Taylor Imhof

Bellevue University | DSC 680

Milestone #1: Project Proposal and Data Selection

10/17/2022

TOPIC

Using historical used car sales data to help prospective buyers negotiate their best price.

BUSINESS PROBLEM

Shelling out the excessive amount of dollar bills required to purchase a new vehicle simply is not realistic for most of the population. Instead, most people will purchase a used vehicle at a much more manageable price point. Several factors contribute to the value of a used car, but perhaps there is a way to help prospective buyers know what they are getting into. Moreover, many shady used-car dealerships likely attempt to take advantage of non-mechanically inclined customers. Such establishments will attempt to push their inventory at much higher prices than they would negotiate. The primary purpose of this project is to train a predictive model that will assist people in the market for a used car to better understand the prices they should haggle for.

DATASETS

The main dataset that will be used to train and validate my predictive model is the "Automobile Data Set" retrieved from UCI's Machine Learning Repository. The dataset's landing page details the background information. Some key highlights include the following:

- Source of the data
    - 1985 Model Car and Truck Specifications
    - Personal Auto Manuals
    - Insurance Collision Report
- An explanation of the actuarial term "symboling."
    - Metrics used in industry to determine a vehicle's risk-factor
- Attribute information
    - The attribute name, as well as its associated range

The data folder also contains a useful <.names> file that further outlines key characteristics of the dataset. Apart from the info on the splash page, the <.names> file also specifies which columns contain missing data and how they are encoded using the "?" character, which is useful for data cleansing purposes. After reviewing the data and its features, I feel this is an excellent starting point for exploring the underlying data trends and attempting to train a predictive model that could be used in a practical car-buying scenario.

METHODS

The first step in any machine learning endeavor is exploring the data. Once I have loaded the data into the desired format (e.g., a Pandas DataFrame), I will begin by performing fundamental exploratory data analysis operations. As mentioned earlier, missing values will need to be handled, and there is likely to be some data that is not in the desired format. After getting the data in the proper format, I will perform any pre-processing steps as necessary before training the model. For example, many features are numeric and are measured on vastly different scales based on the ranges provided. As such, I will likely need to standardize the data so that the model learns more accurately. I will also perform correlation analysis to ascertain which features appear to have more predictive power on the target (price). This step is also useful for gauging whether features express multi-collinearity, which is important to consider when selecting predictive features.

Once the data has been processed, I will start trying to fit a model to the data. As the price target is numerical, I believe a simple linear regression model will serve as an excellent starting point for this project. Once I have trained and validated the model, I will likely better understand the data. Then, I might select a different modeling approach or iterate back to engineer more predictive features.

ETHICAL CONSIDERATIONS

When thinking about how this data could be used for the "bad," I could not help but think about how the predicted prices might influence the used-car market and push out smaller sellers that cannot increase their prices to meet the new "acceptable price" from the model. The driving impetus of the project was to curtail the bad-actor sellers' market hold by driving down the prices or at least encouraging buyers to haggle more when purchasing a used vehicle. It would be helpful if the dataset included a feature that measured the seller's business environment, perhaps average cars sold per day/quarter, to train the model more effectively.

There does not appear to be any location information, but this would also be useful information to explore using predictive analytics. If such information were provided, it would be important to consider each location's respective cost of living. For example, it would be unfaithful to gauge how much a used car sells in Beverly Hills compared with Beemer, Nebraska (and yes, I just Google'd "remote towns in Nebraska" 😀).

CHALLENGES/ISSUES

After reviewing the dataset, one of the larger problems that I feel like I will encounter is not having enough data. There are only 205 observations, and while the features are quite robust, this might not be sufficient data to generalize to real-world inputs. Another issue I might encounter is how to incorporate the new "symboling" feature in my analysis. "Risky-ness" seems a little nebulous to me, as I do not have the actuarial background of those who designed such a metric. Therefore, I might have to do additional research to better understand how this feature might impact the target.

Lastly, one challenge I will likely have to consider is that this data appears to be relatively old according to the provided source information. As such, perhaps I could develop a method of translating the price points to values comparable with today's market.

REFERENCES

Automobile Data Set. (1987). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. Retrieved October 17, 2022, from https://archive.ics.uci.edu/ml/datasets/automobile