YallaMotor

Web-scraping and Linear Regression

Abstract

Data from yallamotors.

Different countries.

New cars vs old cars

EDA on the data

Linear Regression

Design

The features in a car are many whether the car is new or used

The features can be mentioned in the car description, inferred, filled(based on similar) or simply not there

After studying the site it showed that the used cars had different features, as expected, but also different sellers mention/attach different features, of which some are missing...etc

Data

The data from the YallaMotor was scrapped using BeautifulSoup and Selenium.

The process of scrapping was as follows.

First we select a country and get the number of pages it has.

Then, starting with the first page, we get the list of cars.

By getting the list of cars here it means getting the links related to each of the cars in that list.

Once the page is done, it moves to the next page, gets the data, then moves… and so on. Once we have collected all the pages of country C

Data

Finally once we have collected all the data from all the site.

We run the links through the scrape function which extracts the data we want from each page. If the pre-defined feature exists in the page it sets as one if not then zero

The way it was done here is by extracting the data that we want and find. Whatever is missing is then filled with 0 to make all list of each feature the same

Data cleaning - Major

The data scraped had some major issues from padding, seeding, dtype...etc. The padding and seeding are caused by how the function was built which assumes -1 at each location to start the process

The data scraped had some major issues from padding, seeding, dtype...etc. The padding and seeding are caused by how the function was built which assumes -1 at each location to start the process length of each feature. If its not equal. Then its missing that feature as it was not detected. So a zero is placed in that location to fill the missing value making the length of the list equal.

Algorithms:

The required algorithm is Linear Regression. Which was done in two ways. The first on each collection of features vs price. Then based on correlation

Tools:

Numpy, Pandas for data manipulation | BeautifulSoup | Selenium | Firefox

Communication

In addition to the slides and visuals presented, the code is on github bushnag-ai and data is from these sites:

https://ksa.yallamotor.com

https://uae.yallamotor.com

https://kuwait.yallamotor.com

https://bahrain.yallamotor.com

https://qatar.yallamotor.com

https://egypt.yallamotor.com

The End