Car90 Dataset Assessment Summary

James CarMichael

Assumptions taken for the consumer profile

1. The average net compensation for a single person in 1990 was approximately $20,172.11 (pre-tax) according to the Social Security Administration. We'll assume this remains the same over the course of the loan, excluding taxes.
    a. The buyer would be abiding by the following assumptions:
        i. Our consumer does not want to spend more than 15 percent of his/her gross income on the car loan over the four-year period.
        ii. The term for the loan is 48 months.
        iii. Our consumer will put zero percent down on the car.
    b. For simplicity sake, there will be a four percent interest rate.
    c. With these parameters, a loan amount of $12,103.27 serves as an example of a categorically "green" purchase.
        i. Total cost of the car loan with interest would be $13,117.
        ii. Monthly payments would be $273.
2. The buyer would be the average height of an adult male measuring at 5' 9''.
3. The buyer doesn't have any children, so space in the back is not crucial.
4. The buyer lives or often drives in a metropolitan area, so size and maneuverability are valued over power.
    a. An average width of North American parking space is 102.36 inches, while a compact space within city limits is 90.55 inches. Our consumer would like his/her car to be compatible with compact parking spots. This metric is not necessary since all of the cars' widths are below 78 inches.

Classification parameters

5. I broke down the categorization process into three elements: affordability, performance and size.

Category	2 Points	1 Point	0 Points
Affordability	Purchase amount at 15% or less of gross income	Purchase amount between 15% and 20% of gross income	Purchase amount above 20% of gross income or no price listed
Performance	Normalized data less than the 25th percentile	Normalized data between the 25th – 50th percentile	Normalized data greater than the 50th percentile
Size	Normalized data less than the 25th percentile	Normalized data between the 25th – 50th percentile	Normalized data greater than the 50th percentile

6. Final classifications were determined from the sum of the categorical scores shown above.
    a. ***Green - 25***
        i. Aggregate score of 5+
    b. ***Yellow - 38***
        i. Aggregate score of 1-4
    c. ***Red - 48***

    i. Aggregate score of 1-0

7. Unweighted, statistical significance of categories were defined as follows:
- ***a.*** Price is the first factor to consider when deciding to buy a product. These categorical scores were derived from the percentage of total income over the course of a four-year loan. Cars without prices listed were given 0 points in this category.
- ***b.*** Since the consumer is an average city-dweller, he/she will more than likely not need to prioritize power or speed. I identified both horsepower and engine displacement to have a relatively positive correlation with the car's price. From which, I decided to match a higher categorical score to a lower aggregate, normalized score of these two metrics.
- ***c.*** Keeping in mind the parking situations in the city, our consumer would prioritize a smaller vehicle to fit into parking spaces easier and maneuver through traffic better. As such, I noticed that both length and tank size sustain a positive correlation with price and categorized a lower aggregated, normalized score of these two metrics with a higher categorical score.

What obstacles did I have to overcome?

The amount of missing values was one of the largest challenges to overcome. More specifically, the NA values within key metrics like mileage, price, reliability and gear ratio that I thought would be helpful variables within the analysis. At the same time, establishing the parameters for the categories was challenging with the amount of variables provided. Lastly, figuring out some of the variable definitions and how they fit into the overall profile of the vehicle was a challenge.

What other information would you like to have had to aid your analysis?

Other than retrieving NA values for the existing data, I would have liked to expand the data set by pulling in used car and/or more cars' information to create a much larger dataframe. An interesting idea would be to pull in traffic data, parking space sizing and traffic violations data from a metropolitan database from the period.

Beyond bringing in more data, one of the most interesting variables was the 'reliability' metric. I wanted to quantify how reliability was being measured within the dataset, fishing to see if there was any correlation between max engine speed (or any other performance variable) with reliability. I was interested in training a random forest classifier to see if I could further establish how these categories were being generated. From which, there was not enough observations (cars) for the classifier to discern the importance of each element in relation to the reliability category.