DataScientest

Abstract

This document explains the process of the project CO2 emissions in 2019.

Members:
- Theo Bonsu
- Julián Márquez

Report 1: exploration, data visualization and data pre-processing report

Introduction to the project

Context

The main goal of the project assigned is to determine how all the characteristics presented in cars affect the emissions of CO2. One must understand the features of cars as well as the relation among them.

From an economic point of view, it is interesting to find the brands that generate more pollution to reduce the manufacturing so the companies will not get any kind of fine from the government.

Scientifically speaking, it is desirable to have an eco-friendly environment because it is well known CO2 emissions are extremely dangerous for human health, then understanding the reason why one vehicle's brand contaminates more than other, one will focus on those cars with a tendency of high pollution.

Objectives

The purpose of the project is to study the features of every brand of vehicle and to analyze how each one emits more emissions. Some of them will have a huge impact, those are the main features to analyze.

DataScientest

Understanding and manipulation of data

Framework

To be able to do the test it will be used the dataset of 2019 for emissions and the dataset for the technical characteristics of vehicles. It is easily to find it in this source: CO2 and pollutant emissions of vehicles marketed in France - data.gouv.fr , where it has the technical features and for emissions it will be found in: https://www.eea.europa.eu/data-and-maps/data/co2-cars-emission-20 , both are available anytime.

The dataset of emissions is the larger one, about 2 GB of data.

Relevance

The principal variable for the objective is the one which stores the emissions. Furthermore, the weight of vehicles, the width of wheels, fuel consumption and others are highly correlated with emissions.

The target variable is 'Ewltp (g/km)' which indicates how many grams per kilometer a vehicle emits to the atmosphere.

The only limitation one can find is that not all features offer information and others are completely empty.

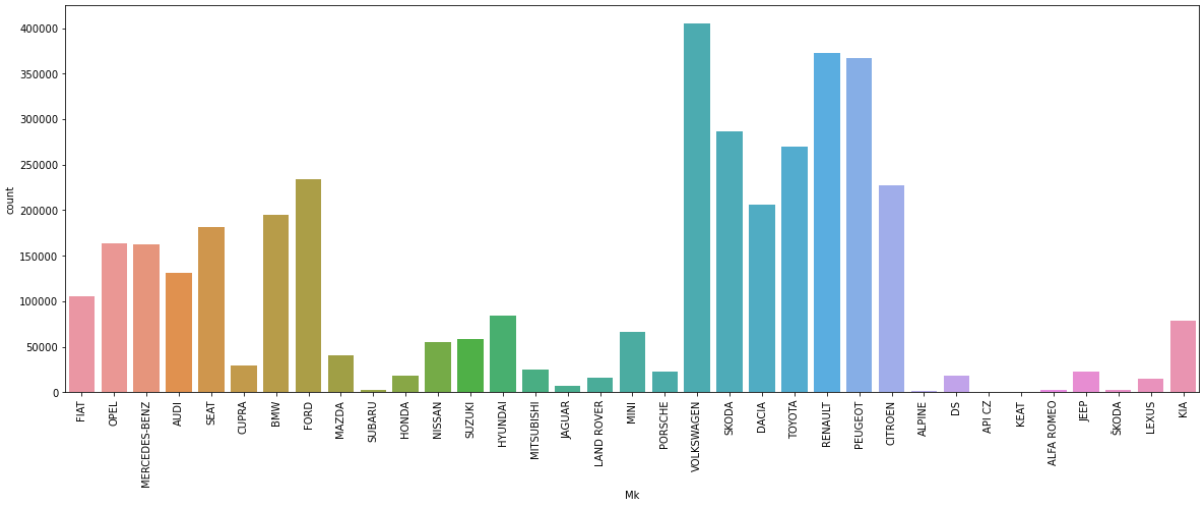
Pre-processing and feature engineering

Before starting the study it is necessary to clean the data. Due to the size of the dataset, it is mandatory to remove those columns that are completely empty, and they do not represent relevant information or correlation with the target.

Those that are not completely empty, they will be filled with the median, so it is easy to drive in the analysis.

