AutoMPG Dataset Analysis

Atrij Talgery

github/progmatix21

Contents

1 Problem statement

Cars are around us for the past many decades. The skill and perfection of manufacturing cars improved over the years. There is a constant effort to improve the fuel economy of cars. Car makers often quote mileage under ideal conditions. However, car users are more interested in fuel economy under actual conditions. The data available to the car user community is usually accessible in the public domain. The problem and challenge is to predict fuel economy using such data. The autoMPG dataset attempts to take up this challenge.

2 Data description

Origin of data Statlib library from the Carnegie Mellon University https://archive.ics.uci.edu/ml/datasets/auto+mpg
Number of Records 398
Missing Values Yes

Year Donated to the repository 1993-07-07

2.1 Attributes and their meaning

1. mpg: continuous (target variable)
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

3 Analysis approach and Model used

We are going to use Orange which is a visual programming environment for data science.

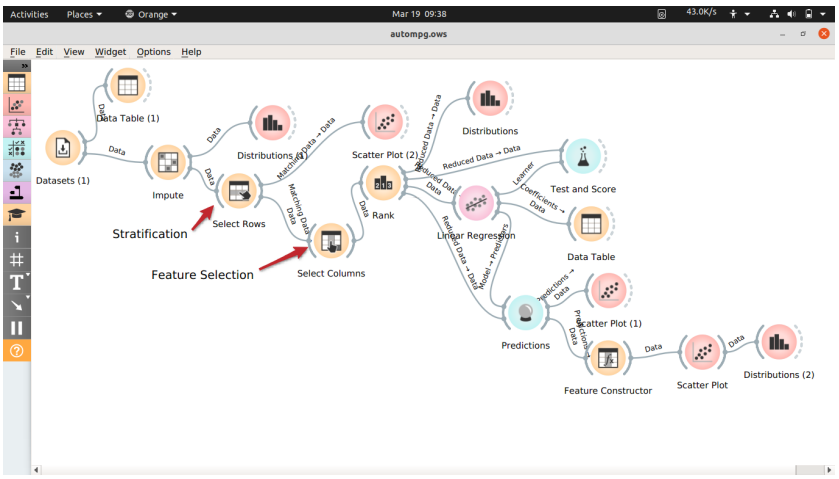

Figure 1: Screenshot of Orange Visual Programming Environment.

3.1 Data preparation

Imputation strategy There are missing values. Rows with missing values are dropped. This does not make much of a difference since we have a relatively large number of data points.

Feature scaling, encoding and normalization have not been done.

3.2 Exploratory Data Analysis

From the Figure 2 we observe that American cars have a distinctly lower mpg compared to European and Japanese cars. We also observe a move towards lighter cars in the later years of 1980 and beyond.

In Figure 3, American cars are numerically greater in the dataset compared to European and Japanese cars. The Figure 3 shows an overall uniform distribution across the years with exceptions in the years 1973, 1976 and 1978.

In Figure 4 we see that American cars are the heaviest and Japanese cars are the lightest. European cars lie somewhere in between.

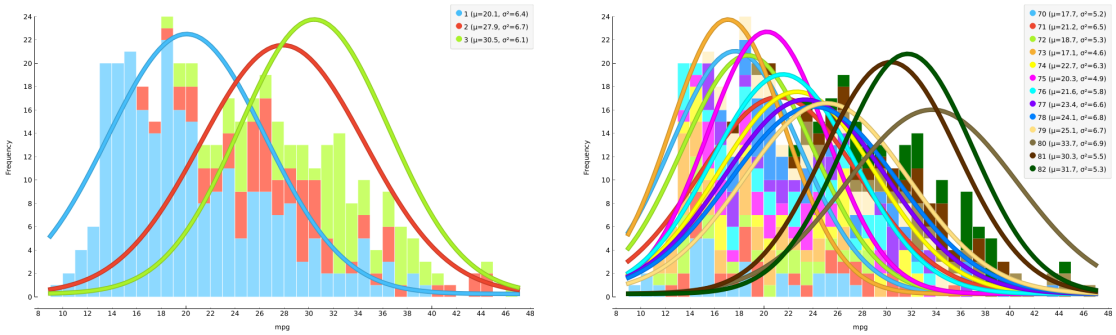

Figure 2: Distribution of target variable (mpg) by Origin.

