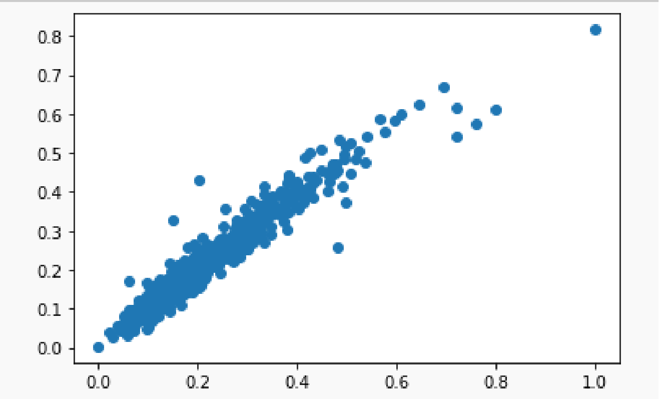
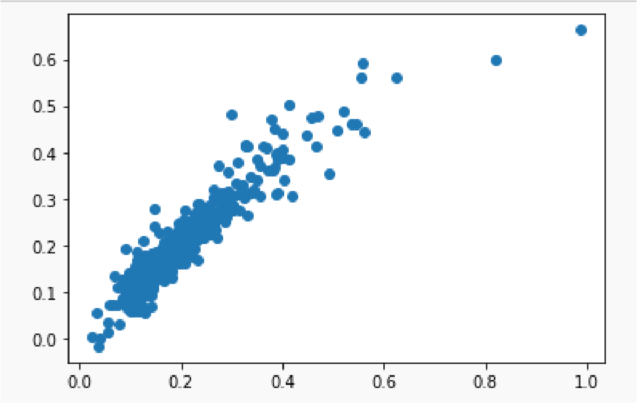
Advance Regression - Part 2

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: In ridge regression when we plot the curve between negative mean absolute error and alpha we see that as the value of alpha increase from 0 the error term decrease and the train error is showing increasing trend when value of alpha increased. When the value of alpha is 2 the test error isminimum so I decided to go with value of alpha equal to 2 for ridge regression.

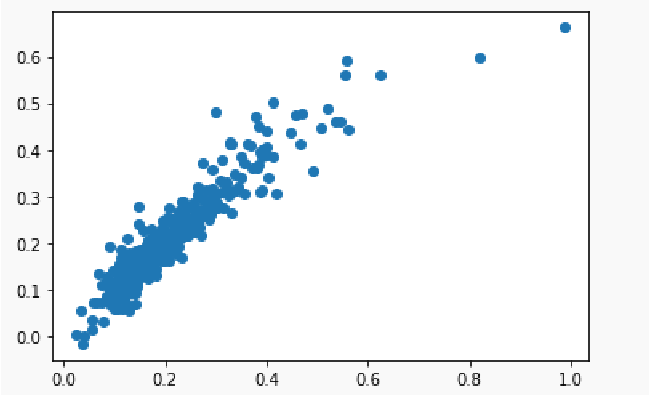


Ridge regression Train set

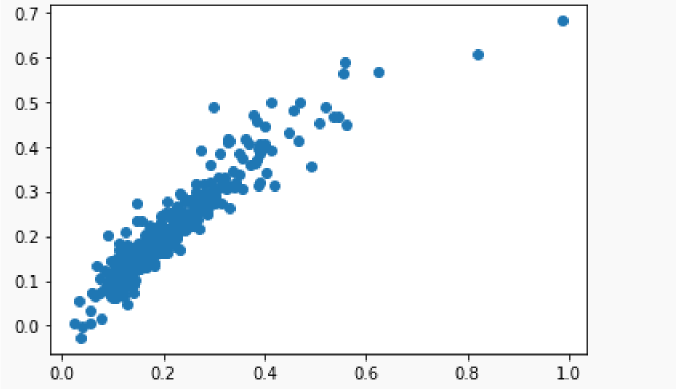


Ridge regression Test set

For lasso regression I have decided to keep very minor value that is 0.01, when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero. Initially it came as 0.4 in negative mean absolute error and alpha.



Lasso regression Train set



Lasso regression Test set

When I double value of alpha for our ridge regression number, I will have to take the value of alpha equal to 10 the model will apply more penalty on the curve and try to make the model more general that is making model more simpler and number thinking to fit every data of the data set. From the graph we can see that when alpha is 10, we get more error for both test and train.
Likewise, when we increase value of alpha for lasso we try to penalize more to our model and more co-efficient of variable will be reduced to zero, when we increase value of our r2 square decreases.
The most significant variable after the changes are applied for ridge regression are given below:-

1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

The significant variable after the changes has been applied for lasso regression are mentioned below:

1. GrLivArea
2. OverallQual

3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage

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: Regularizing coefficients is important to improve the prediction accuracy also with decrease in variance, makes the model interpretable.

Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum or squares should be small by using penalty. Penalty is lambda times sum of squares of the coefficients, hence coefficients that have greater values get penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

Lasso regression, uses a tuning parameter called lambda as penalty is absolute value of magnitude of coefficients which is identified by cross validation. As lambda value increases Lasso shrinks coefficient towards zero and it makes variables exactly equal to 0. Lasso also does variable selection.When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the 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 thefive most important predictor variables. Which are the five most important predictor variables now?

Those 5 most important predictor variables that will be excluded are :-

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageArea

Question 4: How can you make sure that a model is robust and generalisable? What are the implications of thesame for the accuracy of the model and why?

Answer: The model should simple, though its accuracy decreases but it will be more robust and generalizable. It can be understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Implication in terms of accuracy that a robust and generalizable model will perform equally well on training and test data i.e. the accuracy does not change much for training and test data.

Bias is error in model, when the model is weak to learn from the data. High bias means model isunable to learn details in data. Model performs poor on training and testing data.

Variance is error in model, when model tries to over learn from data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid over-fitting and under-fitting of the data.