Assignment Part II

Assignment-based Subjective Questions

Q1 . 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 :

The Optimal value of alpha for ridge & Lasso Regression are :

Ridge - 0.01

Lasso - 0.0001

After doubling the coefficient values in lasso, the most important predictor variables are :

1. MSZoning
2. LotArea
3. Street
4. LandContour
5. Utilities

After doubling the coefficient values in ridge, the most important predictor variables are :

1. MSZoning
2. LotArea
3. Street
4. LandContour
5. Utilities

Q2. 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 :

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.917782	0.913420
1	R2 Score (Test)	0.890028	0.889238
2	RSS (Train)	13.195279	13.895450
3	RSS (Test)	7.925544	7.982439
4	MSE (Train)	0.113683	0.116660
5	MSE (Test)	0.134517	0.134999

- From the above table we can see the model performance by Ridge Regression is better in terms of R2 values of Train and Test data sets, it is better to use the Lasso Model as it more robust because it assigns the 0 values to insignificant features. Lasso Equation that model provides is :

- Log (Y) = Constant + 0.886379(GrLivArea) + 0.430843(OverallQual) + 0.412558(1stFlrSF) + 0.323463(OverallCond) + 0.264828(LotArea) + 0.242423(Neighborhood) + 0.199273(GarageCars) + 0.147726(Utilities) + 0.147569(GarageQual) + 0.143458(BsmtFullBath) + Error term (RSS + alpha * (sum of absolute value of coefficients)
- Suggestions for Surprise Housing is to keep a check on these predictors affecting the price of the house. The higher values of positive coefficients suggest a high sale value. Some of those features are:-
    - GrLivArea
    - OverallQual
    - 1stFlrSF
    - OverallCond
    - LotArea
    - GarageCars
    - Neighborhood
- The higher values of negative coeeficients suggest a decrease in sale value.Some of those features are:-
    - PoolQC
    - YearBuilt

Model Conclusion

- Lasso is better model Since Lasso helps in feature reduction (as the coefficient value of insignificant features became 0)
- Hence based on Lasso, the factors that generally affect the price are the Living area square feet, Overall quality, First Floor square feet and condition of the house, Lot size in square feet, Number of cars that can be accommodated in the garage and the Physical locations within Ames city limits
- The variables predicted by Lasso (Mentioned above)as significant variables for predicting the price of a house.

Q3. 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

Once we remove the 5 most important predictor variables and build a new model, The 5 most important predictor variables are:

1. MSZoning
2. Street
3. LandContour
4. Utilities
5. Neighborhood

Q4. 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 :
The model should be generalized so that the test accuracy is not significant lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much importance should not give to the outliers so that the accuracy predicted by the model is high i.e, Model should not be overfit. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis. Model which is not Robust and Generalised not able to make the accurate and reliable predictions over general data outside our Train & test sets.