PRICING MODEL (CASE STUDY)

DECEMBER 2024

CASE STUDY BRIEF

The new head of Data Analytics asked you to create a new machine learning model to predict vehicle prices in real time in the future. For this purpose, he has provided you with the attached dataset.
In a Jupiter notebook (or similar), create all the steps to obtain a machine learning model.

Please, focus on:

- ❑ Perform EDA (Exploratory Data Analysis)
- ❑ Data cleaning
- ❑ Feature engineering
  - Are there features you think we should ask to include (external sources, other variables you suggest)?
  - Are there features that you think should be created from those already in the dataset?
- ❑ Feature selection
- ❑ Modelling
  - What kind of model should we use?
- ❑ Evaluation of the model
  - What metrics do you suggest we use?

Delivery:
Delivery should be a Jupiter notebook or similar with the various steps specifically noted and discussed with the interviewer afterwards.

Keep in mind that:

- The purpose of the exercise is not to obtain a model with optimal performance, but to show how the various steps of model construction are approached and what kind of logic is behind it.
- When splitting the data set into train/test, Keep in mind that the prices suffer influences from market trends, so they are time dependent.
- For this reason, it is important that the code is well documented.

THE DATASET

The dataset consists of a dummy set of data with approximately 18.5K rows of vehicles sold in a total of 3 years (2021, 2022, 2023). The variable we need to predict is the last one (sellingPrice). Below is an explanation of the columns in the dataset.

Column description:

- **vehicleID**: Unique ID to identify a vehicle.
- **registrationDate**: Date when the vehicle was first registered.
- **kilometers** : Total distance the vehicle has traveled in kilometers.
- **colour**: The exterior color of the vehicle.
- **aestheticGrade** : A score reflecting the vehicle's exterior/esthetical condition (Good, Bad, Very good, etc)
- **mechanicalGrade**: A score reflecting the vehicle's mechanical condition (Good, Bad, Very good, etc)
- **saleDate** : The date when the vehicle was sold.
- **make**: Brand of the vehicle’s manufacturer (e.g., BMW, Mercedes, etc.)
- **model**: Specific model within the make of the vehicle.
- **doorNumber**: Number of doors.
- **type**: Category associated with the shape and structure of the vehicle (e.g., Sedan, Estate, Hatchback).
- **fuel**: The type of fuel the vehicle uses to operate (e.g., Petrol, Diesel).
- **transmission**: The transmission type of the vehicle (e.g., Manual, Automatic).
- **yearIntroduced**: The year when that particular model has been introduced into the market.
- **cylinder**: Cylinder of the vehicle.
- **cubeCapacity**: Capacity of the vehicle.
- **powerKW**: Power (KW) of the vehicle.
- **powerHP**: Power (HP) of the vehicle.
- **sellingPrice**: Price for which the vehicle was sold.