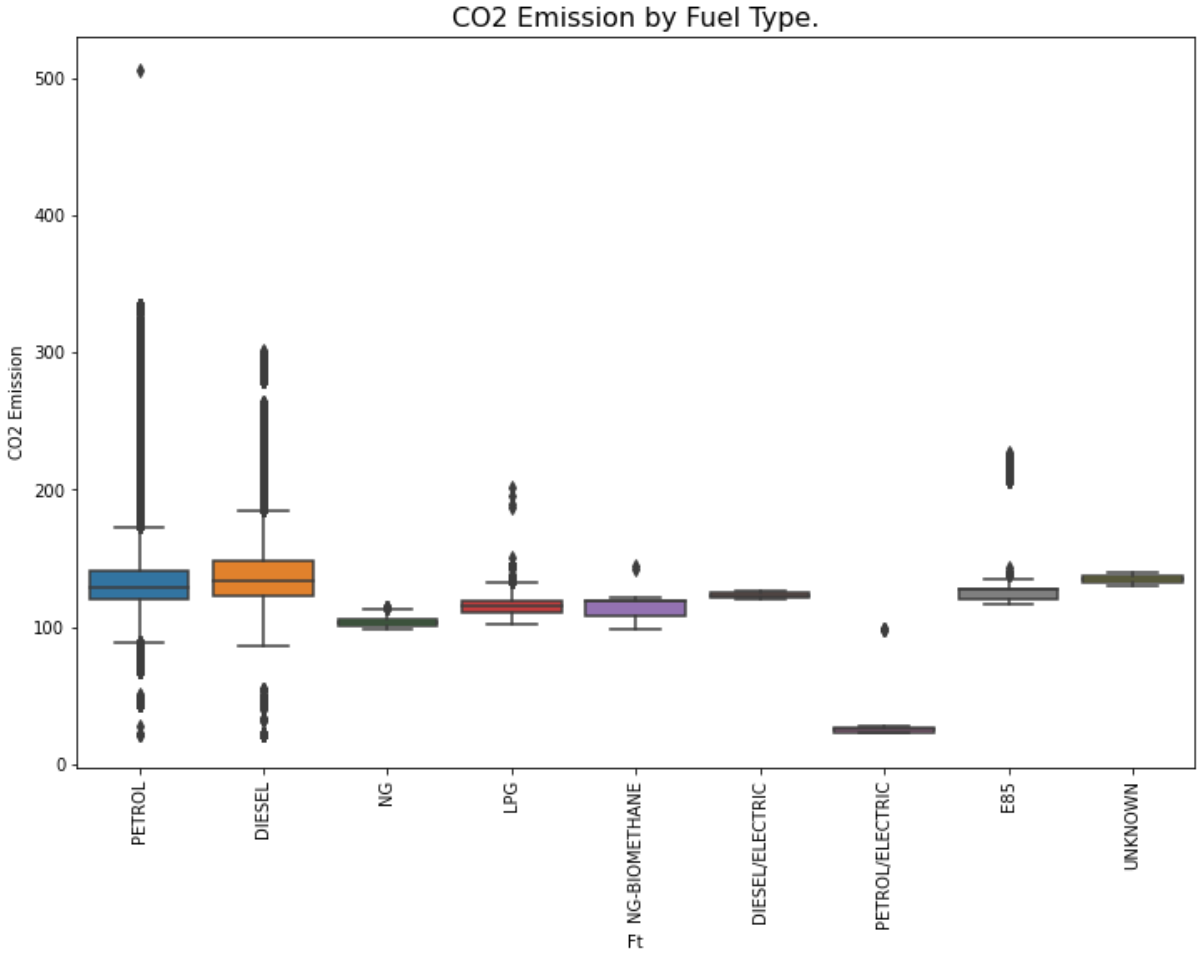
Visualizations and Statistics

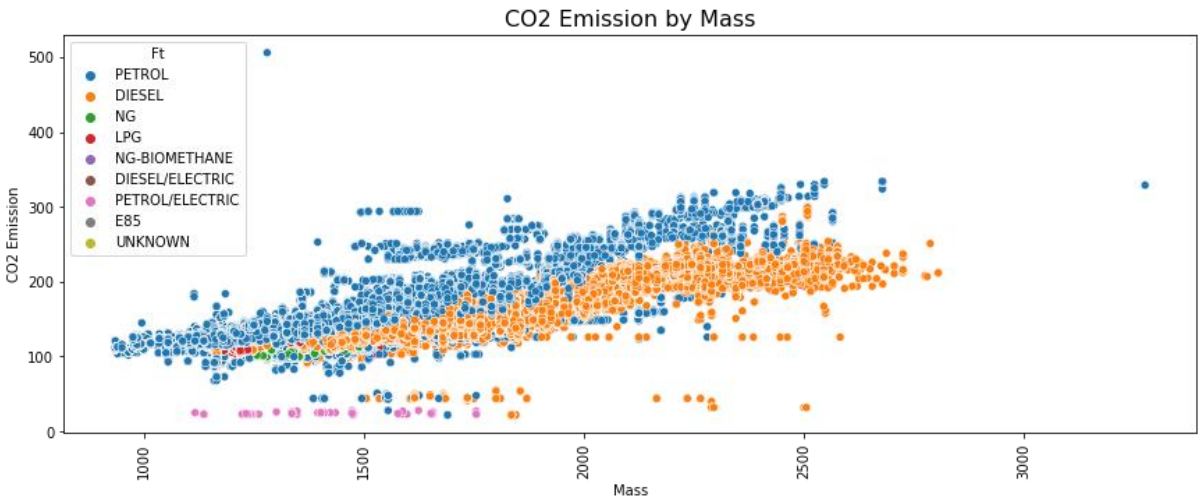
The techniques of representation allow us to see the relations between variables and relations among target and other variables. Here it is few examples:



As seen in the graph, the variables: ep(KW), Mt, m(kg), At2(mm), At1(mm), ec(cm3) and W(mm) are very correlated with the variable of CO2 emissions, which is Ewltp (g/km).

DataScientest



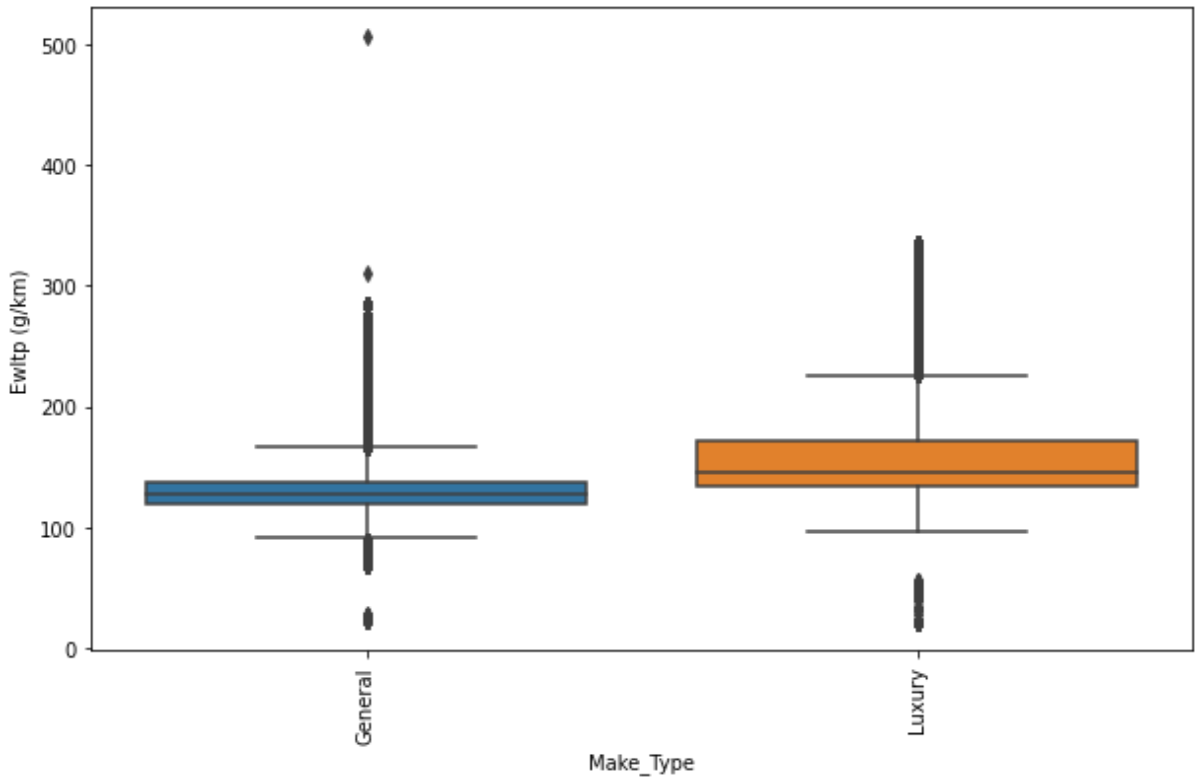


DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

The distribution of the data indicates that some companies produce more pollution than others

Looking at the graphs and the number of values in the dataset, it is reasonably likely to find outliers. This problem will be discussed later.

With all this information, we can create a classification whether the car is considered luxurious or not. The brands considered as 'luxury' are: BMW, AUDI, MERCEDES-BENZ, PORSCHE, CUPRA, JEEP, LAND ROVER, LEXUS and JAGUAR. The ones that are not in this list are considered as general. The next graph represents the classification mentioned above according to their emissions.



The luxurious cars emit more pollution due to their power and consumption of fuel. Considering this information, we can observe that luxury cars do not have as many outliers as the general cars have. It can be easily explained. Wealthy people can afford any car so when their car gets older, they replace it for another better and new one. In contrast, the average person has to acquiesce in having a car for several years. The older the car, the more pollution it emits.

DataScientest

Report 2: modeling report

Stages of the project

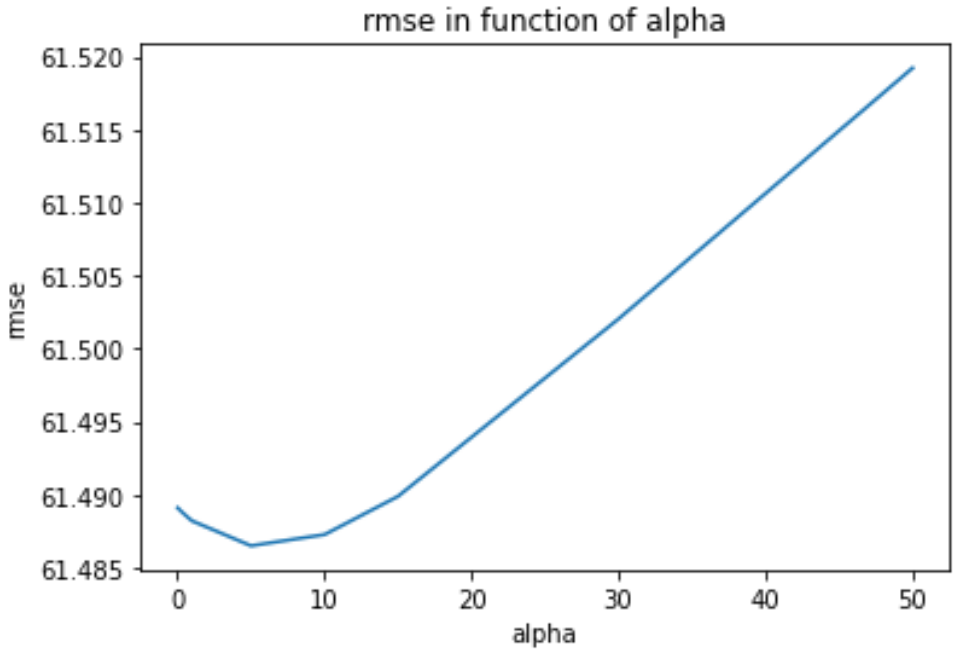
Classification of the problem

The first part will consist of doing a simple linear regression because one wants to predict what kind of cars will emit more pollution. Therefore, the primary task is emission detection by car.

