What drives the price of a car?

By Delil Martinez

Link to notebook: Practical-Application-2-Mod11/assignmentmod11_DelilMartinez.ipynb at main · delilx/Practical-Application-2-Mod11

Executive Summary

The task at hand in this project was to make use of a database with information about used vehicles to gain understanding of the elements that affect their prices.

Initial **observations**

- The set contains valuable information that can sensibly be considered to be a factor in the market's determination of prices for used vehicles, such as type of vehicle, manufacturer, model, year, and condition, to name a few. The total number of columns in the set is 18 (including an 'id' column that identifies the record).
- The size of the dataset is quite overwhelming, with almost half a million rows; however, there is a very large number of missing values in the set, reducing the usable sample significantly.

Preprocessing

Notable steps taken in order to proceed with the analysis:

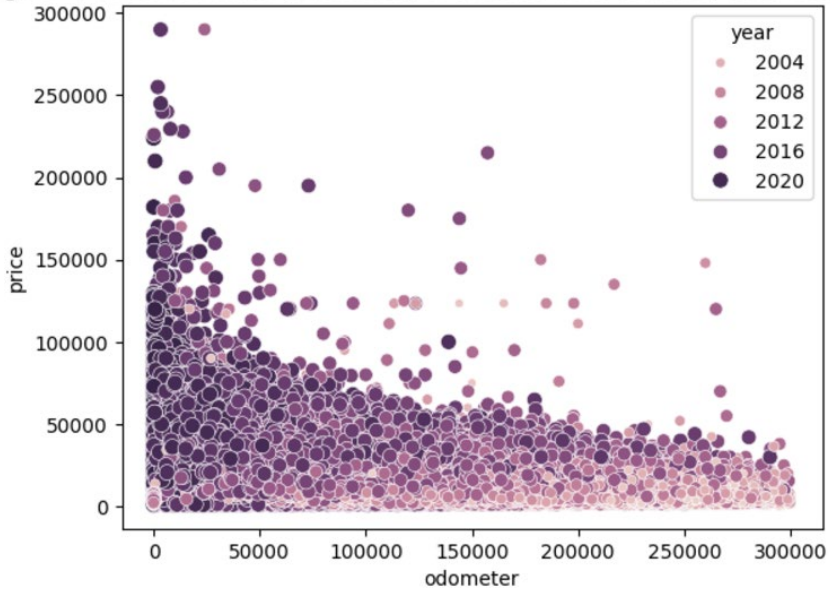
- Attention was restricted to vehicles manufactured after year 2000, as anything older is considered a classic/vintage/collectible vehicle, which constitute a separate category and require a separate analysis.
- Duplicates were eliminated from the set (using information on vehicles' VIN where available, or looking for exact matches otherwise).
- Samples with reported prices of less than $1000 or considered in "salvage" condition were also eliminated, as were samples with prices over $300,000 (assuming these are true values, vehicles with such prices should be analyzed separately).
- Among the categorical features in the set, the type of vehicle turned out to be informative of vehicles' prices, so samples where this entry was missing were also eliminated.
- A transformation was applied to the prices to adjust for skewness. This is accounted for in all the resulting models.

The resulting set contained over 140,000 observations, a healthy size to train and test various models.

Feature Selection

While all the features included in the set can plausibly contribute to the determination of used vehicles' prices, my intention was to keep the models under consideration parsimonious and, ideally, interpretable. Initial exploration of the data suggested that ideal features to make use of are the vehicles' mileage (referred to as 'odometer') and the year of manufacturing, as suggested by the matrix of linear correlation coefficients

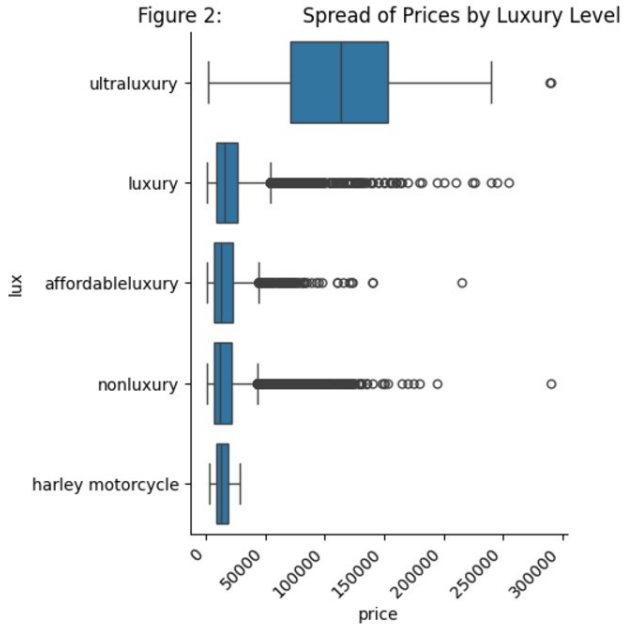
Figure 1: Recorded miles (odometer) vs Price of used vehicles, color-coded by year



