Project Luther

Build Linear Regression Model to Learn about used Car Market

Anjali - October 2, 2018

Goals

- Develop a multiple linear regression model to understand the used car market
- Identify features both continuous and categorical that impact car prices
- Try to estimate car price by a set features
- Develop a supervised regression model which can learn
- Collect sufficient data to be able to train the model (at least 1000 points)
- Quantify how good the model is.

Data

Main source for data scraping - www.carmax.com
Supplement main data source with other data if time permits

Data Features

Continous	Categorical
Original Price	Brand
Year	Model
Miles	Color
Horsepower	Trim Level
Torque ?	Type - Sedan, SUV,…
Engine Size ?	Features - as sum of number or list ?
Fuel Economy (EPA MPG - City / Hwy)	
Wheelbase	

Continous	Categorical
Overall Length	
Width	
Curb Weight	
Seating Capacity	
Cargo Capacity	

? - may not be required

Original Price not available on most used car sites. Add extra data as part of second phase.

Target - Price for car

Min 1000 data points to start with. Look at cars from 2008 onwards only. Increase data points by scrapping for cars in areas near other cities.

Tools

Do web scrapping using Selenium. I don't think I will be using BeautifulSoup as of now.
Use Pandas, Pickling, Seaborn for handling data, storing and quickly visualizing it.
Use stats model, scikit-learn for developing regression model