VANDERBILT | M.S. DATA SCIENCE

Shea Conaway
shea.conaway@vanderbilt.edu

Draft v1 Capstone Project Proposal
Submitted for comment

Please answer the following questions to help guide you in scoping out your Capstone project. Each answer should be about a paragraph for questions 3-7.

1. Who are you working with on this problem?
    a. NA

2. Will you need to meet with me during the semester?
    a. Yes

3. Describe the problem you are solving. Be sure to include why this problem is unique or novel. If you are working in a research group or team, be sure to detail your precise contribution in relation to what the whole group is doing.

    a. The primary goal of my project is to build a predictive pricing model for used vehicles on the individual consumer market. This model could be monetized as a service for either consumers or dealers to assess the market value of a potential purchase or sale.
        i. For a consumer, it's helpful to know what the going market rate is for a used vehicle for comparison shopping and negotiation preparation.
        ii. For a dealer, vehicles are acquired via individual purchase, trade-in, or auction. The ability to price a vehicle and project profit potential has direct value.
    b. Cars are one of the highest-priced single products exchanged in a consumer market. Used cars make up a significant portion of transactions. They are a difficult to price asset due to the highly variable nature of the market, inventories, consumer preferences, and the features and conditions of individual vehicles, among other factors.
    c. My problem is unique because I'll be using modern regression techniques on real sales data to predict market value in the used car market specific to individual vehicles incorporating factors mentioned above.
    d. While the primary goal is accuracy, I would like to apply a range of parametric and non-parametric models, highlighting the benefits and drawbacks of each. Parametric models may still have analytical value to business understanding and/or serve as a robustness check for non-parametric models.

4. Describe the data you need for your project. Be as detailed as you can be here – even including key column names you require is encouraged.

    a. The most readily accessible used vehicle pricing data is scraped from public listings. There are four different datasets on Kaggle.com that I could use (1 2 3 4). All of these datasets have variables in common that should make for good predictors on the price target.
        i. Likely predictive independent variables include miles, year, make, model, trim level, body type, and local market
        ii. The Kaggle datasets vary in number of observations/variables and quality. It is also possible to acquire alternative data via scraping directly (CarGurus / Craigslist) or contracting with a commercial auto data provider (MarketCheck.com / AutoDealerData.com). Further investigation is needed on which to use.
    b. The public listings data has the benefit of easy access but with a major drawback - it remains unknown what the ultimate transaction price was, or whether the vehicle sold at all. I have done additional research on how to address this problem.
        i. By speaking with a sales representative at Cross-Sell.com, I learned that certain state DMVs collect reliable transaction prices during vehicle registration. I'm scoping out the costs of purchasing the pre-packaged data from Cross-Sell.com. But I've also made contact with the Texas, Ohio, and Virginia DMVs to collect the data directly via FOIA requests.
    c. Importantly, I want my datasets to include VIN (a unique identifier for vehicles used by the automotive industry) so that I can pull additional data on each vehicle and join different sources.

5. Describe your approach or primary task. What are you going to do with the data above to answer the question above?

    a. Beyond the data acquisition, cross-referencing, preparation, and cleaning above, my primary task is accurately modeling market price for each vehicle. I would like to compare parametric (linear/non-linear regression) and non-parametric (tree-based) approaches for differences in performance and explainability.
    b. This modeling task will involve assessing variables of importance, additional feature engineering (e.g. market conditions, vehicle-specific local inventories, days-on-market), and model tuning/selection.
    c. Ultimately, I would like to produce an interface for pricing a vehicle based on information a consumer/dealer would normally have when considering a purchase.

6. Describe what you have done so far, along with an estimate of how much time you have invested in this project already.

a. I estimate roughly twelve hours spent on this project thus far. Most of this time has been spent on inspecting the available data and pursuing additional data acquisition.
b. For inspecting available data, I've loaded each of the Kaggle datasets listed to review their variables, completeness (i.e. null values), and coverage (e.g. geographic market, timeframe, sale types).
c. I've put considerable effort towards original data collection. I've spoken on the phone with representatives at MarketCheck.com and Cross-Sell.com to better understand their data products and what other data may be out there. Based on my conversations, I've reached out to state DMVs to get my own transaction records.

7. Describe what you think will be the biggest challenges you will face in executing this project. Identify 1-3 challenges.

    a. As should be apparent, I'm interested in getting high-quality reproducible data. Ideally this would come directly from DMVs or a commercial data provider. If these present insurmountable challenges, I'll have to fall back on the public Kaggle.com data or scraping, which has the downside of list price and arbitrary coverage. It may be possible to adjust list price towards transaction price by analyzing vehicles with price reductions over time (indicating initial prices too far above market value) and dealers with fixed price business models (CarMax and Carvana).
    b. Even assuming high-quality true transaction price data, the target will likely be difficult to predict given the available features. The market value of used cars is dependent on idiosyncratic characteristics, like condition, local inventories, trim, and optional equipment, which may not be in the data. The negotiating strengths of individual buyers and sellers may introduce additional variability. Lastly, the used car market has been disrupted recently by COVID and technology related supply-chain issues (e.g. changing car needs with remote work and electronic chip shortages).
    c. There are existing car-pricing products like Kelley Blue Book and Edmunds.com. I would like to achieve comparable performance and/or differentiate my offering in some way. It's not clear how to do that.