Thomas Ehret

Assignment 5

CS519 SP23

Dr. Cao

Linear Regression Model Comparison

Testing Specifications

The California Housing dataset was used to compare the prediction capabilities of the following models: Linear Regression (LinReg), RANSAC, Ridge, Lasso, and ElasticNet (ENet). An additional comparison of linear vs non-linear regression models was undertaken by implementing the non-linear Decision Tree Regressor(DTReg. The performance of each model was measured by timing the 'fit' process, as well as recording the R2, mean squared error (MSE), and median absolute error (MedAE).

Defaults Parameters:

RANSACRegressor(random_state = 42)

Lasso(random_state = 42, alpha = 1.0)

Ridge (random_state = 42, alpha = 1.0)

ElasticNet(random_state = 42, alpha = 1.0)

LinearRegression (fit_intercept = True)

DecisionTreeRegressor(random_state=123, min_samples_split=25)

Initial Performance

Using the default parameters, the results are shown in table 1.

Metric	RANSAC	Lasso	Ridge	ENet	LinReg	DTReg
Fit Time	103ms	3ms	<1ms	<1ms	<1ms	122ms
MSE	0.79	1.31	0.56	0.20	0.56	0.40
R2	0.40	-0.00	0.58	0.20	0.58	0.69
MedAE	0.41	0.75	0.41	0.69	0.41	0.26

Table 1

Based on the performance comparison report of the linear regression models RANSAC, Ridge, ElasticNet, and Lasso using the California housing dataset, we can draw a few preliminary conclusions:

We can see that Ridge outperforms the other models, with the lowest mean squared error (MSE) of 0.56 and the highest R-squared (R2) score of 0.58. This indicates that the Ridge model provides the best fit to the data and is the most accurate predictor of the median home price. Both ElasticNet and Lasso have higher errors and lower R2 scores, indicating that they do not fit the data as well as Ridge. The unexpected result was RANSAC. I had assumed it would perform well on this dataset. However, the default parameters might have something to do with the poor performance. Given the RANSAC results , similar performance by Linear Regression was expected and that sentiment proved true. Overall, the decision tree regressor outperforms the linear models.

Further Testing

The following parameters were changed during the second phase of testing (only the changed parameters shown). The LinearRegression model is included in further testing as there isn't a statistically significant change by "tweaking" parameters.

RANSACRegressor(min_samples = 0.95, residual_threshold = None)

Lasso(alpha = 0.5)

Ridge (alpha = 0.5)

ElasticNet(alpha = 0.5)

DecisionTreeRegressor(criterion = "mean_absolute_error")

As evidence by the results in table 2, decreasing *alpha* resulted in Lasso and ElasticNEt still performing poorly. While changing the loss criterion of the decision tree increased the computational cost, it didn't produce much of a change in the other metrics.

Metric	RANSAC	Lasso	Ridge	ENet	DT Reg
Fit Time	504ms	3ms	2ms	4ms	8411ms
MSE	0.58	0.94	0.56	0.83	0.44
R2	0.56	0.28	0.58	0.37	0.67
MedAE	0.37	0.65	0.41	0.60	0.26

Table 2

The changes to RANSAC produced slightly better results, but also increased the computational cost dramatically (factor of 5). Although that is not a worry for this dataset, one could assume that a sufficiently large dataset would make this type of cost per instance computationally untenable. Overall, the Decision Tree Regressor still outperforms the other model.

Based on the performance comparison report of these linear regression models, one can say that Ridge is the most accurate predictor of the median house value per this dataset, followed by RANSAC. ElasticNet and Lasso may not be as effective on this dataset, possibly due to the relatively small feature size or the characteristics of the data. The Decision Tree Regressor is a special case because of its non-linearity and the tendency for decision trees to be prone to overfitting.