We have used two models for predictions. The Lasso regression and the ridge regression. Both of them have good performance and the results are practically the same.

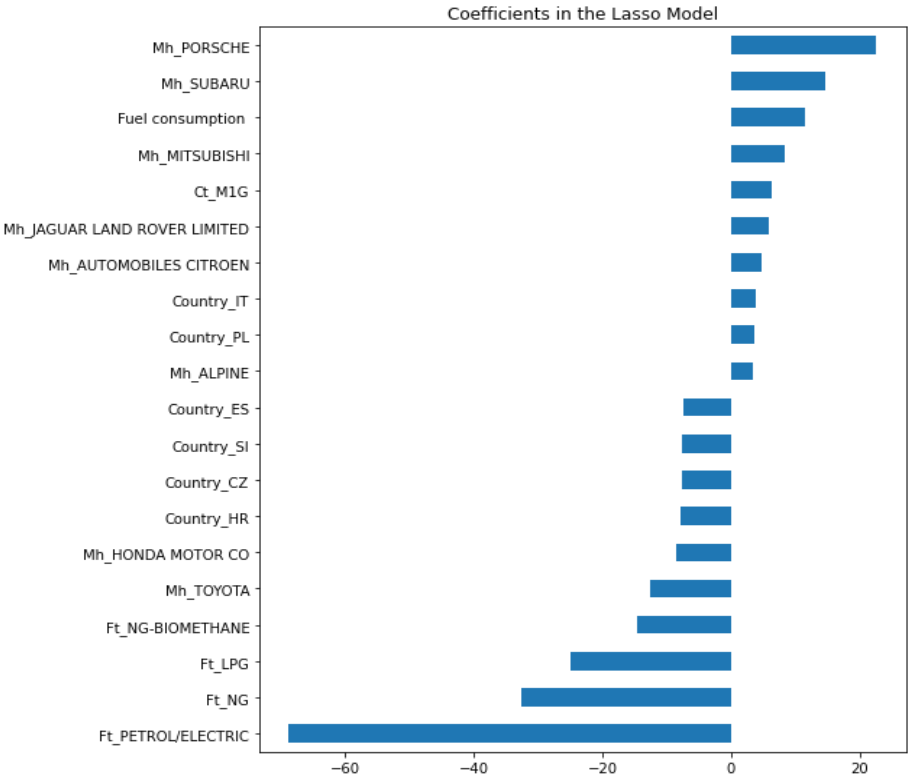
To compare what models predict the best, we will use the function GridSearchCV() because with it we can see what are the hyperparameters that give the best performance. Then, we compare and evaluate each model by cross validation.

Let's take a look first at the ridge model. For a given values of alphas, we have the next result:



The test score for this model is 89.41%. The value of alpha that has less error is alpha = 5, so with this value the test score is 89.42%.

For the Lasso model, we have a score of 89.41%. The following graph shows what variables are most important to the coefficients.



The model has picked up 60 variables and erased 7 of them.

The next stage is to find out what model gets the best performance. To do so, we have used a grid search technique. Implementing the same coefficients, the results are:

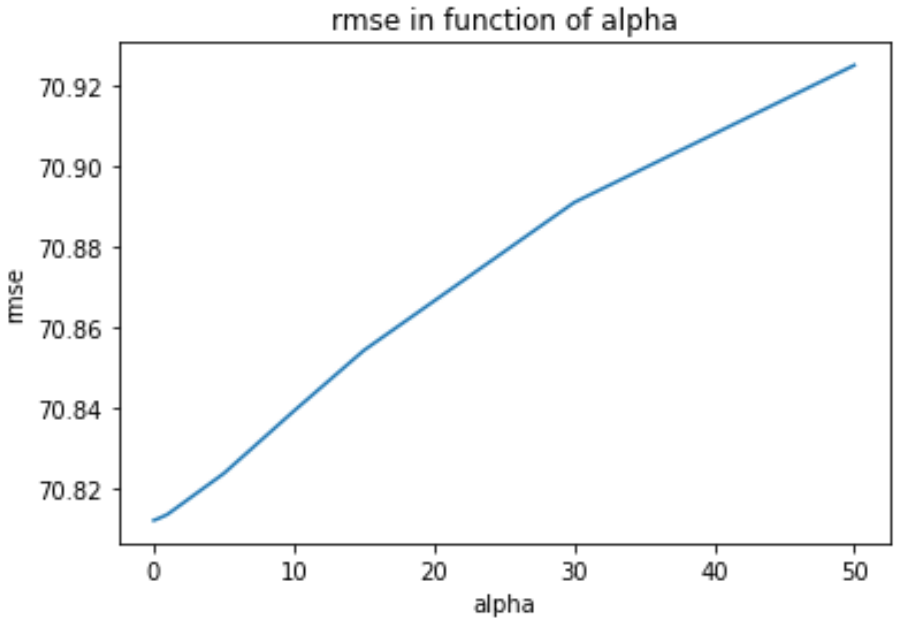
- Best estimator for Lasso: α = 0.005
- Best estimator for ridge: α = 0.01

With these coefficients, the results are:

- Best score for Lasso: 89.26%
- Best score for ridge: 89.27%

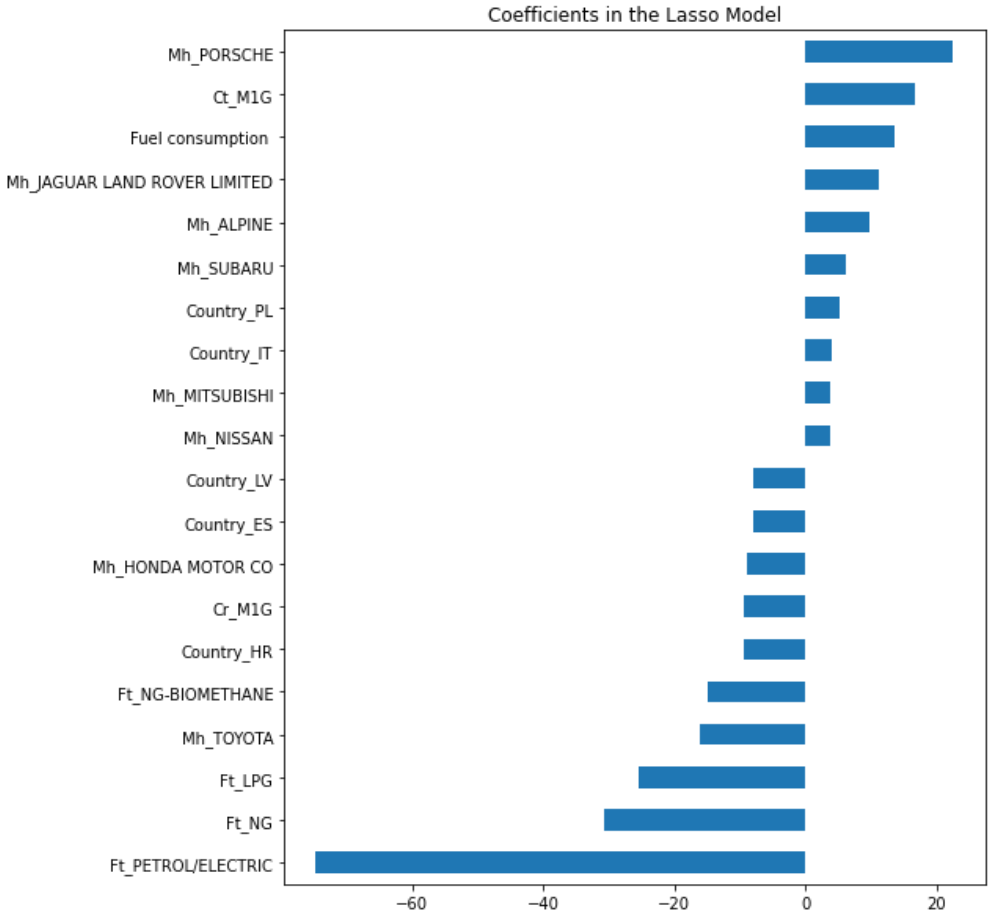
For instance, the ridge model is slightly better.

