CS 519: Applied Machine Learning I

HW5: Regression

Submitted By: Md Ishtiaq Ahmed

Aggie ID: 800606216

Report

Dataset Description:

California housing dataset is used to test all the regressors. This dataset is available in the Sickit-learn library and can be loaded by using fetch_california_housing from sklearn.datasets. The goal of the dataset is to predict median house values based on the available features. The dataset has 8 explanatory variables and 1 target variable. Total instance is 20640.

All the columns and all the instances of the dataset have been used to do regression analysis.

Analysis

I have used a total of 11 regression models to compare the performance of the different regressors' behavior. For linear regression, I have used (i) LinearRegression, (ii) RANSACRegressor, (iii) Ridge, (iv) Lasso, and (v) ElasticNet. For non-linear regression, I used DecisionTree Regressor. Below is the performance summary:

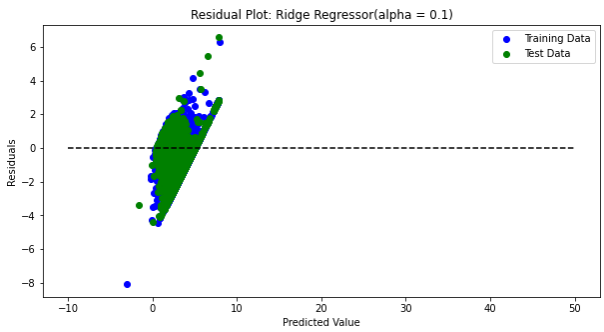
		MSE		R^2		Model Fitting
Model	**Hyper Parameter**	**Train**	**Test**	**Train**	**Test**	**Time (ms)**
Linear Regression	Default	0.608	0.612	0.546	0.534	83.994
RANSAC	min_samples = 50, random_state=1	5.643	0.618	-3.216	0.53	138.921
Ridge 1	alpha =0.1, solver = 'auto'	0.608	0.612	0.546	0.534	17.99
Ridge 2	alpha =1,solver = 'svd'	0.608	0.612	0.546	0.534	4.000
Lasso 1	alpha =0.1	0.683	0.677	0.49	0.485	3.998
Lasso 2	alpha =0.5	1.036	1.014	0.226	0.229	2.999
ElasticNet 1	alpha =0.1, l1_ratio = 0.5	0.66	0.656	0.507	0.501	3.997
ElasticNet 2	alpha =1, l1_ratio = 0.5	1.126	1.103	0.158	0.161	3.998
ElasticNet 3	alpha =1, l1_ratio = 0.1	0.883	0.863	0.341	0.344	2.999
DecisionTree 1 (Non-linear)	max_depth = 5	0.498	0.515	0.628	0.608	47.969
DecisionTree 2 (Non-linear)	max_depth = 20	0.007	0.724	0.995	0.45	135.926

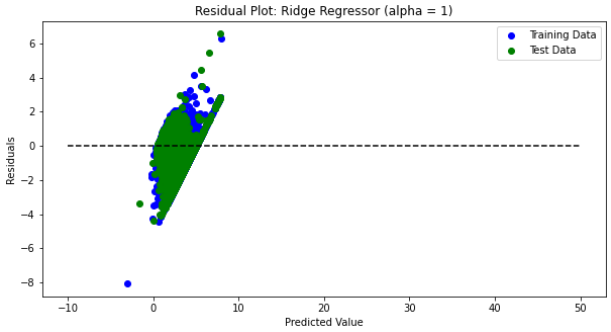
Linear Regression: This model has a moderate performance, with a train MSE of 0.608 and test MSE of 0.612. The R^2 values for both train and test sets are around 0.54, indicating that the model can explain around 54% of the variance in the data.

RANSAC Regressor: Performed poorly on the train set. However, its performance on the test set is better. The better performance on the test set can be attributed to the fact that the model has learned the

underlying relationship from the inliers in the train set. The fitting time for RANSAC is very slow compared to other models.

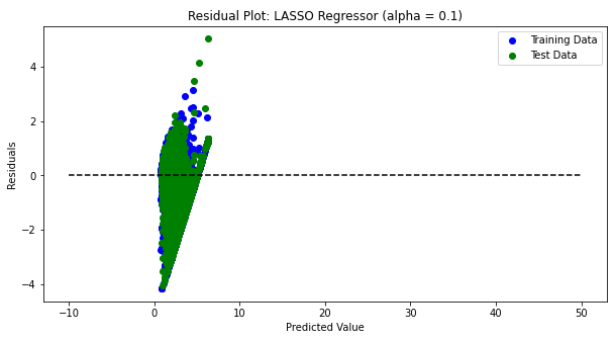
Ridge Regressor: This model's performance is identical to the linear regression model. But the fitting time is much faster than linear regression. There was no effect of changing the alpha and solver parameter values on MSE and r^2. Both of the ridge models do not seem to overfit as there is no significant difference between train and test performance.