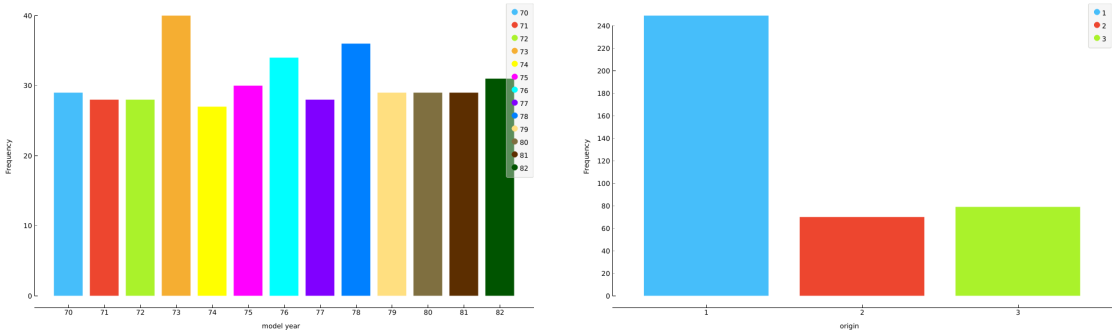

Figure 3: Model Year Histogram and Origin Histogram.

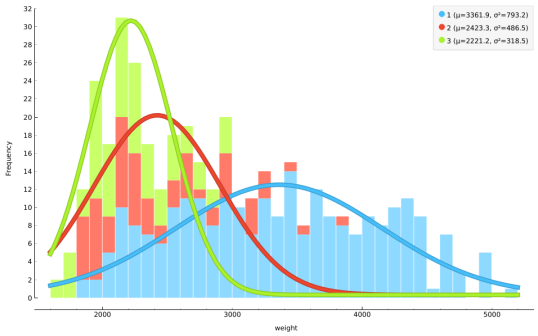

Figure 4: Weight by Origin.

We note that American cars have a different character with respect to the target variable. Also that later model cars are clearly lighter and more fuel efficient.

3.3 Model Used

We use a linear regression model for this data set as most of the predictors are continuous. The dataset source description also recommends this approach.

Since the target variable shows distinct distributions, along two categorical variables: `origin` and `model year`, we will need to take this into account while modelling.

Feature ranks

Feature ranks tell us the most important features with respect to the target variable. Here, we find that `weight`, `horsepower` and `displacement` have a greater influence on the target variable as seen from Figure 5. Indeed, these will be the features we choose

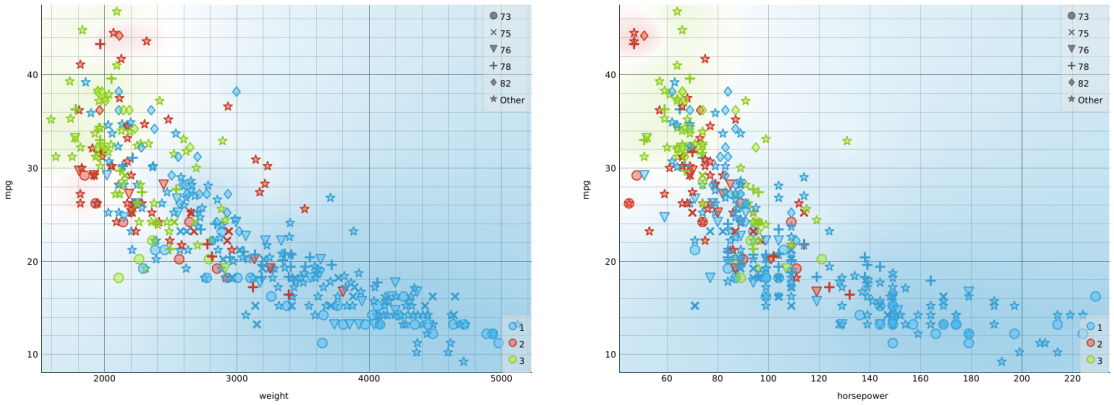

Figure 5: mpg vs Weight and mpg vs Horsepower.

for analysis.

