# Regression Models Application Strategy

Brayden Yates

University of the Cumberlands

MSDS-530: Fundamentals of Data Science

Dr. Jason Turner

10 February 2024

## Regression Models Application Strategy

Car owners looking to sell their current vehicle tend to have difficulty identifying an appropriate price for their vehicle. Services like Kelly Blue Book often offer inflated private-party values to consumers, leading to additional challenges in selling their vehicle. There are many variables in used vehicle sales that impact the overall value of a vehicle, and many of these are not accounted for in services like Kelly Blue Book. Using publicly available data, I intend to create a model utilizing linear regression to identify the value of a given vehicle based on a number of attributes.

Vital information to determining an appropriate listing price include the year, make, and model of the car, along with the number of miles driven, the type of fuel it requires the type of transmission, the type of sale, and the number of previous owners of a given vehicle. As a general rule, newer, low mileage vehicles tend to sell at a higher price than older, or high mileage vehicles. A newer vehicle often means a more modern aesthetic, alongside increased safety features and luxury additions. Make and model is critical in many different ways: a Ferrari is certainly going to be valued higher than a Daewoo! Beyond the extremes, Honda and Toyota vehicles tend to hold more resale value due to their reputation for long lifespans and lower maintenance requirements. An older, moderately-high mileage Toyota Camry is seen by frugal individuals as a “gold standard” vehicle. A higher mileage vehicle is more likely to require additional maintenance in order to be roadworthy—even in “gold standard” vehicles like the Camry and the Accord.

Fuel costs tend to play a lesser, but still significant role in vehicle pricing. The price of Diesel Fuel has increased dramatically since the Great Recession, leading to higher fuel costs. Diesel vehicles also tend to have a lower fuel efficiency, leading to further increased fuel costs. Dealers also tend to charge a higher rate for their vehicles as compared to private-party sellers. For those selling their vehicles, private-party sales tend to be more profitable than trade-in’s at the dealership. Additionally, vehicles with several owners, especially in a newer vehicle, tends to lower the price. This is most likely due to the fact that vehicles with several past owners are more likely to be "lemons" rather than reliable.

With regards to the strategy, data collection is essential. It is impossible to create a model of anything without data. Even those with some knowledge and context behind how different attributes affect vehicle prices would be unable to set an accurate price for a vehicle without any kind of baseline data regarding the vehicle's value!

After collecting the data, it is necessary to process the data. This includes deleting or estimating missing features in the dataset. A simple example of estimation could be this: A vehicle has the model 350z, but no make or year listed. Based on the information provided, we can know with certainty that this vehicle's make is a Nissan, and it's year must be between 2002 and 2008. If this same vehicle was missing a transmission type, this would be more of a challenge, as the 350z was more frequently sold as a manual transmission vehicle, but automatic transmission options were available and weren't necessary unpopular (autoevolution). This data must also be converted into a numerical format such as label encoding to ensure that scikit-learn or other linear regression tools are able to appropriately use the data.

It is also important to perform exploratory data analysis on the dataset to identify independent and dependent variables. This would likely incude scatter plots as a baseline to identify variables that are and are not related. While it makes sense that mileage and value are strongly correlated, the number of previous owners and price may not be so strongly correlated.

After performing some exploratory data analysis, it is necessary to choose the appropriate regerssion algorithm for the dataset. Linear regression is a simple but effective algorithm, however the popular python library *scikit-learn* offers decision trees and random forest algorithms, which could be more effective in building the model. Additionally, tensorflow and neural network libraries could be more effective for this particular dataset. After choosing a model algorithm, one must split the dataset into a training and testing set in order to evaluate how well the model performs. While this does decrease the amount of training data available, there is not any other way to ensure that your model is behaving appropriately with the data it's been given. The model must then be evaluated using metrics such as mean squared error and R-squared to ensure it is providing appropriate appropriate predictions for a given vehicle.

# References

*Nissan 350z*. autoevolution. (n.d.). https://www.autoevolution.com/nissan/350z/