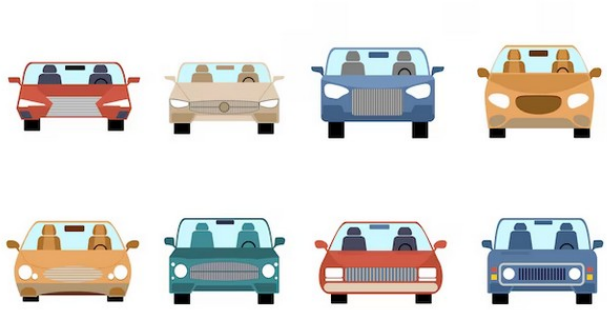
Automobiles price prediction project



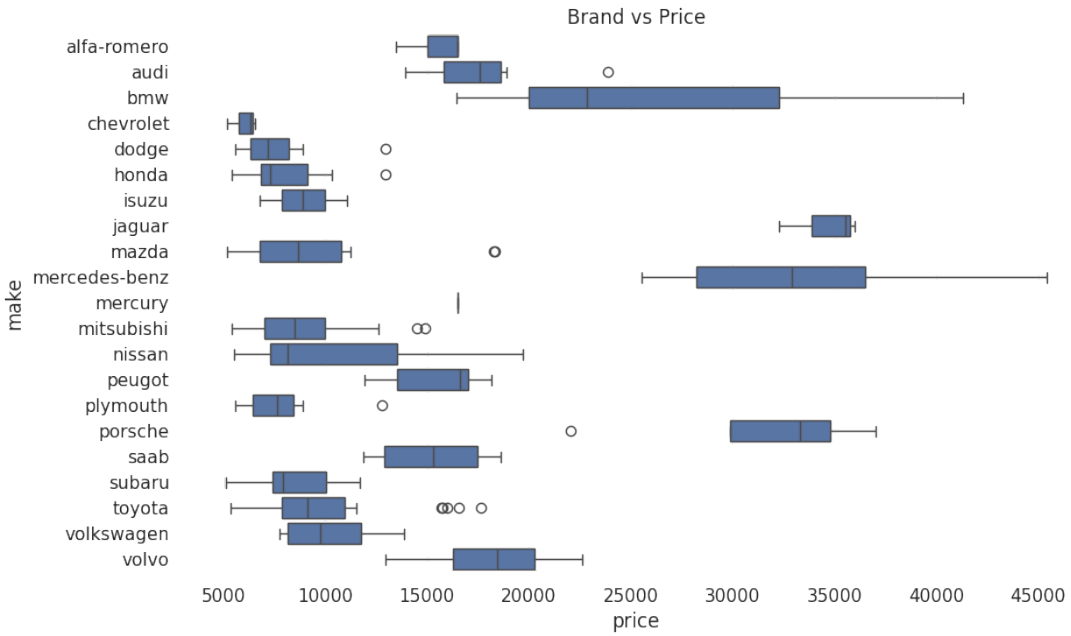
Here the methods used and the results squired from the project will be demonstrated.
In the notebooks along with the code one could find more detailed explanations about the steps and decision processes.

