Machine Learning: Regression

Author:
Adhok Katti

Data Description:

This data consists of information about the various communities in the USA. It is a cleaned data sectioned from the 1990 US LEMAS survey and the 1995 FBI UCR. Since it is already cleaned, there are no missing values present in the dataset. Some of the highlights are:

1. The dataset contains 101 features and is pre-split into training data, testing data, and validation data.
2. All the variables are of the 'float64' data type, no other data types are found in the datasets.
3. The training and validation datasets have 298 observations each and the testing dataset has 299 observations.
4. All the variables in training, testing and validation datasets were used in building the regression models, to predict violent crime per capita.
5. All of the variables have decimal values ranging from 0.00 to 1.00. I, therefore, did not apply any further transformations, such as standardization, on any of the training, testing, or validation datasets since we are not interested in the colinearity, or the problems thereof, present in the variables in the data.
6. On the training dataset, a lasso and ridge regression with the same alpha value (α=0) is trained initially. Then a validation dataset is then used to determine the best alpha value, and the model's performance is compared to the validation and testing datasets.

All the regression methods use simple terms, without any higher-order, interaction terms, and a straightforward methodology is employed, referred from Statsmodel and Scikit-learn's documentation pages.

Linear Regression:

The linear regression model is performed on the dataset using Statmodel's Ordinary Least Squares.

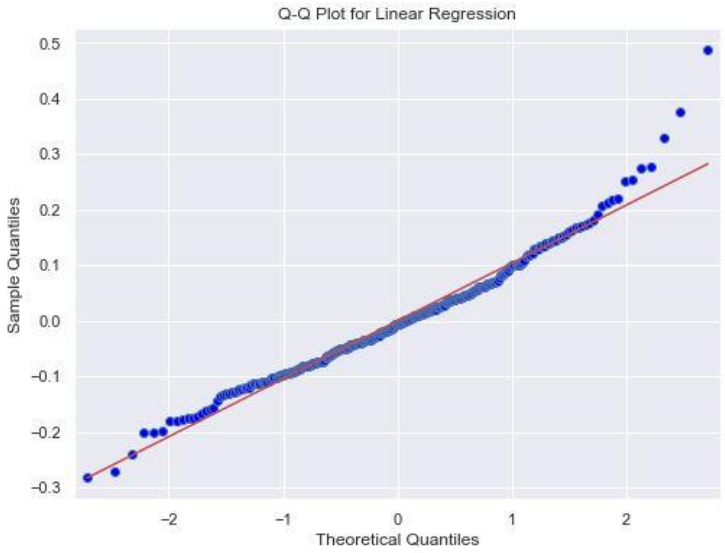

Figure 1: Q-Q plot

As we can observe, the underlying assumption of linear regression, that the residuals follow Gaussian distribution, does not hold good. We see a large deviation from the line of best fit towards the top and bottom residuals. However, most of the residuals in the middle section fall on the regression line.

	RMSE	**R-squared**
Training Data	0.104	0.759
Testing Data	0.153	0.784
Validation Data	0.155	0.817

Table 1: Linear Regression Statistics

The linear model suffers from overfitting to the training dataset as the error rate increases on the testing and validation datasets. Though the accuracy of the model increases in testing and validation datasets, the low training error with high testing and validation error results in the overfitting of data. Better data transformation with feature selection on the model will result in better overall results and can help in an optimum linear regression model.

Lasso Regression:

	RMSE	**R-squared**
Training Data	0.104	0.758
Testing Data	0.156	0.566

Validation Data	0.152	0.635

Table 2: Lasso Regression Statistics

α Value	R-squared
0	0.816
0.001	0.722
0.0001	0.801
0.00001	0.815
0.000001	0.816

Table 3: α Value Statistics for Lasso Regression on Validation Data

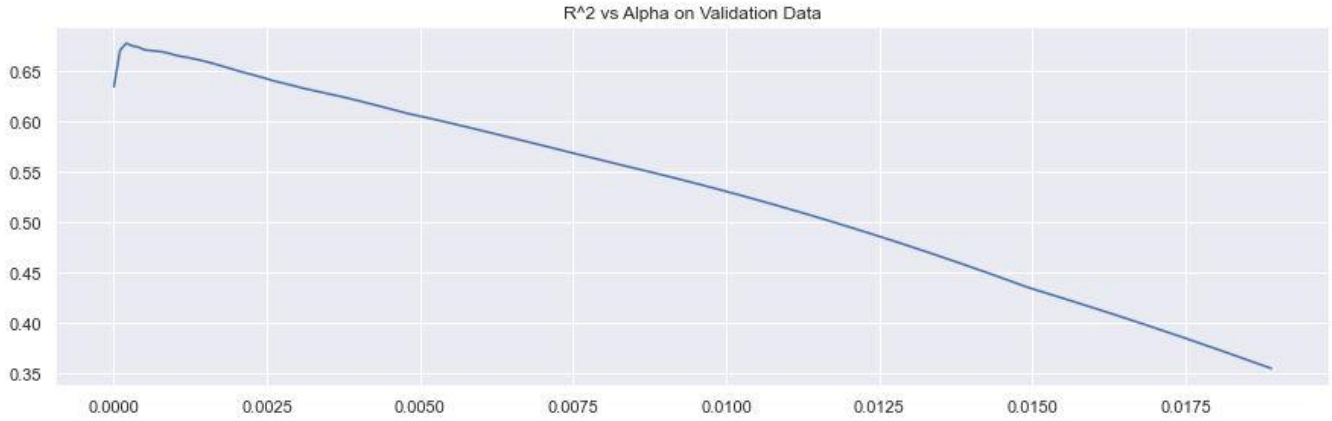

Figure 2: R^2 vs Alpha for Lasso Regression

Table 2 shows the general statistics of the lasso regression on all three datasets. It is distinctly evident that the model is overfitting to the training dataset and the error rate increases significantly on testing and validation datasets respectively. We also observe that the lasso model suffers a high accuracy loss even with small changes in the L1 weights from Table 3 and Figure 2. The model converges to a constant very quickly as the sum of absolute values becomes higher. Hence, the best value for an alpha that is selected for the model is 0.00001.

Ridge Regression:

	RMSE	R-squared
Training Data	0.104	0.759
Testing Data	0.153	0.553

Validation Data	0.155	0.621

Table 4: Ridge Regression Statistics

α Value	R-squared
0	0.817
0.001	0.817
0.01	0.815
0.5	0.781
1	0.766
10	0.713

Table 5: α Value Statistics for Ridge Regression on Validation Data

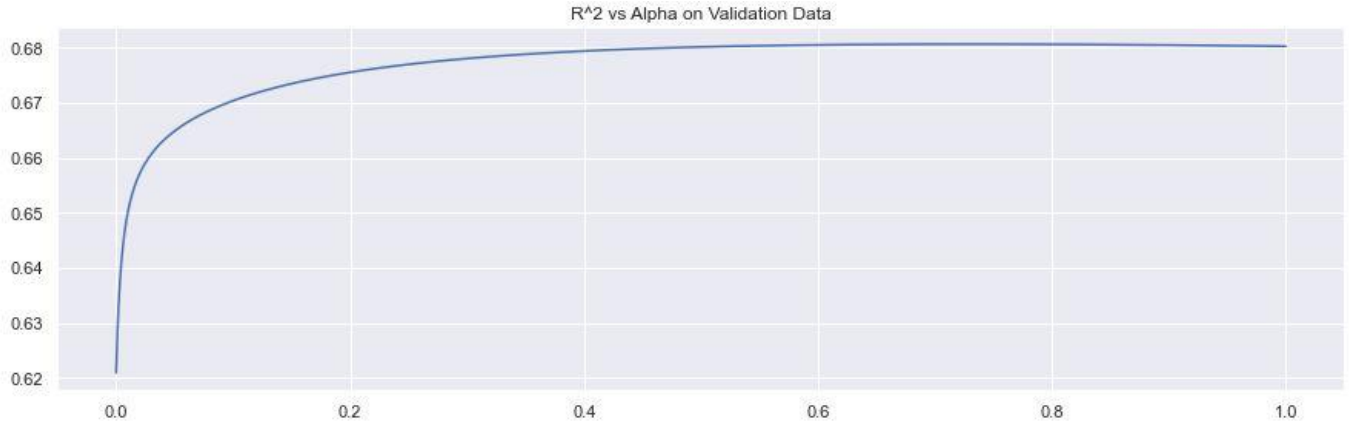

Figure 3: R^2 vs Alpha for Ridge Regression

Similar to Lasso, Ridge regression's accuracy suffers from overfitting to the training dataset as seen in Table 4 and the model's accuracy decreases when the alpha level increases as shown in Table 5. However, the accuracy of the Ridge regression tends to increase as the alpha level increases from 0 to 1 and eventually starts to decrease for higher levels of alpha. The best alpha level selected from the validation dataset is 0.001.