Confounding variables

The features `origin` and `model year` are clearly our confounding variables as they are not among the top-ranked features for analysis.

3.4 Modelling Approaches

Approach 1: Feature Ranking, Stratification, No regularisation We build the regression model using our top-ranked features are predictors and stratify our data set along the two confounding variables: `origin` and `model year`. We also run a regression without stratification. This gives us five scenarios for regression.

Approach 2: L1 (Lasso) Regularisation only This approach does not use stratification and lets us observe, the trend in the regression coefficients as degree of regularisation increases.

3.5 Cross validation approach

We use 10-fold cross validation. There is no separate test data set.

4 Results (Discussion)

4.1 Approach 1 (Feature Ranking, Stratification, No regularisation)

The table below gives the values of R^2 for the various stratifications with our top-ranked features predictors.

Table 1: Values of R^2 by stratification

Stratification	R^2
American Cars	0.728
Non-American Cars	0.435
Models Pre-1980	0.768
Models 1980 onwards	0.502
All cars	0.700

4.2 Approach 2: L1 (Lasso) Regularisation only

We consider the entire dataset without stratification. We observe that without regularisation we get a R^2 value of `0.85`. This optimistic figure is definitely a sign of overfitting. This can be confirmed from the values of coefficients for relatively insignificant features such as `model year`.

As we increase the weightage of L1 regularisation(λ) we observe that relatively less relevant features drop off. A λ value of `0.5` gives us a model with our top-ranked features as the predictors and an R^2 value of `0.7`. This value agrees with the corresponding value we got in the first approach.

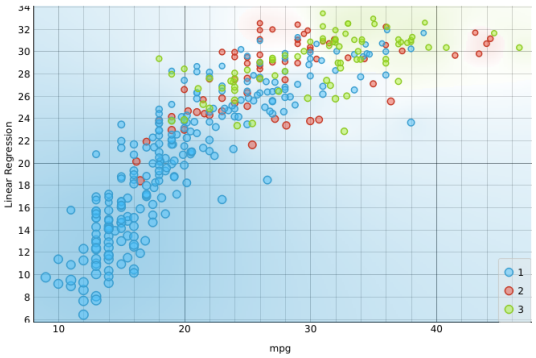

Figure 6: mpg predicted vs mpg true.

4.3 Analysis of Residuals

Figure 7 shows the scatter and distribution plots of residuals by origin. They are centered mostly around 0.

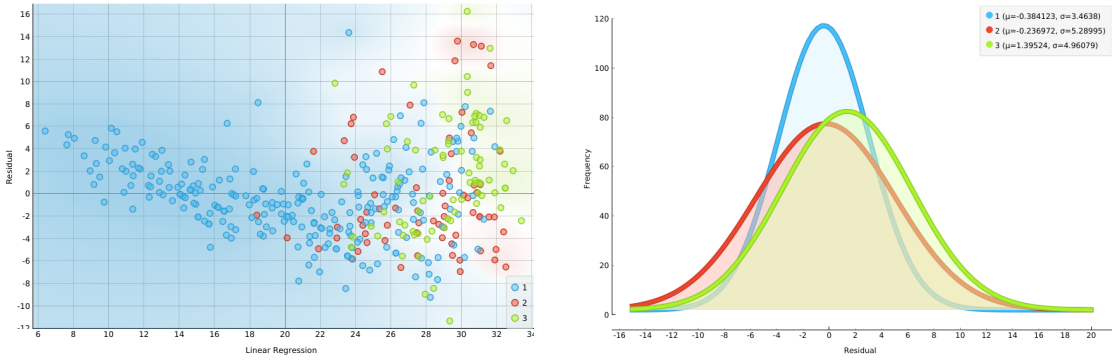

Figure 7: Residuals: Scatterplot and distribution by Origin.