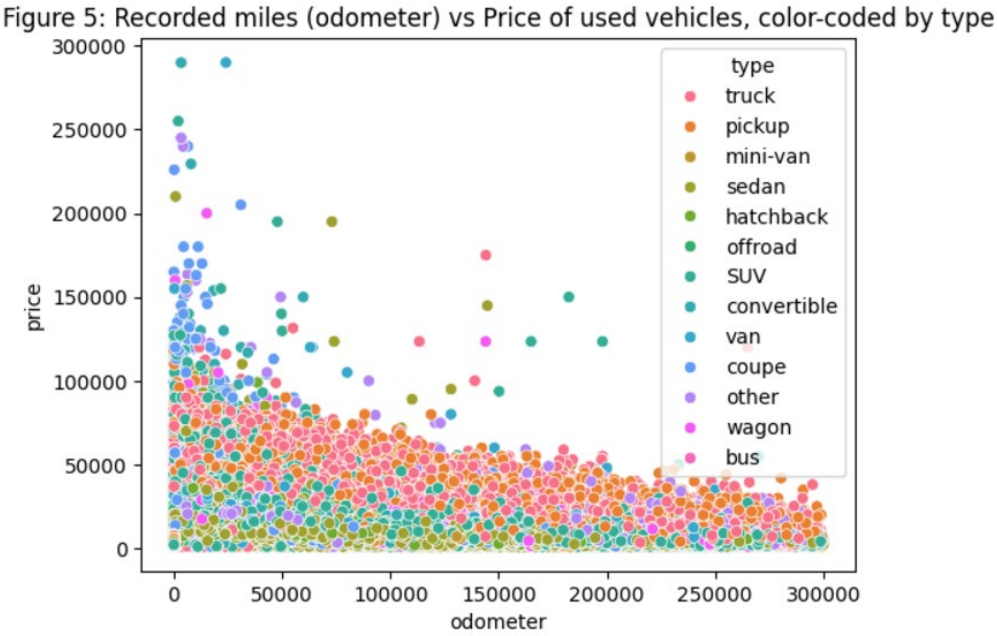
In terms of categorical features, it seemed intuitive to explore the effects of the brand name ('manufacturer'), though this variable has too many possible values to create a clear picture. Thus, an artificial variable called 'lux' was created to classify the manufacturer names by luxury level. This classification of manufacturers by luxury level was based on the information found in https://autotribute.com/luxury-car-brands/:

	Luxury level			
	ultraluxury	**luxury**	**affordable luxury**	**nonluxury**
Manufacturers included in the set	Aston-Martin, Ferrari, Lamborghini, McLaren, Bugatti, Bentley, Rolls-Royce, Maybach	Mercedes-Benz, BMW. Audi, Porsche, Alfa-Romeo, Jaguar, Land Rover, Tesla. Mini, Cadillac, Rivian, Maserati, Genesis, Volvo, Polestar, Rover, Morgan	Lincoln, Buick, Acura, Infiniti, Lexus	Ford, Chevrolet, Jeep, GMC, RAM, Dodge, Chrysler, Pontiac, Mercury, Saturn, Volkswagen, Fiat, Toyota, Honda, Nissan, Hyundai, Subaru, Kia, Mazda, Mitsubishi, Datsun

Here is a visual presentation of the differentiation in prices according to luxury level:



Aditionally, the type of vehicle was also identified as a potentially important contributor in determining its price, as suggested in the following graph:



Results

After fitting a total of 15 models with different configurations of the variables described above, the one considered optimal is the one that minimized the mean squared error (MSE), a measure

of average discrepancy between the model's predicted prices and the actual prices in the test portion of the dataset.
This optimal model takes into account the two quantitative features (the mileage recorded in the vehicle's odometer and the vehicle's age) in a strictly linear way, and the two categorical features (the type of vehicle and the manufacturer's luxury level) through auxiliary 'dummy' variables.

The mathematical structure of the model is a linear regression with a somewhat complex formulation, given by

$$\ln(\text{price}) = 11.319158 - 0.000003 \cdot \text{odometer} - 0.85702 \cdot \text{age} + \beta_{\text{lux}} + \beta_{\text{type}},$$

Where β_{lux} and β_{type} represent fitted coefficients corresponding to the luxury level of the vehicle's manufacturer, and to the vehicle's type, respectively.

This model, once converted to the original scale, suggests a baseline price of $e^{11.3192} =$ \$82,384.91 for a hypothetical new vehicle (age = 0 and miles in odometer = 0). The model then uses multiplicative factors to adjust the price in the following way:

- For every 1000 miles in the odometer reading the price is scaled down by a factor of $e^{-0.003} = 0.997$ –this can be interpreted as vehicles lose about 0.3% of their value with every thousand miles traveled.
- For each additional year of age of the vehicle, the price is scaled down by a factor of $e^{-0.0857} = 0.9179$ –the vehicle's value decreases by about 8.21% each year.
- Depending on the manufacturer's luxury level and type of vehicle, the price is scaled up or down by the factors given in the following table:

Factors for scaling price according to lux level and type

		LUX (β_{lux})				
		ultraluxury	luxury	affordable luxury	nonluxury	harley motorcycle
TYPE	(β_{type})	1.249662	-0.124267	-0.190621	-0.456046	-0.478728
truck	0.495536	5.72703532	1.44957296	1.35650969	1.0402801	1.01695005
pickup	0.469433	5.57947675	1.41222433	1.32155887	1.01347701	0.99074806
offroad	0.457244	5.51188131	1.39511521	1.30554816	1.00119872	0.97874514
bus	0.316092	4.78628243	1.21145849	1.13368229	0.86939823	0.8499005
convertible	0.092798	3.82844992	0.969021	0.90680939	0.69541395	0.67981811
other	-0.01723	3.42955323	0.86805604	0.81232644	0.62295687	0.608986
van	-0.04524	3.33482765	0.84408	0.78988967	0.60575056	0.59216557
coupe	-0.09807	3.16324057	0.8006495	0.74924743	0.57458284	0.56169684
SUV	-0.11162	3.12066463	0.78987308	0.73916286	0.56684918	0.55413663
wagon	-0.27308	2.65536733	0.67210144	0.62895221	0.48233084	0.47151376
mini-van	-0.32465	2.52190609	0.63832099	0.59734049	0.45808845	0.44781504
sedan	-0.45705	2.20916143	0.55916202	0.52326356	0.40128034	0.39228094
hatchback	-0.50417	2.10747174	0.53342329	0.49917727	0.38280904	0.3742239

Blue values are factors greater than one –meaning the price is scaled up for those particular combinations of luxury level and type of vehicle, whereas red values are less than one –meaning those combinations of luxury level and vehicle type are associated with a scaling down of the price.

Actionable items: Vehicles to keep in inventory/Vehicles to avoid

The work above results in the following insights:

1. Obviously, newer vehicles fetch higher resale prices than older ones.
2. All ultraluxury brands of vehicles produce high selling prices. When possible, dealers should hold these in inventory as they tend to have high prices associated with them.
3. Trucks, pickups and offroad vehicles are good assets for dealers' inventories, as they are associated with high prices.
4. At the opposite end of the spectrum: hatchbacks, sedans, minivans, wagons and even SUVs tend to associate with lower prices, as suggested by the factors that are all less than one in the table.
5. Coupes, vans and convertibles, not surprisingly, are better if they are from luxury manufacturers than if they are not; the same is true for convertibles.
6. In hindsight, while "Harley-Davidson" was detected as a label in the manufacturer feature and not in type, it would have been better to code these samples as "motorcycle" type, overriding whatever entry was present in the set. This would have made the results of the analysis more sensible. Nonetheless, since Harley-Davidson is considered a luxury brand of motorcycles, we can presume that having them in a used car dealership would also be desirable.