Regression

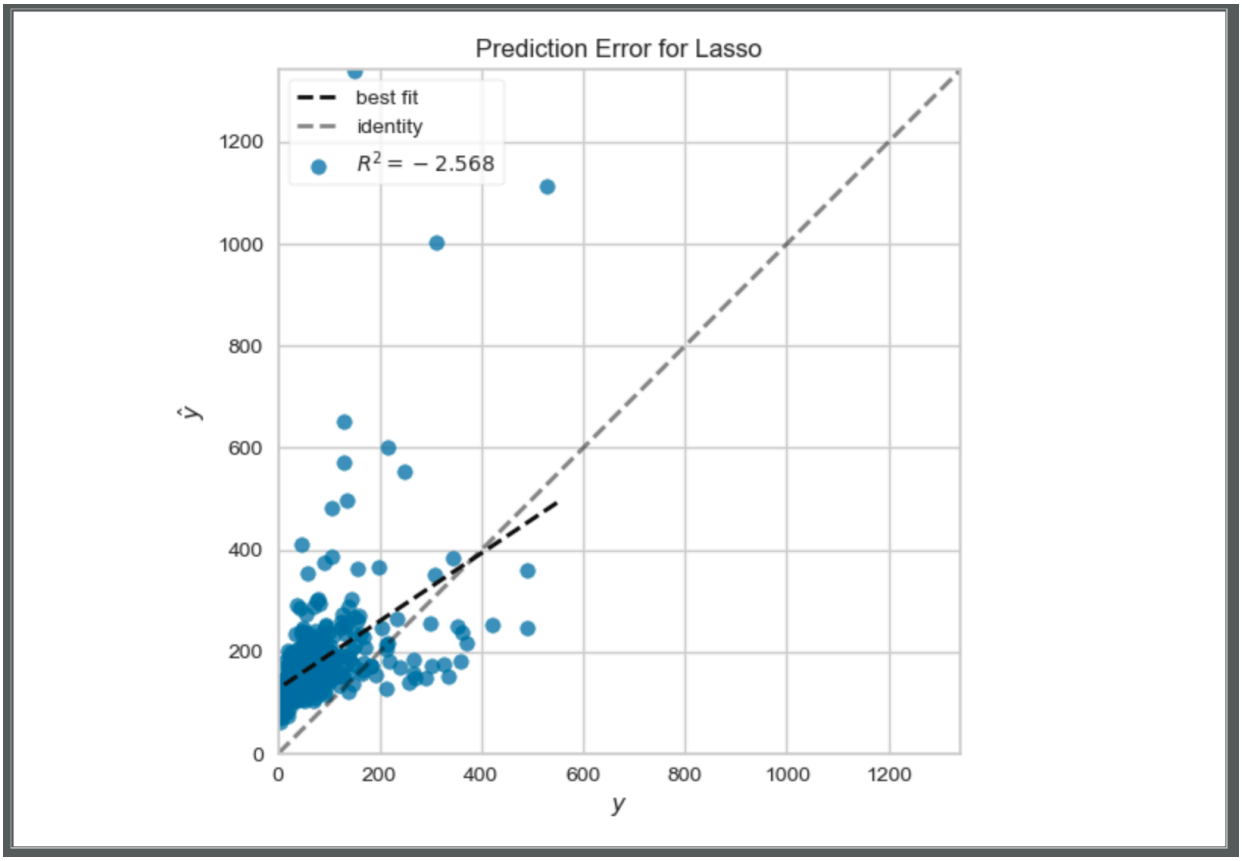
1. Data

We chose three numerical variables: duration, release time, and view as input data to predict regression variables - comments. We have about 2,500 pieces of data, and we selected 500 of them as test data, which is 20% of the total amount of data.