5 Conclusion and caveat

Weight, Horsepower and Displacement are the most important predictors for the target variable `mpg` across all makes and all years.

Looking at the stratified results, we get a better fitting model for American cars and cars pre-1980 than we get for non-American(European and Japanese) and cars of 1980 and later.

It is quite intuitive that the heavier weight of American cars contributes to their lower mpg.

However, the lower R^2 value for non-American cars indicates that there must be other predictors for Japanese and European cars that are not captured in the dataset.

As an aside, from Figure 10 it is evident that increasing horsepower of American cars was contributing to a lower acceleration! Perhaps the higher horsepower engines were used to haul the heavier American cars.

Caveat: There are more American cars in the dataset than there are European and Japanese cars. This would have an effect on the result where ever the stratification is not based on origin.

6 Appendix

6.1 Interesting Observations about the dataset features.

Figure 8 shows cars of 1980 and beyond show a distinct move towards lighter weight and better mpg.

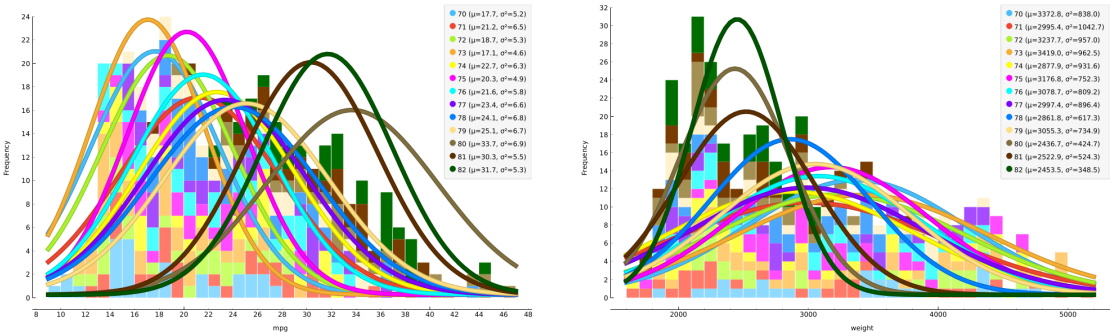

Figure 8: mpg by Model Year and Weight by Model Year.

From Figure 9 we see that American cars have higher displacement, higher horsepower, lower mpg and yet offer the same acceleration. Japanese and European cars offer the same acceleration for lower displacement, lower horsepower and higher mpg.

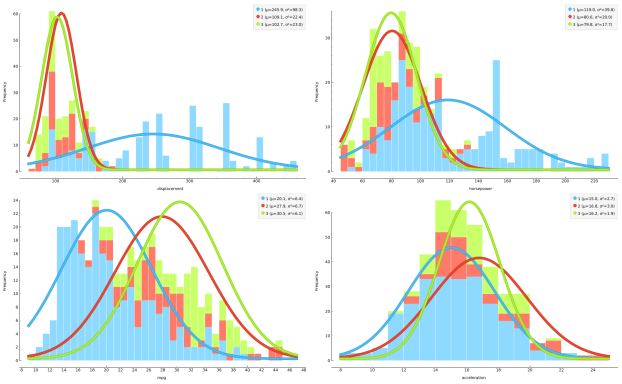

Figure 9: Clockwise from top left, Displacement by origin, Horsepower by origin, MPG by origin, Acceleration by origin

In Figure 10 American cars show a lower acceleration with increasing horsepower.

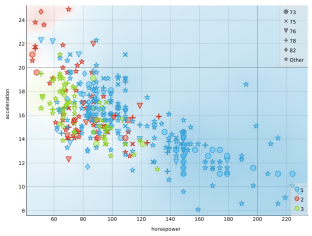

Figure 10: Acceleration vs Horsepower