0.019949
KitchenQual_Ex	0.019787
HouseStyle_2.5Unf	0.018952
MasVnrType_Stone	0.018388
PavedDrive_P	0.017973
RoofMatl_WdShngl	0.017856
PavedDrive_Y	0.016840

Out[127]:

	Ridge Co-Efficient
Total_sqr_footage	0.169122
GarageArea	0.101585
TotRmsAbvGrd	0.067348
OverallCond	0.047652
LotArea	0.043941
CentralAir_Y	0.032034
LotFrontage	0.031772
Total_porch_sf	0.031639
Neighborhood_StoneBr	0.029093
Alley_Pave	0.024270
OpenPorchSF	0.023148
MSSubClass_70	0.022995
RoofMatl_WdShngl	0.022586
Neighborhood_Veenker	0.022410
SaleType_Con	0.022293
HouseStyle_2.5Unf	0.021873
PavedDrive_P	0.020160
KitchenQual_Ex	0.019378
LandContour_HLS	0.018595
SaleType_Oth	0.018123

Lasso:

	Lasso Co-Efficient
Total_sqr_footage	0.202244
GarageArea	0.110863
TotRmsAbvGrd	0.063161
OverallCond	0.046686
LotArea	0.044597
CentralAir_Y	0.033294
Total_porch_sf	0.028923
Neighborhood_StoneBr	0.023370
Alley_Pave	0.020848
OpenPorchSF	0.020776
MSSubClass_70	0.018898
LandContour_HLS	0.017279
KitchenQual_Ex	0.016795
BsmtQual_Ex	0.016710
Condition1_Norm	0.015551
Neighborhood_Veenker	0.014707
MasVnrType_Stone	0.014389
PavedDrive_P	0.013578
LotFrontage	0.013377
PavedDrive_Y	0.012363

Out[134]:

	Lasso Doubled Alpha Co-Efficient
Total_sqr_footage	0.204642
GarageArea	0.103822
TotRmsAbvGrd	0.064902
OverallCond	0.042168
CentralAir_Y	0.033113
Total_porch_sf	0.030659
LotArea	0.025909
BsmtQual_Ex	0.018128
Neighborhood_StoneBr	0.017152
Alley_Pave	0.016628
OpenPorchSF	0.016490
KitchenQual_Ex	0.016359
LandContour_HLS	0.014793
MSSubClass_70	0.014495
MasVnrType_Stone	0.013292
Condition1_Norm	0.012674
BsmtCond_TA	0.011677
SaleCondition_Partial	0.011236
LotConfig_CulDSac	0.008776
PavedDrive_Y	0.008685

We don’t see much difference in the top predictor variables as the alpha value is already small. So, doubling it doesn’t make much change.

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:

Optimal lambda values

1. Ridge – 2
2. Lasso – 0.0001

Mean Squared Errors –

1. Ridge ~ 0.00184
2. Lasso ~ 0.00186

Both Lasso and Ridge have almost same MSE.

But I would choose **Lasso** over Ridge as it helps in feature reduction by setting the coefficients of less important features to zero.

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:

5 most important predictor variables in Lasso model are –

1. Total_sqr_footage
2. GarageArea
3. TotRmsAbvGrd
4. OverallCond
5. LotArea

After making the model in Jupiter notebook by removing these variables we find –

R2 value ~ 0.73

MSE value ~ 0.0028

We see that the R2 value has dropped significantly and also MSE has increased.

New Top 5 predictors are –

1. LotFrontage
2. Total_porch_sf
3. HouseStyle_2.5Unf
4. HouseStype_2.5Fin
5. Neighbourhood_Veenker

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 should follow Occam's Razor principle while designing the models –
"We should keep the model as simple as possible without compromising on accuracy significantly". Simpler models are more robust and generalizable as they have not "memorized" the dataset.
Keeping the model simple also helps us avoid problem of "Overfitting".
We can avoid "Overfitting" by ensuring that the model has similar performance on the training and test data set.
To achieve this fine balance of correct level of complexity in models so that we have a robust and generic model but not too naïve, that is it of no use -
we use **Regularization** techniques.
A balanced model will have best combination of Bias and Variance (i.e. the intersection point in Bias Variance Curve) which ensures the **least Total Error**.


Optimal Model Complexity
Total Error
Varianc
Bias^2
Model Complexity