1. Data Cleaning Part

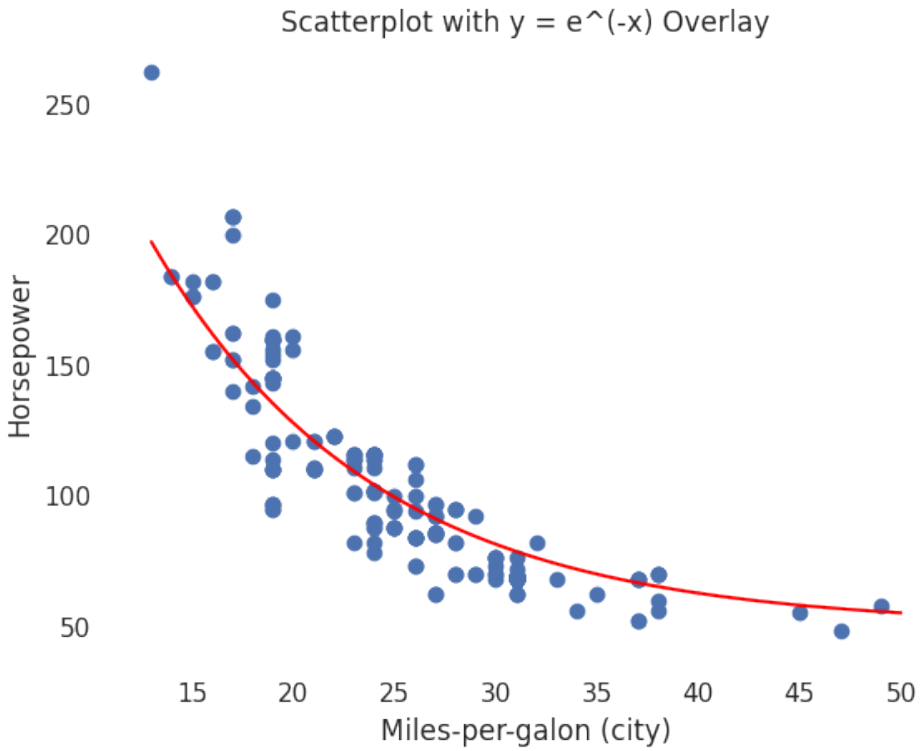
The data had substantial amount of missing values, they were replaced using the kNN algorithm, taking care to preserve the data distribution of the dataset. Few more methods were shown for comparison.
Also the datatypes of some features were not set properly, so this was fixed.

2. Exploratory Data Analysis (EDA) Part

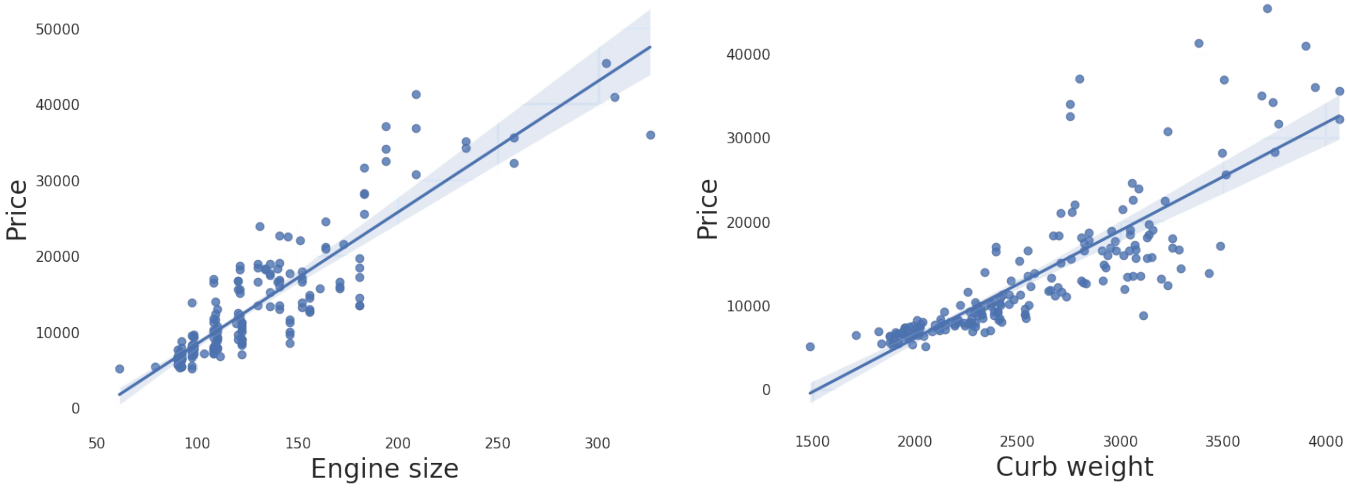
A number of interesting findings were made, some of them illustrated in the following 3 plots.