To make it more difficult, we are going to show the same interpretability with less correlated variables. In this case, we have erased *ep (KW)* and *m (kg)*. Applying the same techniques what we get is:



The test score for this model is 87.10%. The value of alpha that has less error is alpha = 0, so with this value the test score is 87.86% which is lower than the other model with all correlated variables. Despite this, the result is still accurate.

For the Lasso model, we have a score of 87.88%. The following graph shows what variables are most important to the coefficients.



The coefficients have a little variation for every variable.

With grid search technique, implementing the same coefficients, the results are:

- Best estimator for Lasso: α = 0.005
- Best estimator for ridge: α = 0.01

With these coefficients, the results are:

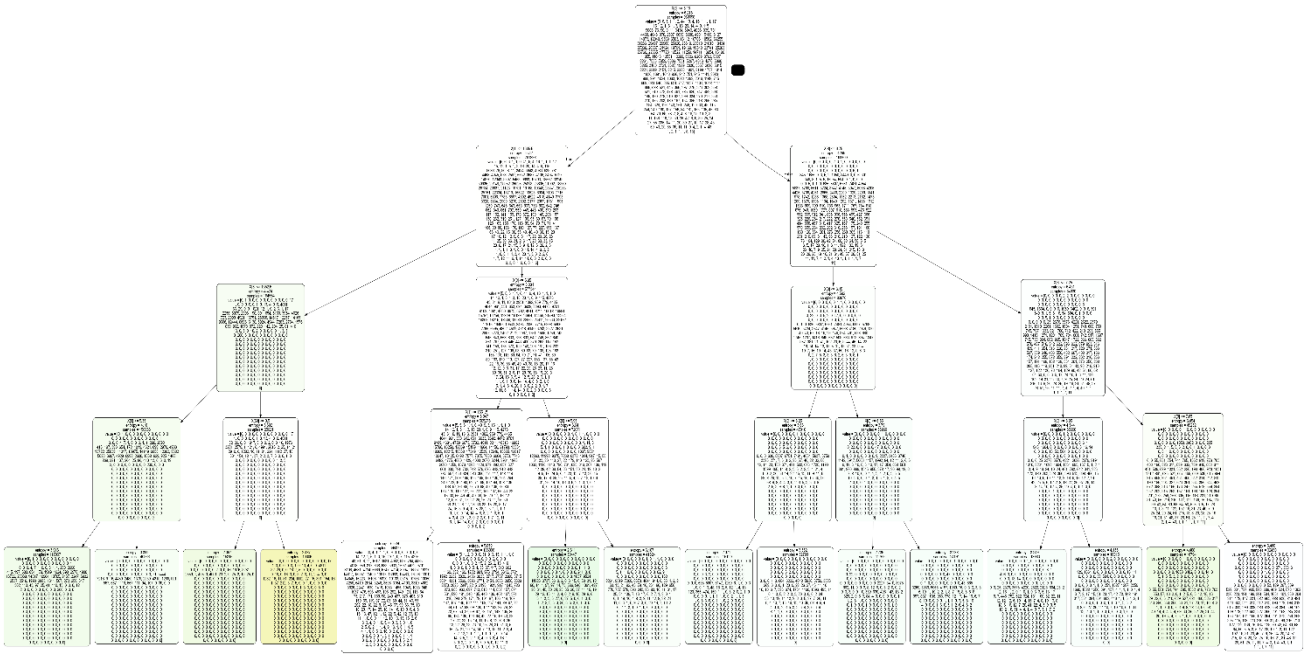
- Best score for Lasso: 87.094%
- Best score for ridge: 87.098%

Again, the ridge model is slightly better and both results are lower than the previous because of the lack of correlated variables.

The final step is the classification with decision trees. After using a reduction of the data, about 30 % of total, the score with all correlated variables and without all correlated variables are:

- Precision of decision tree with all variables: 14.103%
- Precision of decision tree without all variables: 12.76%

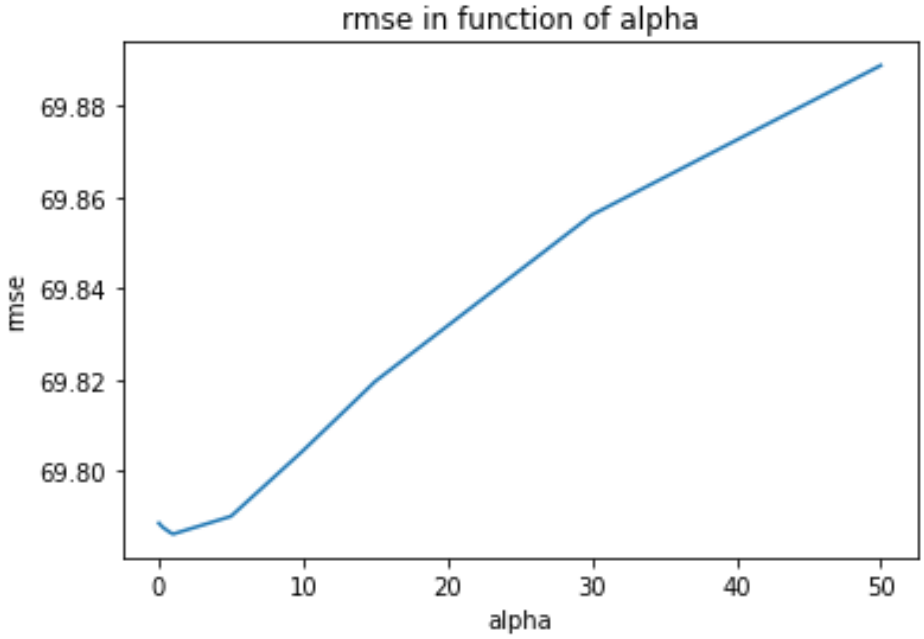
These values are very low because we have used only 30 % of all data.



