Machine Learning Project

Data: Car-price

* This is a Regression Data.
* The data set has 25 features
* The Target value is the price of the data

Pre-processing

This dataset has 25 features, some of them have missing values.

The Missing values in numerical features were replaced by Statistical methods.

The Missing values in Categorical features were replaced by One Hot Encoder.

Some features with low variance were removed.

Now the Data is splited into 80 -20 % for Training and Testing.

The data is standardized by Standard Scaler from sk-learn.

Modelling

The Data is fitted into following models

* Linear Regression, this data has many features , so linear regression gives a bad train, test scores
* Ridge Regression, this model gives a Train score of 0.9560, and a Test score of 0.6908, this is overfitting
* Lasso, this model gives a Train score of 0.97660, and a Test score of 0.5798, this is overfitting
* Elastic Net, this model gives a Train score of 0.8560, and a Test score of 0.5606, this is overfitting
* SGD, this model gives a Train score of -0.1560, and a Test score of -0.2908, this is a Bad model
* Decision Tree, this model gives a Train score of 0.9560, and a Test score of 0.7108, this is overfitting
* SVR, this model gives a Train score of -0.1560, and a Test score of -0.2908, this is a Bad model
* Random Forest Regressor, this model gives a Train score of 0.9760, and a Test score of 0.7228, this is overfitting
* Gradient Boosting Regressor, this model gives a Train score of 0.9860, and a Test score of 0.6808, this is overfitting

Final Model

After Reducing the dimensionality into 15 by PCA, the model gives a good value