Lovemore Tenha
Project4 Report

In this report, we explore the performance of various regressors on 2 different datasets; the housing.data.txt dataset and the all_breakdown.csv dataset from the California Renewable Production. We use the regressors from the scikit-learn library in python, and an implementation of the normal equation version of the Linear regressor. The regressors are Linear Regressor, Ransac, Ridge, Lasso and a Non-Linear regressor with degree 3.
The following Results were obtained after applying the regressors on the housing dataset.

Regressor	**R^2_train**	**R^2_test**	**MSE_train**	**MSE_test**	**Runtime**
Linear	0.73	0.75	0.27	0.25	0.001
Ransac	0.73	0.75	0.27	0.25	0.002
Ridge	0.73	0.75	0.27	0.25	0.00
Lasso	0.0	-0.005	1.01	1.01	0.00
Non-Linear	1.0	-522.6	1.33	523	0.22
Normal Equation	0.64		0.26		0.00

Prediction Performance and Runtime for each classifier on housing dataset
All regressors, except for Lasso, had relatively high R2 value on the training set. The same regressors also had high R2 values on the testing data with the exception of the non-linear regressor, with an extremely low R2 value. The non-linear model, which is more complex, overfits the training set and thus results in poor prediction performance on the testing set. As expected, the Normal equation model had the fastest runtime and the Non-Linear model had the slowest runtime due to an additional step of adding polynomial features before linear regression. Surprising, lasso had a relatively poor performance on both the training and testing sets.

The following results were obtained after applying the classifiers on the California Renewable Production dataset.

The dataset had missing values, which were filled with the value 999.

Regressor	**R^2_train**	**R^2_test**	**MSE_train**	**MSE_test**	**Runtime**
Linear	0.13	0.12	0.89	0.88	0.010
Ransac	0.13	0.12	0.89	0.88	0.020
Ridge	0.13	0.12	0.89	0.88	0.004
Lasso	0.0	-0.0005	1.02	1.00	0.003
Non-Linear	0.34	0.33	0.67	0.67	0.87

The Non-Linear model had the best predictive performance on the renewable production as reflected by both R2 and MSE values. The values of the metrics were also consistent across both the training and testing sets suggesting that there was no overfitting. As expected, the non-linear model also had the highest runtime. Linear, Ransac, Ridge had relatively low predictive performance suggesting that a linear model is not the best model for the dataset. Lasso, had the worst predictive performance across both the training and testing sets.