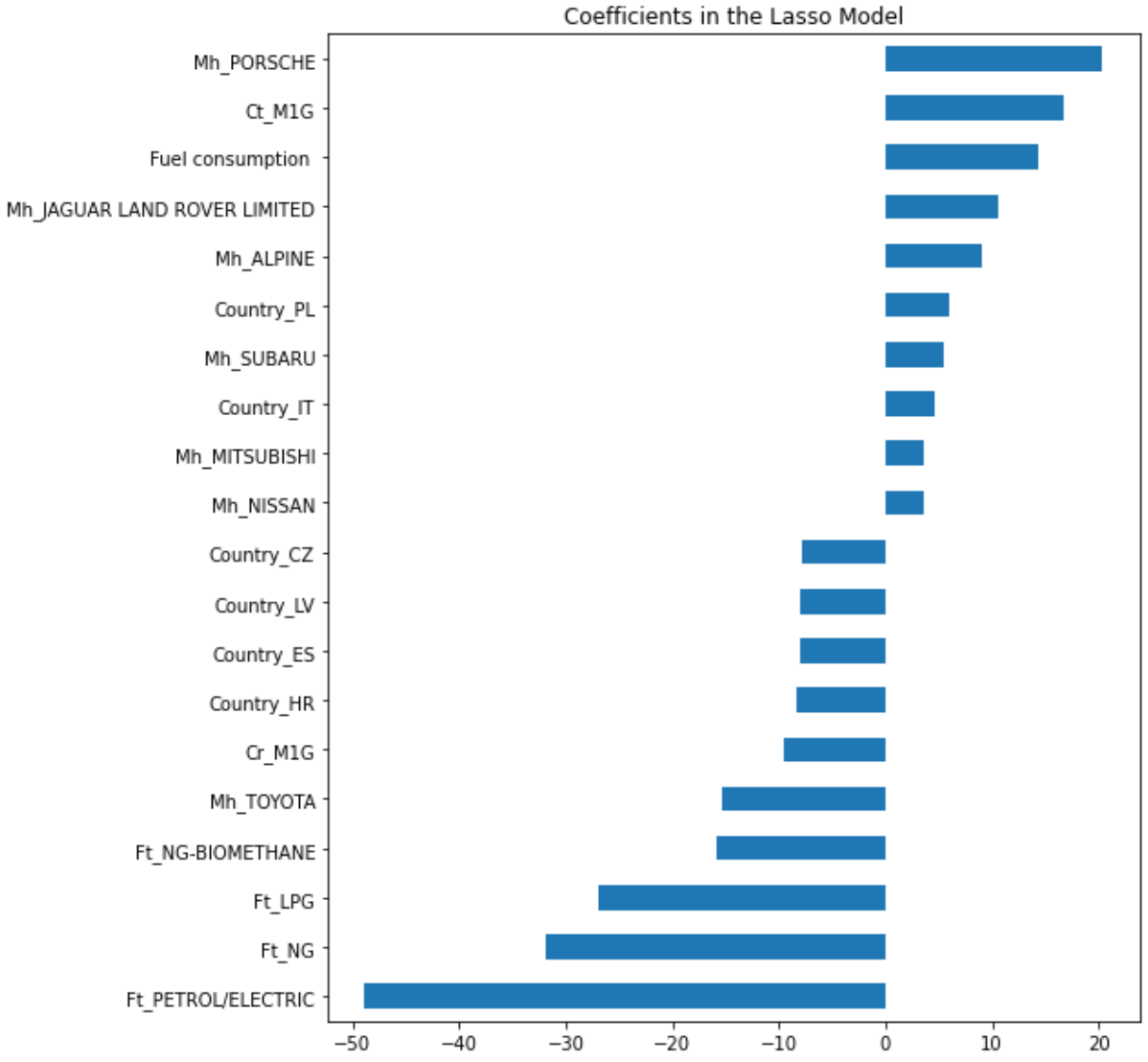
(We have to explain all this)

The second part is doing the same without correlated variables. We have deleted *'ep (KW)'* and *'m (kg)'* (explaining these variables).

Firstly, we applied the models for regression. For the ridge regression, we get for train score of 87.54% and for the test score of 88.03%. With all variables the result was 89.41%. We can conclude that modelization was fine and accurate.



For the Lasso regression we had 87.55% for the train score and 88.04%. Comparing these with all variables, the results are still fine and accurate. In this case, the Lasso model has picked 58 variables and has eliminated 5.



Model choice and optimization

The models for the first analysis are Lasso and ridge regressions. We are going to use both, and we will compare the models and choose the one with lower errors. To select one of them, we are going to use grid search and cross validations methods. Afterwards, the model used is the decision tree. We have selected ridge regression because it has higher precision, and it takes less power of the computer to make predictions in our case.

For the decision tree, we must study the documentation so we can use the proper criteria and get the precise results. For that reason, we have selected ‘entropy’ for criterion, then we have divided the dataset in two samples to take less time to implement the model.

Interpretation of results

On one hand, the very first interpretation we can see shows that Lasso and ridge regression almost the same way. Comparing the score of both models, the ridge model is slightly better.

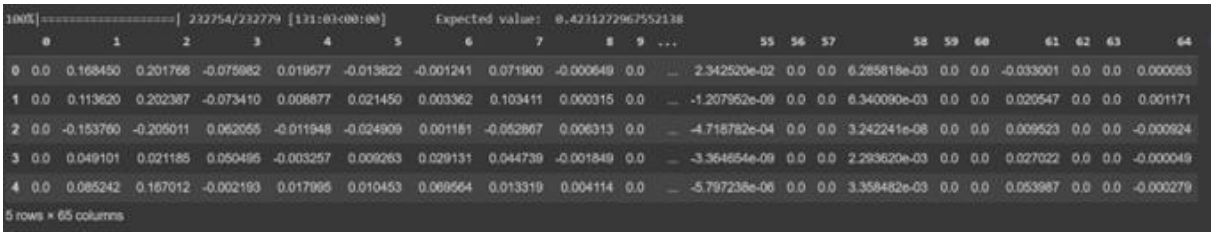
Although we have erased correlated variables, the model is still good and gives us excellent results if we consider we are using the third part of the total data.

On the other hand, the decision tree reveals a poor scoring on the classification. We conclude it due to the few data we have used for testing the model, otherwise, it would be necessary to use a powerful computer to do it in a reasonable period of time.

In order to obtain better results we have introduced grid search and cross validation techniques for optimization. Afterwards, with results in hand, we have applied the better parameters.

We have introduced a classification based on the level of emission by car. If one car has emitted more than 134 grams per km (that number is the average of emissions for all cars), it will have the value 1, otherwise it will have 0 if it has emitted less than 134 g/km.

The interpretability technique used is the SHAP method.



