Predicting Used Car Prices Using Craigslist Data

Keshav Mantha

Problem Statement: Can we accurately predict a used car's resale value using existing data of used car sales?

Context: Craigslist is one of the internet's oldest resale listing websites and marketplaces. The site has a plethora of listings around the country for used items put on sale. The car resale (ie. used car sales) industry is worth nearly $100 billion in the US alone. There are nearly 300 million registered vehicles in the country, and nearly 250 million licensed drivers. Drivers that sell or trade their vehicles to a dealership or wholesaler lose at minimum 11% of value, which represents thousands of dollars of further depreciated value to the car owner. By being able to predict the resale value of any given car, we can restore some power to both the car owner and the car buyer by ensuring everyone can get a fair price for their used car.

Criteria for Success: If we can achieve a strong fit between our train and test data by splitting up the car-search results from Craigslist, then we can predict the resale price for cars that are "currently" used that someone would want to sell.

Scope of Solution: These prediction models could be used by the every-day car used car buyer to predict what the price of their vehicle might be in a given number of years. It could also be used as a validation tool against currently existing used-car market benchmarks, like the Kelly Blue Book.

Constraints: Our dataset comes from Craigslist, so it is a public repository of listings, meaning our it would be limited to listings by individuals without deep knowledge of car sales. The dataset also does not seem to indicate the price cars were sold for (if at all).

Stakeholders: Any individuals either looking to buy/sell used cars, or any companies involved in the vehicle reselling business.

Notable Features: Year, make, and model of the vehicle, as well as listing price, are some of the most important features in this data. Condition and price are also important considerations. Based on the vastness of our data, and the numerous factors which can influence resale value, we can predict that a supervised RandomForest regression machine learning model may fit this data in helping us form our future predictions.

Data Sources: Craigslist data on used car sales from kaggle:
https://www.kaggle.com/austinreese/craigslist-carstrucks-data