Here we can see the price ranges, means and quartiles of different brands.



Here we see an inverse exponential relationship b/w engine horsepower and the mileage/gallon.



The price seems to be linear with the size and power of the car (which themselves are correlated linearly).

Here is some takeout from the EDA:

* Station wagons and hatchbacks seem to be on average cheaper then all the other cars
* turbo engines tend to be slightly more expensive
* Rear wheel drives are often more expensive then the rest
* Ohcv engines are the most expensive on average and ohc and ohcf ones the cheapest on average
* Mpfi and idi seem tbe the most expensive fuel systems, while 2bbl and 1bbl the cheapest
* There is a positive correlation b/w the number of cylinders and the price of the vehicle

* city-mpg is proportional to highway-mpg
* curb-weight seem to be an excellent indicator, as it correlates strongly and linearly with the:
    * width
    * length
    * wheel-base
    * horse power
    * engine size
    * price of the vehicle (this on negatively)
* miles/gallon seems to be inversely proportional to the horsepower/engine size/weight
* heavier cars seem to have more rear-wheel-drives and turbo engines on average
* also heavier cars have bigger engines

* The most expensive brands are: jaguar, mercedes-benz, bmw and porsche
* The cheapest brands are: chevrolet, dodge and honda

Major conclusion:
* The bigger the car, the bigger engine it needs and the more likely it is to use more advanced and expensive engine with rear-wheel-drive. All of the above, plus the car brand label are major determinants of the price (and correlate positively-linearly with the price).

MCA was performed, to reveal commonalities b/w various features in various aspects:

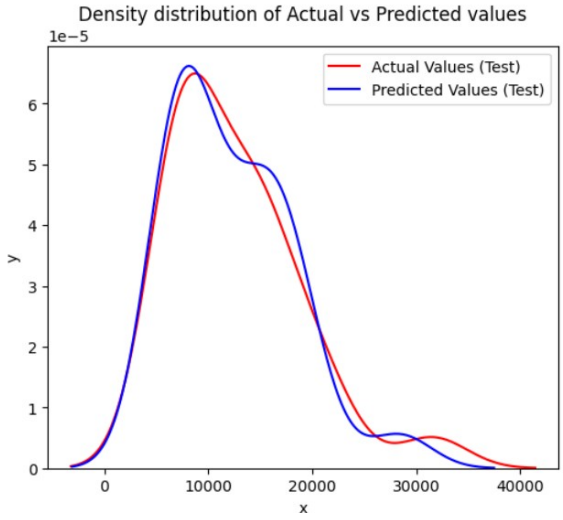
* Jaguars tend to have the biggest engines
* Porsche tend to put the engine in the back
* Convertibles and hardtops seem to have high Stroke
* Mercedes cars are correlated with high wheel base and big, heavy chassis
* Volvo and Peugeut seem to have mostly engine type I
* Bigger cars seem to have turbo engines
* Bigger bore correlates with lower horsepower
* Toyota seems to make more 4wd then the rest, they make light chassis and have low normalized_losses, also they have mostly economic 4 cylinder engines
* Two-door cars are stringly correlated with OHCF engines
* Honda seems to prefer the 1bbl and 2bbl fuel systems, they have good MPG and low width of the chassis
* Chevrolets seem to favour short chassis with 3 cylinder engines
* Isuzu and Mitsubisi have mostly SPFI fuel system
* Japanese cars tend to be on the samller size and more economic
* Mercury make very low RPM cars
* BMW seem to gave lower MPG and rear-wheel-drives
* MPFI fuel systems have high Bore

3. Machine learning part

The cleaned data, was prepared and all the non-numeric features one-hot encoded.
A random forest regression model was chosen to describe the data.
It was chosen as such, because decision trees are great at catching non-linearity in the data and because Random Forests can handle small datasets fairly well.

The model was hyper tuned with Bayesian parameter sweep (because its better then the rest).
Distributions of the predicted variable were compared and scores derived.

A score of 92.7% accuracy was achieved.