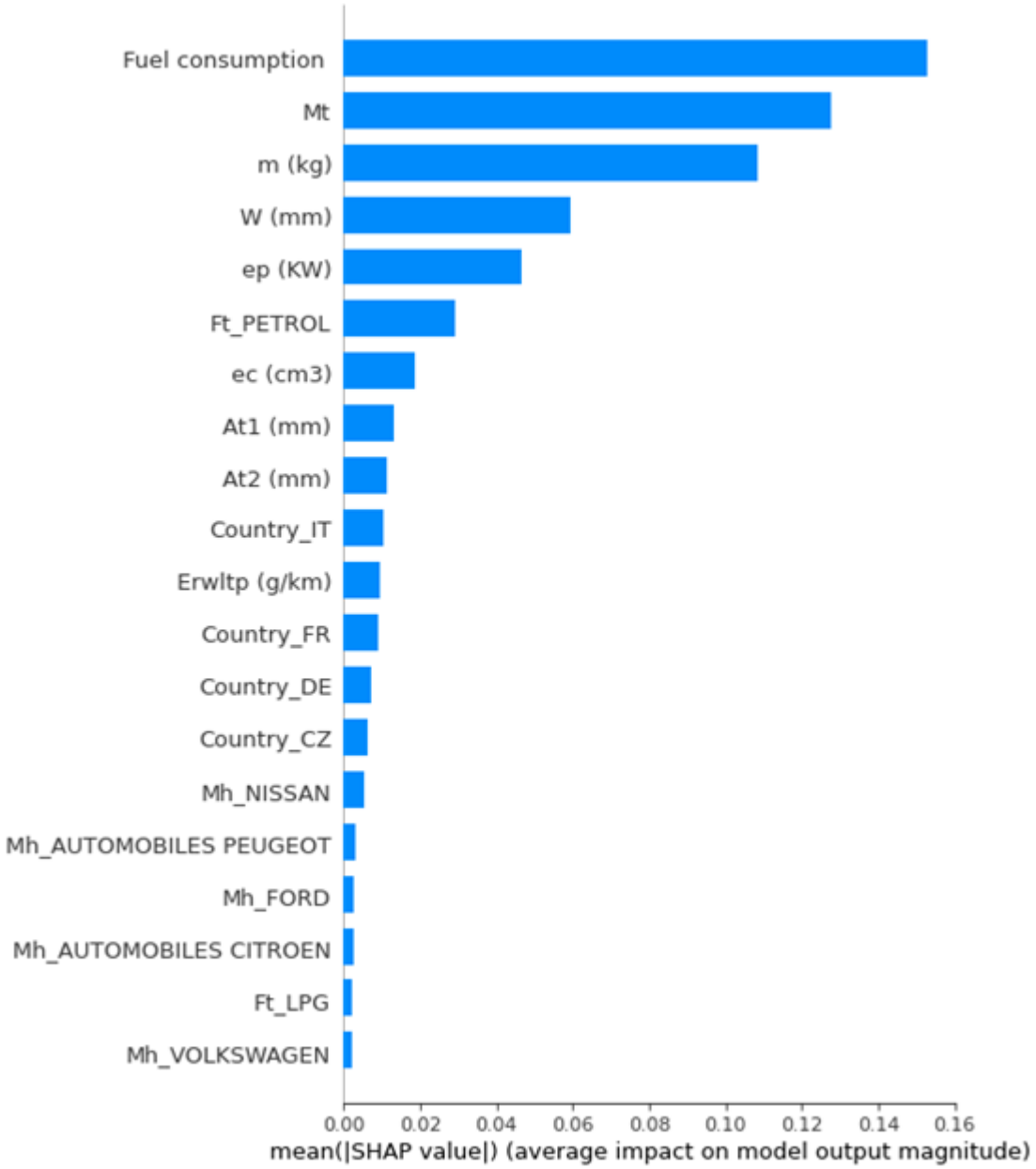
	0	1	2	3	4	5	6	7	8	9	...	55	56	57	58	59	60	61	62	63	64
0	0.0	0.168450	0.201766	-0.075982	0.019577	-0.013822	-0.001241	0.071900	-0.000649	0.0	...	2.342520e-02	0.0	0.0	6.285818e-03	0.0	0.0	-0.033001	0.0	0.0	0.000053
1	0.0	0.113620	0.202387	-0.073410	0.008877	0.021450	0.003362	0.103411	0.000315	0.0	...	-1.207952e-09	0.0	0.0	6.340090e-03	0.0	0.0	0.020547	0.0	0.0	0.001171
2	0.0	-0.153760	-0.205011	0.062055	-0.011948	-0.024909	0.001181	-0.052867	0.006313	0.0	...	-4.718782e-04	0.0	0.0	3.242241e-08	0.0	0.0	0.009523	0.0	0.0	-0.000924
3	0.0	0.049101	0.021185	0.050495	-0.003257	0.009263	0.029131	0.044739	-0.001849	0.0	...	-3.364654e-09	0.0	0.0	2.293620e-03	0.0	0.0	0.027022	0.0	0.0	-0.000049
4	0.0	0.085242	0.167012	-0.002193	0.017995	0.010453	0.068564	0.013319	0.004114	0.0	...	-5.797238e-06	0.0	0.0	3.358482e-03	0.0	0.0	0.053987	0.0	0.0	-0.000279

Every row belongs to a single prediction made by the model of regression, in our case, ridge regression. Each column, indicated numerically, represents a feature used in the model. The values in the table represent the contribution of each feature to the prediction.

Negative values have negative values, which indicates it would predict 0 (that means a value of emission less than 134 g/km, as explained before). Positive SHAP values show positive impact, leading the model to predict 1.

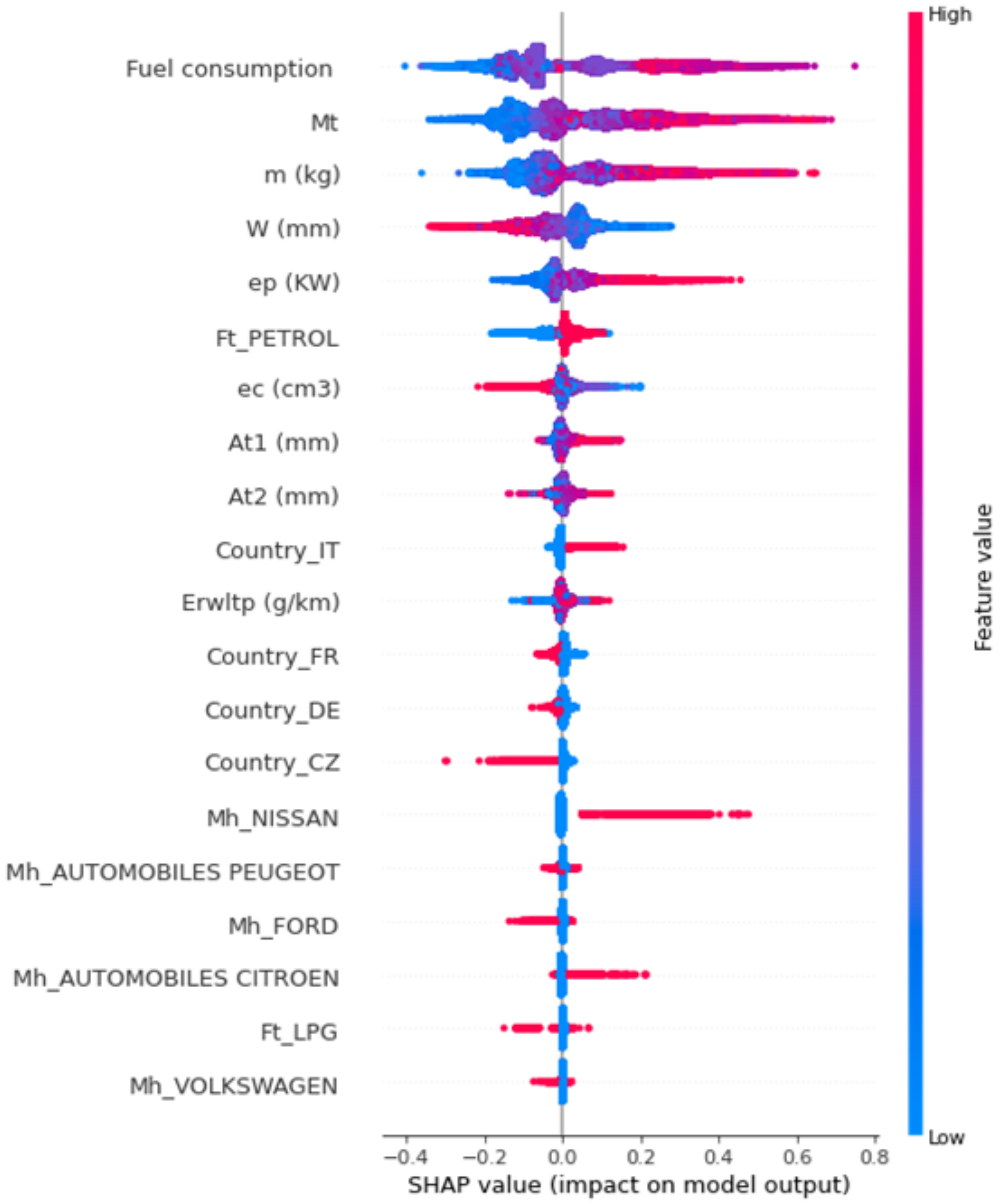
According to the next figure, we can see the summary plot to see the importance of the variables and how they influence the target. As expected, the fuel consumption has a big impact. Surprisingly, the Mt variable is the second one with

a huge impact. Mt stands for “mass in running order”, it corresponds to the mass with passengers or any object inside the car. Of course, the more weight the car carries, the more pollution it emits.



