CASE STUDY 051
[Python]
Car Prices Linear Regression Model

Difficulty Level: 3 of 3

Suppose you are a data scientist working on a project for an insurance company that is trying to understand how the characteristics of a car could impact its price.

To perform the task, you received a dataset containing 205 observations representing cars with 26 variables defining its characteristics. This dataset was downloaded from the University of California, Irvine Machine Learning repository (the original file and description can be downloaded at https://archive.ics.uci.edu/ml/datasets/automobile)

You also received a description of the dataset including the definition of each variable.

Your manager asked you to build a model that implements machine learning linear regression on the received data. To get the work done, she asked you to:

1) Read the dataset and perform an Exploratory Data Analysis

    a) Read the file

    b) Check the column names and first rows

    c) Check for missing values

    d) Check the datatypes

2) Read the dataset again, specifying that the dataset doesn't have a header row, indicating the column names and the correct interpretation of missing data

    a) Read the file

    b) Check the column names and first rows

    c) Check for missing values

    d) Check the datatypes

    e) Analyze the categorical data

3) Remove unused columns

4) Deal with missing data

    a) num_of_doors column

    b) bore column

    c) stroke column

d) horsepower column

e) peak_rpm column

f) price column

5) Deal with categorical columns

a) num_of_cylinders column

b) Other categorical columns

6) Split your data into train (80%) and test (20%) data, and separe the dependent variables of the independent variables

a) Split the original data into train and test datasets

b) Separate your dependent variable of the training data

c) Separate your dependent variable of the test data

7) Train and execute your model

a) Create the linear regression object

b) Train the model using the training sets

c) Make predictions using the testing set

8) Assess the performance of your model

a) Print the R-Squared of your model

b) print the comparison between the prediction of the model and the actual data

Good luck!

Difficulty note: this is a difficult assignment. Do not be surprised that there will be lots of nuances we have not covered off in the courses. But just like in the Real Life – there will be things training has not prepared you for and you will need to do research to find how to solve the problems at hand. If you get stuck, check the clues file.