Author: Kumaraguru Muthuraj.

Oct22 Cohort: MLAI

Surprise Housing Assignment on Ridge and Lasso regularization.

Assignment Part-I – Answers are in the notebook towards the end.

Assignment Part-II

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?

Answers: Lambda for Ridge and Lasso - **The lambdas for Ridge and Lasso are 1.5 and 0.0001 respectively**. The R2 values for Ridge and Lasso doesn't change drastically after doubling the alphas. ***Refer notebook for code.***

Top 5 Betas from Ridge Regularization
OverallQual 0.111483
Condition2_posn 0.105463
1stFlrSF 0.098451
2ndFlrSF 0.098262
YearBuilt 0.082551

Top 5 Betas from Lasso Regularization
Condition2_posn 0.247190
1stFlrSF 0.161111
2ndFlrSF 0.156739
OverallQual 0.137956
YearBuilt 0.083046

There are changes to the coefficients in magnitude, order and sign when the alpha is doubled.

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?

Answers: I stick to the alphas of Ridge and Lasso (1.5 and 0.0001). Of these 2 models, Ridge gives a train and test R2 of 0.89 and 0.86 whereas Lasso gives 0.90 and 0.84. Since the accuracy is superior with Ridge, we should be choosing Ridge model, but since there are 40 features, interpretation is difficult. I will choose to go with Lasso model because it has selected 28 features of 40 which is easy to interpret. If we had chosen say 30 or lesser features from RFE, the experiments yielded bad R2 values. Hence I choose 40 features from RFE and finally ended with **Lasso model.**

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: Note that when top 5 features identified are missing, it doesn't mean we will just pick the next 5 in the betas list. We have to remove these columns from test and train data, rebuild the model and pick the top 5. Applying this for Linear Regression to pick 40 features followed by Lasso, we get the following top 5 features. ***Refer notebook for code***.

Top 5 Lasso regularized model
GrLivArea 0.189910
RoofMatl_wdshngl 0.117735
TotalBsmtSF 0.108552
KitchenQual_fa -0.082891
BsmtQual_ta -0.077302

'R2Train': 0.84, 'R2Test': 0.82.
The predictive power has reduced a lot but is still accurate.

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: Consider a linear regression model is not regularized. This falls in the zone of lowest bias and highest variance (in the bias-variance plot). We want to move to the zone towards left, where the bias and variance are just right, i.e., lowest total error. For this, we apply regularization and adjust the lambda starting 0. As lambda is increased, we move to the left in the bias-variance plot. By this, we make sure bias increases and variance reduces. This implies that the accuracy on test data increases. Robustness means low bias and generalizable is low variance. We cannot achieve both and hence a tradeoff. The model might not be accurate with the optimal zone, but the variance is definitely much lesser. To sum up, with a model that is highly regularized, the accuracy reduces but variance is gained. This means the model is not overfit and more generalizable.