The next graph is the directionality impact. The x-axis stands for SHAP value, while the y-axis represents all the features. Every point on the chart indicates

one SHAP value for a prediction. We can conclude that higher value of 'Fuel consumption' leads to higher prediction of emitting high volume of CO2. The same is for the other features related to the mass of the car and its horsepower.



As previously compared the insights, the model is well enough to predict emissions.

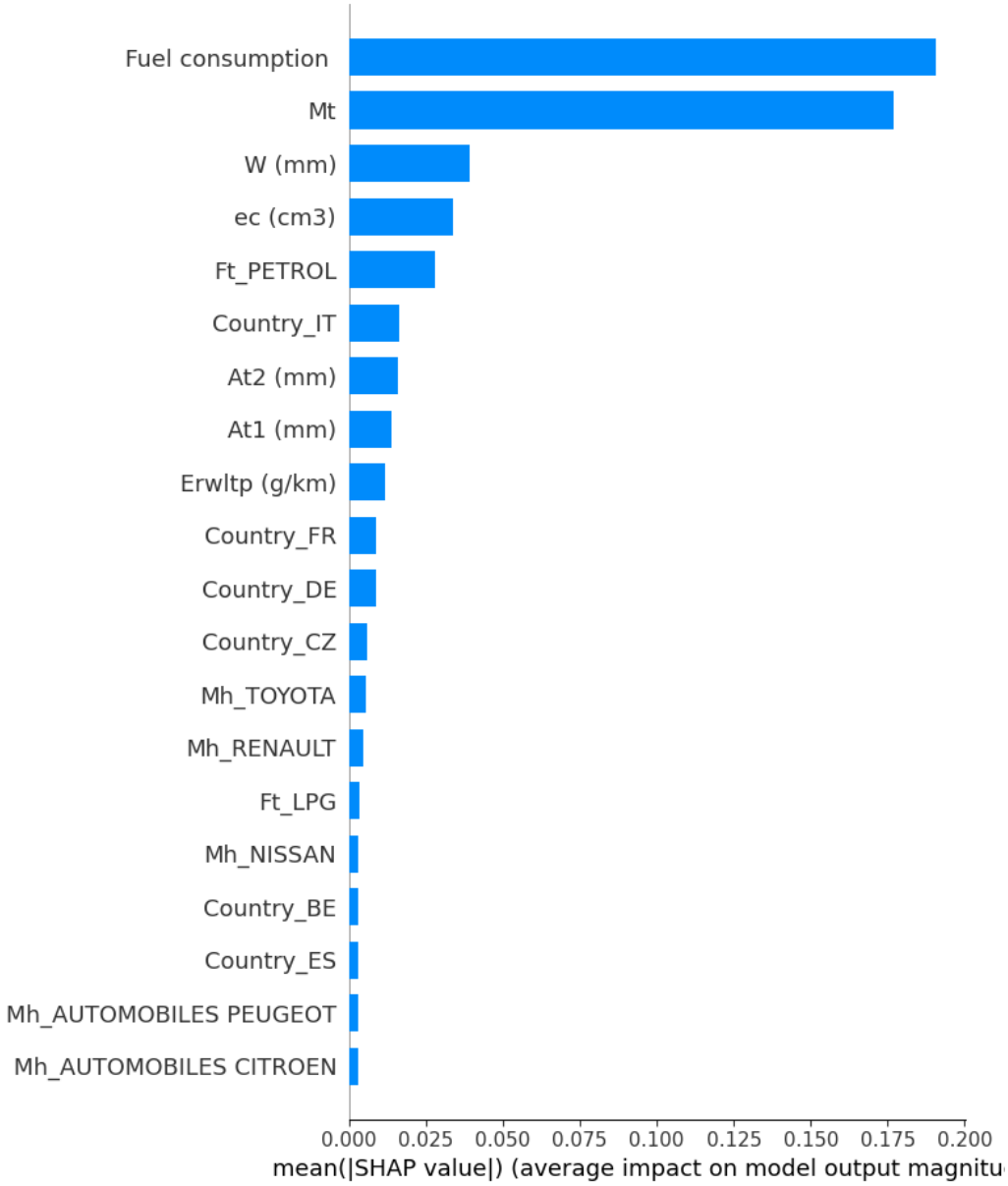
Individual predictions are shown below. 'Fuel consumption', the mass and the width of wheels have positive impact on the predictions

The regression models provided us an excellent result before using interpretability techniques, avoiding overfitting problems, so it has not generated improvement performance.

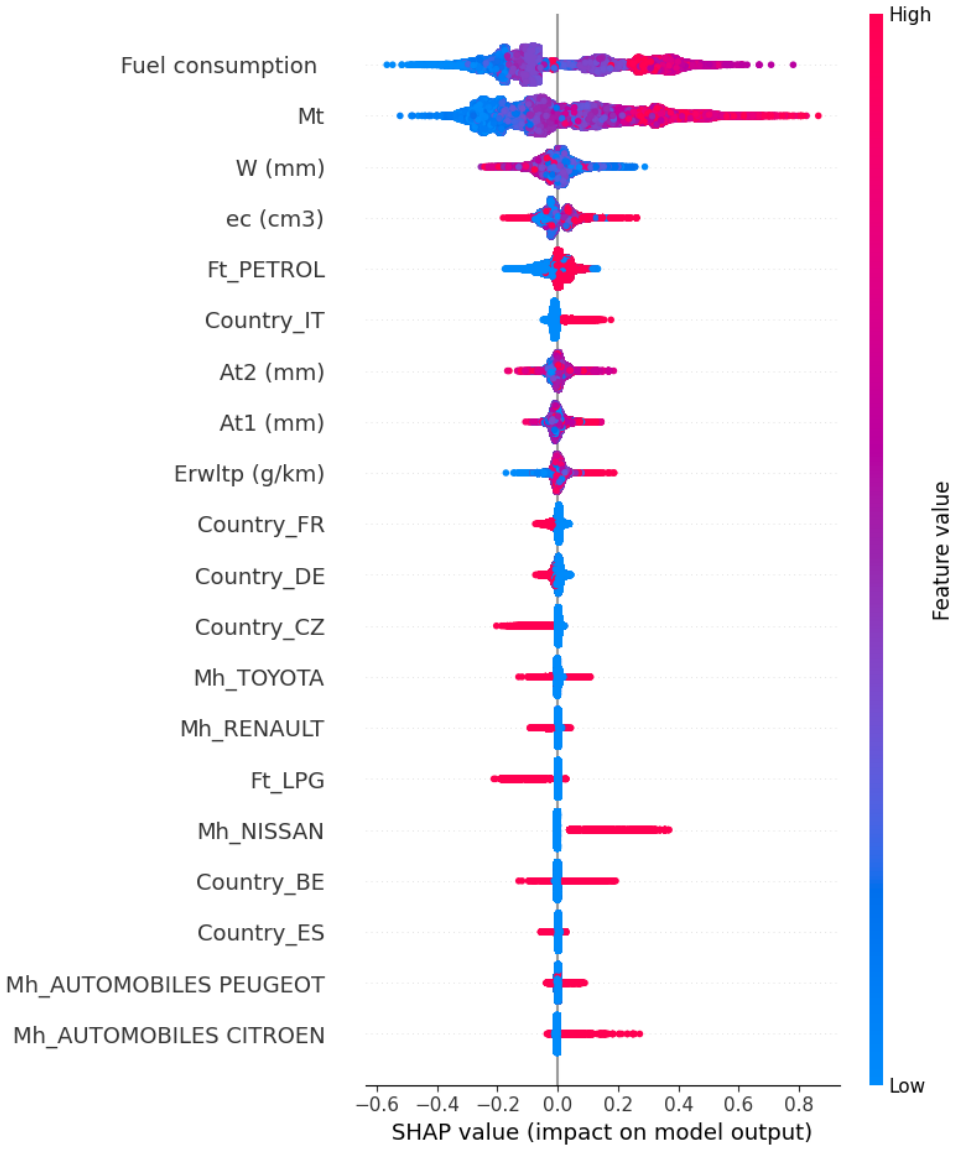
Interpretability without the correlated variables

One question that someone can ask is: what does it happen when we remove the correlated variables?

