CAR PRICE ANALYSIS

A Chinese automobile company, Geely Auto, aims to enter the U.S. market by setting up a production facility there to manufacture vehicles locally and compete with U.S. and European counterparts. They have hired an automotive consulting firm to understand the factors influencing car prices, particularly in the American market, as these may differ significantly from those in the European market.

Our objective was similar to that of a consulting firm: we cleaned the dataset to minimize duplicates and selected the quantitative variables that were of interest. All variables used are in American units.

1) Linear Regression

First, we scaled the dataset because some variables, such as price, had significantly different magnitudes compared to others. We then conducted linear regression using the following independent variables:

- Horsepower
- Car length
- Compression ratio
- Engine size

The choice of these variables was empirical and based on intuition.

We observed that horsepower and engine size were particularly significant with respect to the dependent variable, price. Additionally, the VIF values were generally below 5, indicating no excessive collinearity among variables, and both R-Squared and Adjusted R-Squared values were around 0.8, indicating a strong model that explains much of the theoretical variance. The p-value from the F-test was very low, so we rejected the null hypothesis that all betas are zero.

A bidimensional graph showing the linear dependency reveals that all variables, except for the compression ratio, are linearly distributed with price.

What is the Compression Ratio? What Does It Depend On?

The compression ratio is the ratio between the cylinder volume when the piston is at the bottom dead center in the intake phase and the cylinder volume when the piston is at the bottom dead center in the compression phase. This ratio is always greater than 1 and, in fact, depends not on the car's price but on the type of fuel it uses.

Gasoline cars have a compression ratio between 0 and 9, while diesel cars have a compression ratio greater than 20. This is because diesel engines lack spark plugs and require a higher compression ratio to increase the temperature inside the cylinder, allowing diesel fuel to ignite through compression.

2) Cluster Analysis

Cluster analysis further strengthened our hypothesis that price could be a good independent variable.

In hierarchical clustering, we used the following variables:

- Wheelbase

- Car length

- Horsepower

- City MPG

- Price

- Curb weight

Using the elbow method, we identified four clusters in hierarchical clustering, which were relatively homogeneous in size. One cluster, smaller in size, stood out as it consisted of more expensive and prestigious cars.

In non-hierarchical clustering, using the K-means method, we chose to create three groups. The rightmost group (the smallest) comprises cars with higher prices, heavier weight, higher fuel consumption, and greater horsepower. This pattern was also observed in the PCA.

We then applied the PAM (Partitioning Around Medoids) algorithm to verify our analysis. This algorithm selects the optimal clustering method, and we evaluated the silhouette for group numbers ranging from 2 to 6. The best result was obtained with three groups. We noted that the clustering achieved with the PAM algorithm closely resembled the one obtained with K-means.

3) PCA

Through PCA, we observed that variables like "curb weight," "price," and "city MPG" were correlated with the first principal component. This makes sense because, moving from left to right, price, weight, and length increase, and fuel consumption also rises. The “city MPG” variable represents the miles per gallon the car can travel, and it is inversely correlated, suggesting that as price increases, fuel consumption also rises.

Therefore, we can interpret the first principal component as the "prestige" or "luxury" of the vehicle.

The "wheelbase" variable plays a significant role in explaining the second principal component. Wheelbase represents the distance between the front wheel's axis and the rear wheel's axis on the same side. Unlike pure car length, wheelbase is not just a length measurement; a shorter wheelbase is associated with greater maneuverability and better weight transfer, while a longer wheelbase provides greater stability during turns and, above all, more interior space.

As a result, the wheelbase can explain the second principal component, creating a new variable that can be interpreted as how well-suited a vehicle is to be a utility vehicle.

CONCLUSIONS

In conclusion, the factors influencing the price of American cars are undoubtedly cabin size, car size, power, and fuel consumption. Geely Auto should focus on producing large vehicles with efficient fuel consumption to compete locally and cater to U.S. consumers' preferences.

Alberto Sartini

Leonardo Galassi