Modeling Process and Result:

Since this is a regression problem, we construct our model in 4 different ways including lasso regression, support vector regression, neural network and KNN methods. Then we compare their performance based on the score of the regression and the median relative absolute error of training dataset and testing dataset.

Firstly, we used KNN methods. However, regardless of the leaf size as a hyper parameter we choose, the score of the regression is not ideal (less than 0.5) and the MRAE indicators are high.



Next, we try the Support Vector Regression methods with different kernel types including rbf, sigmoid and linear. Among these types, linear kernel performs best as it produces highest regression score and lowest MRAE values.



Then we use the neural network method to predict the count of students. However, great volatility occurs in the plot. The performance of neural network seems to be unstable. The likelihood of overfitting is very high.



Lastly, we use the lasso regression method. Its performance showed in the plot is quite similar to the performance of support vector regression using linear kernel.



We can draw a conclusion here that the relationship between the target and features is likely to be linear based on the performance of the various methods we used.

While the score of regression is very high, we are also concerned whether the count of students last year will be a dominant feature that contributes to the most of the change of the target variable. Thus, we also fit a model excluding any other variable except the count of students last year. There is the performance of these model



From the plots above, we find that the score is greatly reduced by nearly 10%, which indicates that the other features did add some additional valuable information to our model.