Abstract

The data concerning the fuel consumption features of vehicles is collected from UCI Machine Learning Repository and analyzed for pattern recognition to predict the “miles per gallon” feature of the vehicles based on their discrete and continuous attributes. A bagged version of multivariate adaptive regression splines (MARS) model is used tuning its parameters to predict the “miles per gallon” variable. Predictions on the test set are obtained with RMSE of 3.4 for mean value of 23 for the outcome variable.

Aim

The aim of this project is to predict the mpg (miles per gallon) of the vehicles using the data collected from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Auto+MPG).

Introduction

This dataset is a slightly modified version of the dataset provided in the StatLib library in line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The original dataset is available in the file "auto-mpg.data-original". The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes.

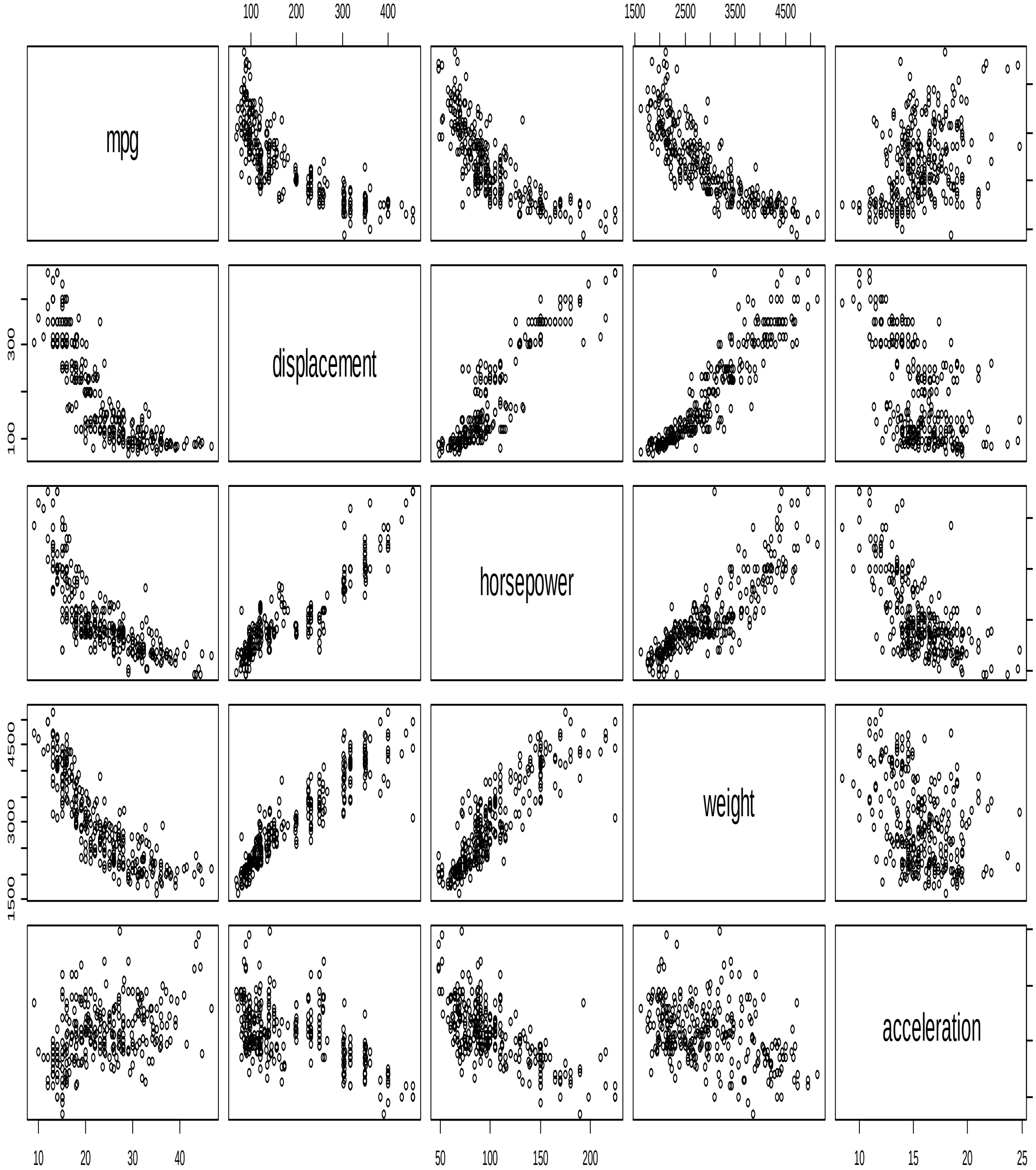
Summary Statistics

Table 1: Overview of numeric variables in the dataset

	n	mean	sd	median	min	max
mpg	276	23.65	7.89	22.75	9	46.6
displacement	276	191.17	101.61	146	68	455
horsepower	276	103.15	37.06	92	48	225
weight	276	2963.48	832.37	2831.5	1613	5140
acceleration	276	15.67	2.69	15.5	8.5	24.8

Exploratory data analysis

Figure 1: Pairwise relationships between variables


mpg
displacement
horsepower
weight
acceleration