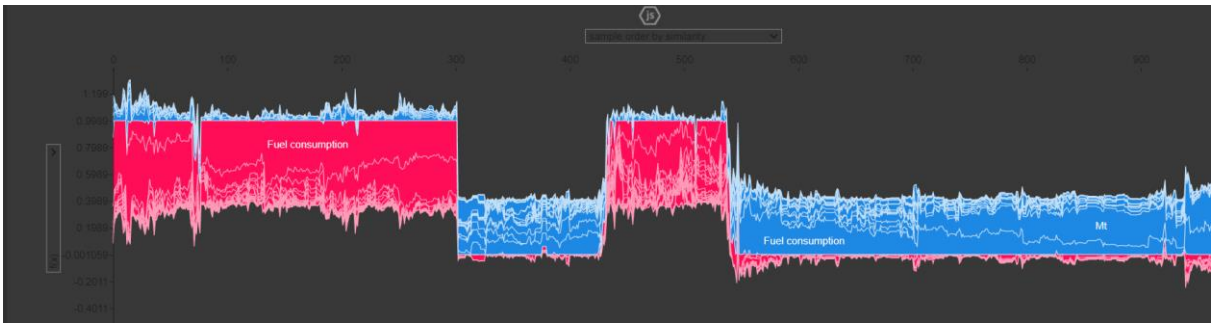
In the image beneath this paragraph, the variables that have more impact are Mt and W (mm). The second one is the width of the wheels, the higher the value of the width, the more consumption the car has. In Europe it is common to have a set of wheels for summer or winter. Depending on the season, the width may change the fuel consumption.



The same interpretation for the shap values. Every point is a prediction. In this case, the feature W (mm) has negative values, so it means it would predict that cars do not emit more with more width of the wheels.

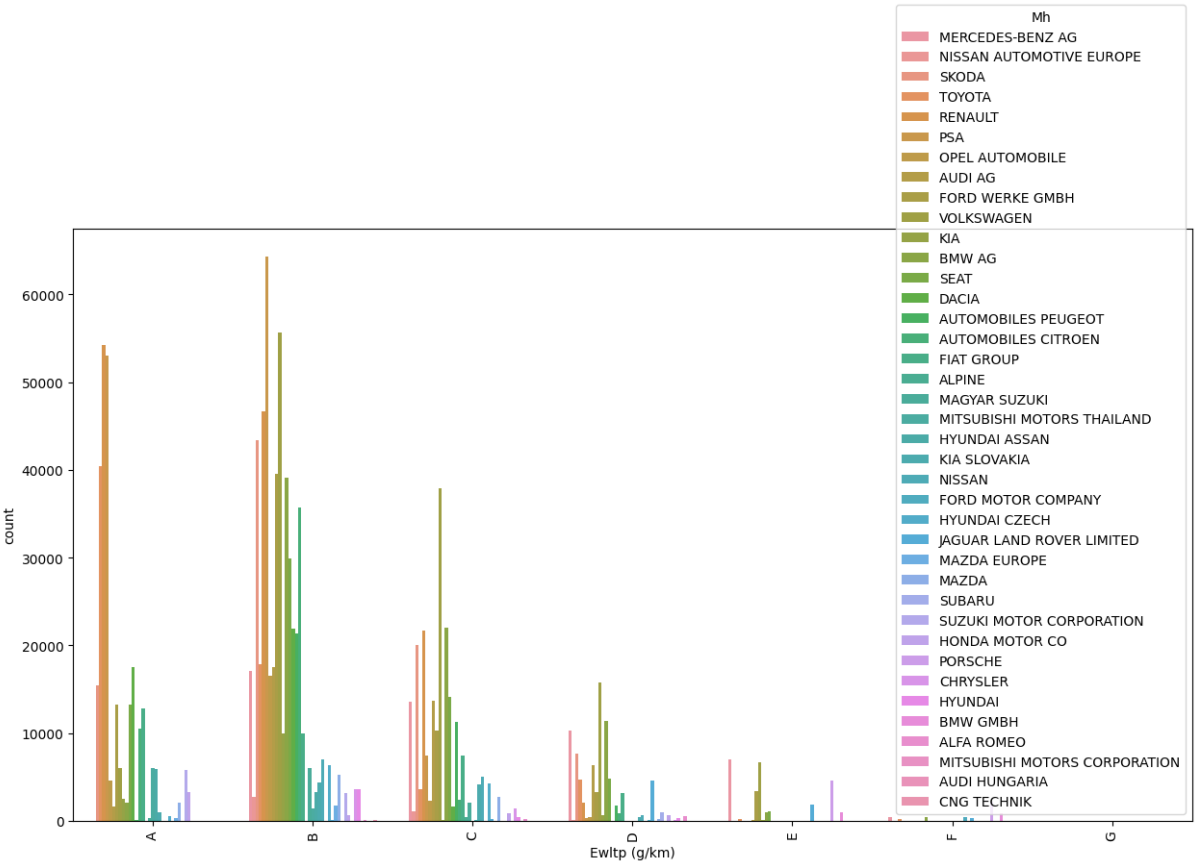


The final part is the individual impacts.



The final part we are going to explain the process of creating the application to display beautiful graphs and the results. To do this, we use streamlit.

A final classification is shown beneath. Comparing the brands of the cars, we can divide into labels depending on the emissions. For example, if a car emits less than a certain value, in our case 134 g/km, it will have the label ‘A’. The more emitted, the higher the label.



The library streamlit allow us to create an application that make easy to share beautiful and custom web apps in order to show a project of machine learning of data science in general. Simple but powerful.

Assessment methods:

DataScientest

Professional scenario: based on a proposed solution, the candidate will have to produce a summary report including: the explanation of the choices of AI solutions implemented, the interpretation of the results, the evaluation of the reliability of the algorithms and an optimization proposal.

Final report :

Conclusion drawn

Difficulties encountered during the project

Despite being a large dataset for an introductory course of data science, the difficulties are few. The most important variables were full or at least, almost filled with values.

The main problem was the computational power. For every model implemented it took hours to display a result. Occasionally it could pass the whole day running the cell without a result.

To resolve this problem, we had to sample the data to achieve the goals proposed. That refrained us from finishing the project as soon as we wanted.

The theoretical skill we needed to use was plotting the decision tree. It was not even hard, and it only required learning the documentation given in the bibliography.

DataScientest

Report

Continuation of the project

Bibliography

- *Python Machine Learning,* Sebastian Raschka, Vahid Mijalilli, ed Marcombo
- https://scikit-learn.org/stable/modules/generated/sklearn.tree.plot_tree.html
- https://www.geeksforgeeks.org/decision-tree/
- https://graphviz.org/documentation/
- https://pydotplus.readthedocs.io/

- https://ourworldindata.org/co2-dataset-sources

Appendices