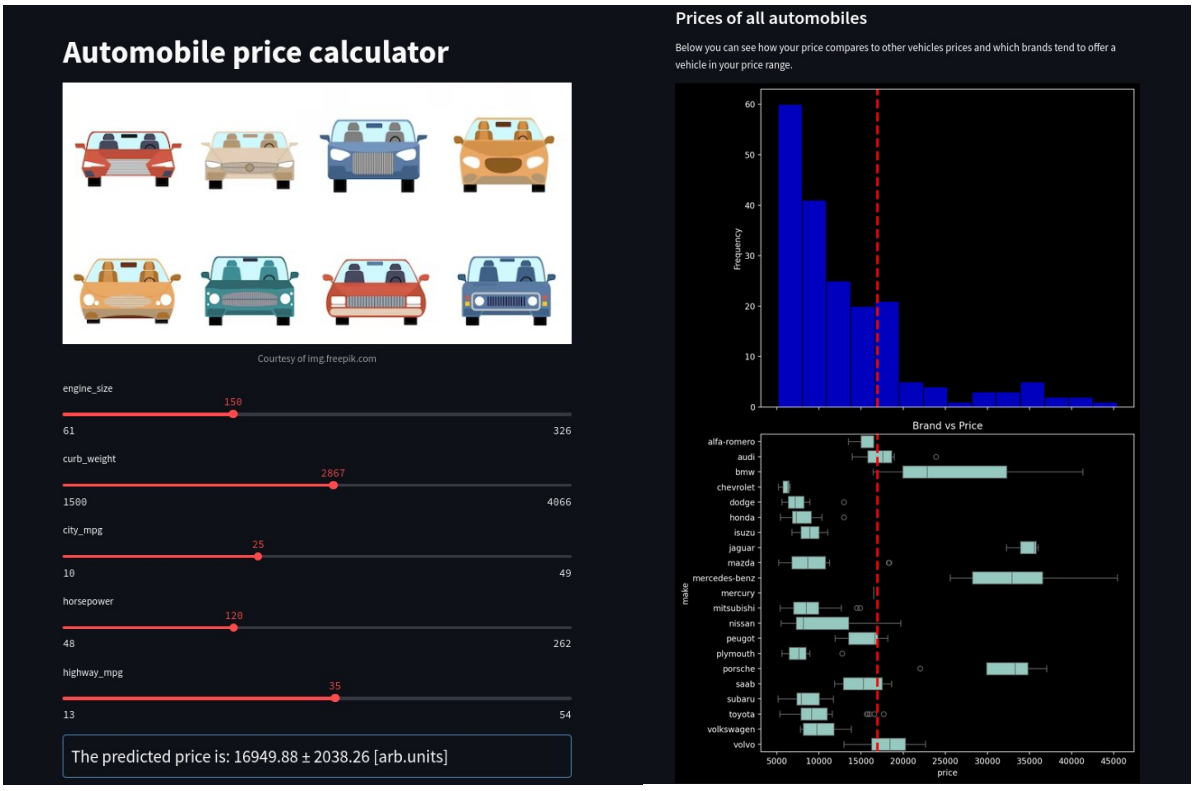


4. Streamlit app

An app was made, it consists of a calculator that takes in the most important parameters like: ['engine-size','curb-weight','city-mpg','horsepower','highway-mpg'].

And it outputs the hypothetical price of a car with such parameters.

Also it gives you information about how this price compares to the other prices and which brands tend to offer such priced vehicles.



5. High-level plan of action

To use the above constructed prototype as a backbend of B2B application, few general steps should be observed:

- a docker container will be created for the project
- a script that automatically takes new data and cleans and prepares it would be created
  - optionally another script that trains and hyper-tunes the model on accumulated/new data can be constructed that periodically runs and outputs a new pickle model file
- a script that generates the inference from the model given some data (which will be preprocessed by the previous point)
- and HTTP (Rest) API will be used to accept the queries from the user (or business), run them trough the scripts and send the inferred data back