SUMMARY

Data set details

1. Dataset file: ‘cars_price.csv’

2. This is a regression data set.

2. The data set has 206 samples.

3. There are 25 features.

4. The target variable is the price of the car.

OVERVIEW OF THE ATTRIBUTES

symboling
normalized-losses
make
fuel-type
aspiration
num-of-doors
body-style
drive-wheels
engine-location
wheel-base
length
width
height
curb-weight
engine-type
num-of-cylinders
engine-size
fuel-system
bore
stroke
compression-ratio
horsepower
peak-rpm
city-mpg
highway-mpg

PERFORMING THE FOLLOWING TASKS IS INVOLVED

1. Initially I Imported all the Libraries whatever I want
2. Reading the Dataset used by pandas
3. Then perform all required steps to analyze the data frame like shape, describe, info, etc.,
4. In this dataset data points are having both numerical and categorical values, so I performed encoding the categorical values also.

5. Some of the pre-processing techniques and visualization were done on the dataset.
6. List of columns was dropped due to those having too many unique values. So I thought it's better to remove it for my model performance.
7. I used Heatmap to visualize the strength of correlation among the variables on numerical features only.
8. While using isnull() to find the appearance of null values. But it shows no null values are having. So as a Data Scientist, exploring myself to find the appearance of the null values. My intention becomes true. Yes, some null values like ?, -1, -2 are having the dataset. Hence, it might be encoded for our model's good performance. Here I used a simple Imputer for replacing the null values by using the strategy 'mean' on the "price" feature others are replaced by nan values. As I expected null values are now handled.
9. Now Declaring feature column and target in X and y respectively.
10. Split data into separate training and test set.
11. I will carry out feature engineering on different types of variables for transforming raw data into useful features that help us to understand our model better and increase its predictive power.
12. Dropped unrequired columns from the dataset.
13. This is a Regression dataset, so I build relevant models like Linear Regression, Ridge, Lasso, SVR, Decision Tree Regressor, Random ForestRegressor, KNeighbors Regressor. Amongst that Decision Tree Regressor gives the better score. Hence, I took the Decision Tree Regressor as the final model.
14. Some visualization is applied between the Train score and the Test score.
15. The following metrics were used in 'Final_model' on the test data (mean squared error, mean absolute error, R2 score).
16. Decision Tree Regressor was tunned by the hyperparameter libraries Gridsearch CV AND Randomized Search CV.

Conclusion

In this project 7 regression models using on car price dataset. These are LinearRegression, Ridge, Lasso, SVR, DecisionTreeRegressor, RandomForestRegressor, KNeighborsRegressor. Amongst that Decision Tree Regressor gives the better score.