Mitchell Marfinetz

Milestone 1 Submission

Context

Why is this problem important to solve?

The used car market is emerging with significant importance. In India, the used car market has grown larger than the new automobile market. Demand has shifted, and we are seeing consumers replacing their old cars with pre owned vehicles. New automobiles are priced based on fairly deterministic measures called OEM's (Original Equipment Manufacturer). There are many more factors that go into determining the price of a used car, which is an important factor for the emerging trend, for both buyers and sellers.

The objectives:

What is the intended goal? Build a model that can successfully and accurately predict the price of a used car.

The key questions:

What are the key questions that need to be answered? What factors are of importance when coming to such a conclusion. What methods best fit my approach? What model best fits the problem?

The problem formulation:

What is it that we are trying to solve using data science? We are trying to come up with a model that will accurately predict the price of a used car using regression. The model will be trained on data, taking various variables and factors into account. It will be able to make predictions on data it has not seen.

Observations and Insights:

The Data consists of 13 columns. Columns S.No., Year, Kilometers_driven are of the integer type, Name, location, fuel_type, transmissions, owner_type, are of the object type, while mileage, engine, power, seats, new_price, and price are all float type.

The median value for car year in this data set was 2014. The model year of the car ranges from 1996 - 2019.

There are 2041 unique values in the column name, 11 in location, 5 in fuel type, 2 in transmission, and 4 unique values in owner type.

The majority of cars from the database had one owner, was from 2012 or newer, was a manual transmission, and was from either Mumbai, hyderabad, coimbatore, kochi, or pune.

The price tends to increase as power increases, power decreases as seats increase, the cars have similar values ranging for kilometers_driven from approx 9-12.Price also has a positive correlation with year.

Engine and price have a positive correlation, as well as power and price, year and mileage have a slightly weaker correlation. Kilometers driven and year have a negative correlation, as well as engine and mileage and power and mileage.

On average, Cars in Coimbatore and Bangalore are the most expensive, on the other hand cars in Kolkata tend to be the cheapest.

Proposed approach

Potential techniques - What different techniques should be explored? Some techniques that could be explored for this problem are linear regression, k-means clustering, or potentially a gaussian mixture module.

Overall solution design

What is the potential solution design? The solution design is to build a multiple linear regression model, that takes independent variables x, and makes a prediction y, for the used cars price.

Measures of success

What are the key measures of success? The model must not be over fitted or under fitted on the data, it is also important to be aware of potential bias.