2. Analysis

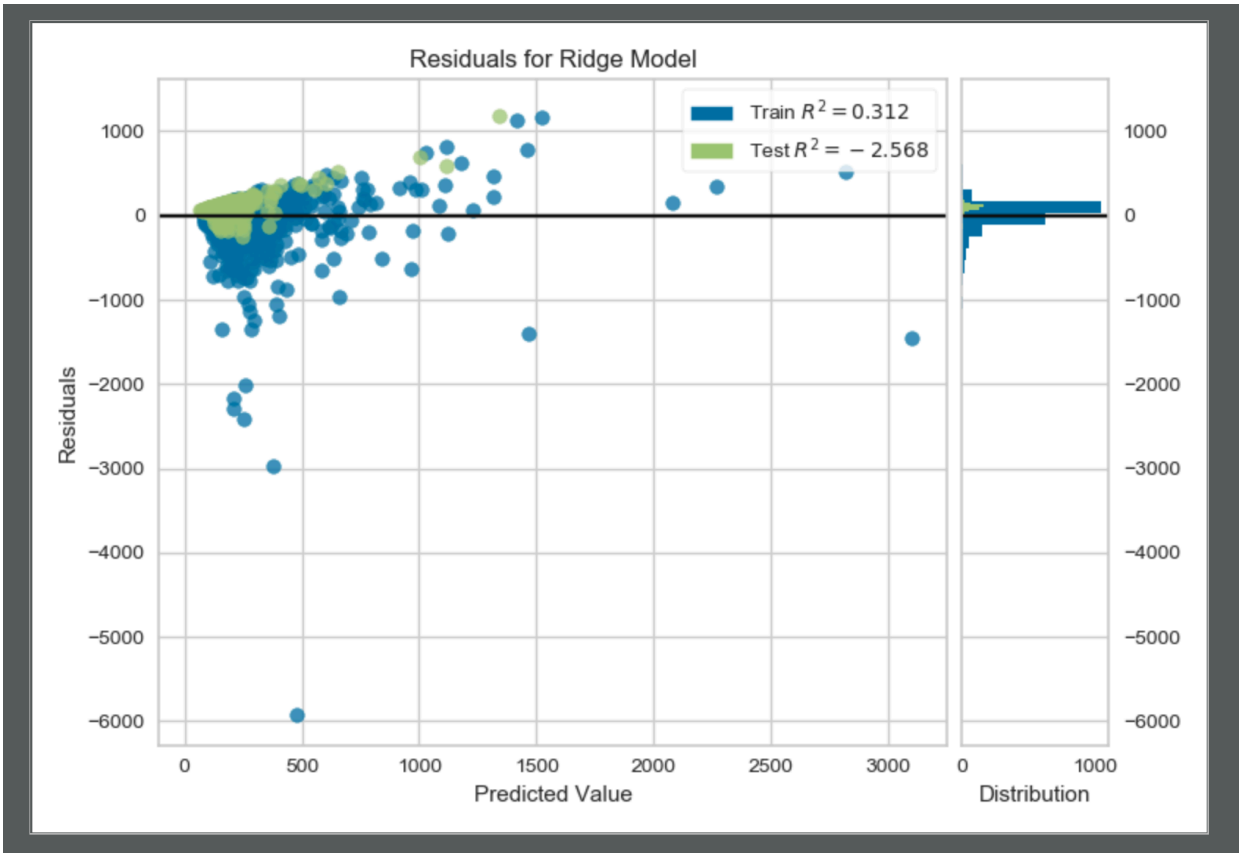
After Sklearn's linear model calculation, we obtained a linear model, and we calculated the error of this linear model using Yellowbrick.

> 1. Prediction Error Plot



A 45 degree angle means that the observed and predicted values are equal. From the graph we find that the predicted values calculated by our model are usually larger than the observed values.

> 2. Residuals Plot



The residual graph shows the relationship between the predicted value and the residual value. It can be seen that the residual value of our predicted value and real value is small, and the predictive model has achieved good results.

> 3. MSE & RMSE

1) Linear Regression

We calculated the MSE and RMSE of the linear regression model and obtained the following results:

```
MSE: 19320.57814808802
MSE(Calculate MSE according to the formula): 19320.578148088014
RMSE: 138.99848253879617
```

MSE is 19320.578, and RMSE is 138.998.

2) Random Forest Regression

We also calculated the MSE and RMSE, while we are using the random forest algorithm to get a regression model. The results are different while choosing different 'max_depth' in the algorithm. The best result is shown as follow :

```
MSE: 8513.061755800001
MSE(Calculate MSE according to the formula): 8513.061755799996
RMSE: 92.26625469693674
```