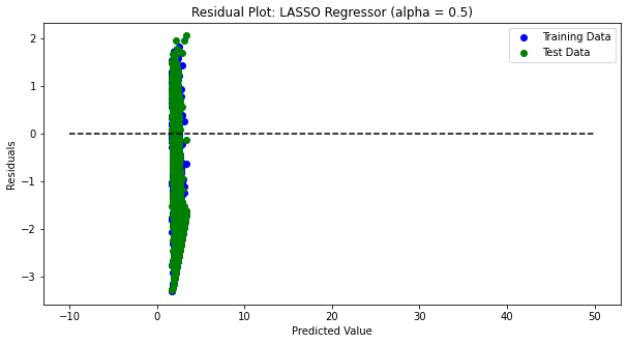




From the residual plots, we can also see that both ridge models are giving exactly the same performance. Also, we can see some outliers as well.

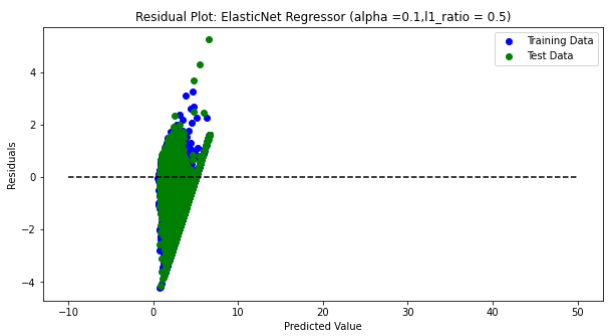
Lasso Regressor: Lasso 1 has slightly worse performance compared to Linear Regression, with train and test MSE values around 0.683 and 0.677, respectively. The R^2 values are around 0.49 for both train and test sets. Lasso 2 displays a worst performance than Lasso 1. Basically, the performance of the lasso regressor is degrading with respect to increased alpha value. This model is also very fast. No indication of overfitting here as well.

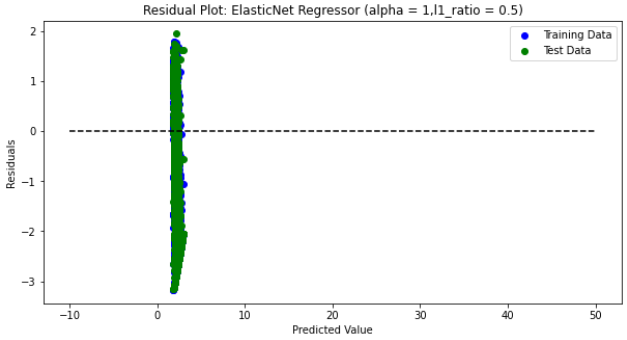


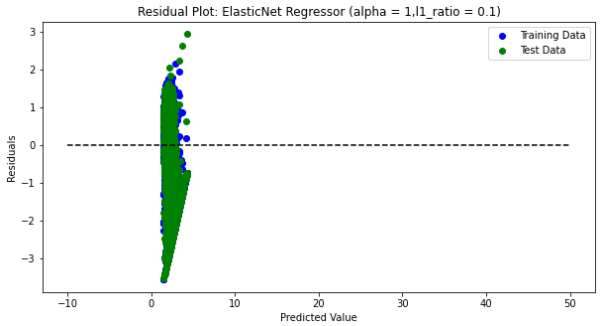


It is also observed from the residual plot that the second lasso model performed poorly and the model is not capturing the underlying relationship between the explanatory variables and target variables effectively.

ElasticNet Regressor: This model performed poorly with a higher alpha value (alpha =1) even if we changed the l1_ratio. ElasticNet 2 has a poor performance while ElasticNet 3 performed slightly better. ElasticNet 1 performed better than the other two and has similar performance like linear regression. Fitting time is also faster for all three Elasticnet models. This model also not showing any indication of overfitting.







If we check the residual plots, we can see that the data points are not randomly distributed. With l1_ratio 0.5 , the residual plots look more like the residual plots of the lasso regressor.

DecisionTree Regressor (Non-linear): DecisionTree 1 with max_depth = 5 appears to perform the best among all the models with MSE 0.515 and R^2 0.608. Also, as expected it requires more time for the model to fit the data. The performance on the train and test sets is similar which suggests that the model is not overfitting.

Unlike all the models, DecisionTree 2 with max_depth = 20 is showing clear indications of overfitting. This model has an extremely low train MSE of 0.007 and a high train R^2 value of 0.995, indicating an almost perfect fit on the training data. However, its performance on the test set is significantly worse compared to the train set, with a test MSE of 0.724 and an R^2 value of 0.45. This model's fitting time is also the largest among all due to its higher depth.

Below are the residual plots of both DecisionTree models: