

Proposal: Predicting Used Car Price

Yihao Zhu, Xinci Liu, Lirou Luo, Liming Zhao

October 4th, 2020

1 Background

The price of used cars are influenced by a lot of factors, including used year, mileage, its original price, its brand and model. Although there are a lot of car-dealing websites providing price for different types of used cars, these data seem complex and relatively independent. It is difficult to understand the relationship between these data, such as the impact of a specific factor on the price. In this project, we want to figure out what factors play an important role in determining the price of used-cars by providing clear and visualized data.

2 Question

What are the primary factors that influence the pricing of used cars in the U.S.? How can this information be used to the benefit of a customer?

3 Why Important?

The current pandemic is altering the used-car industry: transactions are shifted away from the dealership lots and largely digitized. Long-distance communication between buyers and sellers likely results in a reduced ability to communicate on details and negotiate on price. It becomes essential for prospect buyers to know how to best utilize the information provided online to judge the value of their prospect investments. Buyers should understand how each feature is factored into the pricing of a used car, and whether their targeting vehicles possess features that would enable them to carry values into the foreseeing future. With these information, prospect car buyers can then evaluate dealer's offers and make better buying decisions.

4 Dataset Description

We plan to use datasets for used-car information on Kaggle: <https://www.kaggle.com/jpayne/852k-used-car-listings> which provides general information on 1.2 million used cars, such as price, year of manufacturing, mileage, location and model. Considering that the data only provide the price for used-cars in 2017, the current prices of those cars are definitely lower than the prices in dataset. However, assume that the Manufacturer's suggested retail price (MSRP) of a same model basically keeps consistent from year to year, the results we obtain from the dataset in 2017 can be applied to the dataset gathered this year. That is, it is the usage year rather than the register year that dictates the price. Therefore, we first evaluate the price of used cars from that year. Then, we collect new data by scraping from <https://www.car.com/> and <https://www.truecar.com/> to validate our results.