Advanced Regressision Assignment

[Problem Statement 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?

Answer:

For ridge regression the optimal value of alpha is 50

For lasso regression the optimal value of alpha is 100

When alpha is doubled for lasso regression

a) MSE or train set increased

b) R2 Square on train set decreased , R2 on test data too slightly decreased

Most important predictor variables are

- OverallCond
- BsmtFullBath
- GarageArea
- Neighborhood_OldTown
- Neighborhood_NridgHt
- Condition1_PosA
- 2ndFlrSF
- HouseStyle_2.5Fin
- TotRmsAbvGrd
- ExterCond

When alpha is doubled for lasso regression

a) MSE or train set increased

b) R2 Square on train set decreased , R2 on test data too slightly decreased

c) number of non zero coefficients reduced to 81 from 112

Most important predictor variables are

- Exterior1st_AsphShn
- Neighborhood_OldTown
- Neighborhood_NridgHt
- Neighborhood_Timber
- BsmtFullBath
- HouseStyle_2.5Fin
- RoofMatl_Membran
- Exterior2nd_MetalSd
- Neighborhood_Edwards
- GarageArea

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 test accuracy of Lasso Regression model(85.60) is slightly higher than that of Ridge Regression model (85.20). Also the number of features reduced to 112 for Lasso regression. Since Lasso regression is giving simple model with better accuracy, the Lasso regression model will be chosen to apply

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 : The Five most important predictor variables in lasso model are

- Exterior1st_AsphShn
- Neighborhood_OldTown
- Neighborhood_NridgHt
- Neighborhood_Timber
- BsmtFullBath

After removing the above predcitors the most important varaibles are

- Exterior1st_BrkComm
- BsmtHalfBath
- RoofMatl_Membran
- Neighborhood_SWISU
- HouseStyle_2.5Fin

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?
To make model robust and generalization it should perform well in training and testin set. To ensure that

1. Model should avoid underfitting by increasing more data or increasing complexity of model by adding more features, allowing nonlinearity through data transformation (ex: Polynomial degree). This ensures model performas very well on training data
2. Model should avoid overfitting. Regularization methods such as Lasso and Ridge should be applied. This reduces the complexity of the model
3. Bias & Variance tradeoff should be applied to get optimal value

When model is robust and generalized it may perform little worse in training set as bias is increased. But since the model is generalized and variance is

decreased the model performs well in test data set.There for the accuracy difference between training and test dataset is less