Machine Learning: supervised algorithms and application for predicting the energy efficiency of vehicles

Pedro Venancio
Estela Mari Ricetti Bueno
MAHLE

Email: pedro.venancio@mahle.com; estela.bueno@mahle.com

ABSTRACT

Decisions in the industry are increasingly becoming data driven and the technological advance in communication provides huge data generation and availability, but only after being filtered and processed it can be transformed in useful information aiming to build knowledge and drive to intelligible conclusions. Machine learning algorithms are developed to this objective, these algorithms have wide applications range coupling a great power of information processing and prediction generation based on patterns. The automotive sector needs predictability and can take relevant advantages on future tendency predictions. This work presents an application of machine learning algorithm to estimate the energy efficiency of vehicles, expressed in MJ/km, to understand the potential application of some components in new vehicles. Some supervised machine learning algorithms are applied to predict the energy efficiency of a vehicle, based on constructive characteristics, such as mass, power, volumetric displacement, transmission, among others, aiming to obtain the highest possible reliability. Python programming interface was used, together with a database of passenger and light commercial vehicles, and comparisons were made among these algorithms aiming at this search with better performance.

INTRODUCTION

The efficiency of vehicles is one of the most latent themes within the automotive sector, largely due to the environmental concern surrounding the emission of pollutants. Over the years, various technologies have been incorporated into vehicles in order to increase efficiency. The measurement of the efficiency of the vehicle can take place in different units, among them the measurement in mega joules per kilometer, also called specific consumption.

The possibility of estimating specific consumption (or efficiency) from basic characteristics is an interesting way to guide the development of vehicles. It is possible, from some input data, to estimate the specific consumption of the vehicle with different configurations of powertrain and aggregates, using as a basis information from the powertrains of existing vehicles. This estimate becomes even more interesting when put from the perspective of the law that regulates energy efficiency for passenger cars and light commercial vehicles, Route 2030. From the specific consumption value, obtained through a prediction, and the mass, one can fit the vehicle within the lines described in the program and identify the possibility of obtaining a discount of 1 or 2 percentage points in IPI (tax on industrialized products) in Brazil.

To estimate the specific consumption, machine learning (ML) concepts were used. Machine learning is a subset within artificial intelligence that focuses on using data to mimic the way humans learn, using dynamic algorithms capable of making decisions, learning, and gradually improving their accuracy.

Aiming to generate knowledge based on the data available and artificial intelligence technology, the present work focuses on presenting the development of a machine learning algorithm to predict the energy consumption, in mega joules per kilometer, of a passenger or light commercial vehicle. The work also presents some concepts related to machine learning, which are necessary for understanding the choices made during development.

MACHINE LEARNING AND ALGORITHMS

The term "machine learning" was first employed in 1959 by Arthur Samuel, an engineer at MIT (Massachusetts Institute of Technology), to define a field of study that allowed computers to learn to solve problems without having a specific programming for it. The basic concept that supports this process is to develop algorithms that, based on existing data, can adjust the parameters of a mathematical model to predict a situation of the same nature as the data provided. Machine learning algorithms are mainly based on linear algebra and statistics, obviously having computer programming as the main tool. Since its creation research field has expanded significantly and the advancement of computational hardware technology has allowed a significant increase in the size of databases and the number of model parameters, as well as the adoption of models with high processing requirements.

The implementation of a machine learning process is composed of well-established phases, as described in the CRISP-DM (Cross-industry standard process for data mining) flowchart Figure 1.

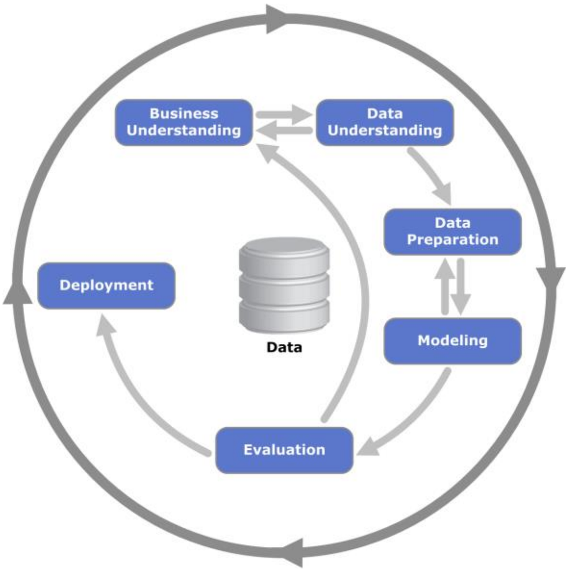

Figure 1. CRISP-DM Flowchart [1]

The implementation of a machine learning process begins with understanding the business, clearly establishing the problem to be solved and what information is necessary and available to obtain the solution. This initial step should be done in conjunction with the experts in the business itself and the data scientists and engineers who will manage the machine learning process. The conclusion of this first step is the choice of the type of machine learning model to be adopted that fits better to the situation to be solved.

The next step is the preparation of data for the machine learning modeling process. Often this step is the one which requires more time of the process because it requires laborious work of selection, clean up and organization of data to start the modeling step. In the modeling phase several algorithms can be adopted because it is not possible to establish a priori a perfect model for each problem, in addition to the very statistical nature of the method brings the need for checks of error residuals and statistical correlations.

After choosing an implemented model and whose results meet the statistical criteria, the evaluation stage of the model is reached, making its pilot application within the business, that is, to meet the desired real forecast. This phase of initial testing is critical to identify overfitting, excessive deviations from predictions, response time, among others. This phase connects with the initial phase of understanding the business, because at this stage it is possible to identify any non-attendance of the model to real needs and allows corrections returned to the development lopping. After a set of tests, the model can be implemented for use as a tool for the business. Throughout the period of use the model can and should be fed back and improved, in such a way that it constitutes a dynamic tool.

The machining learning algorithms can be classified basically in three groups that are described in the forward bullets:

- Supervised algorithms: The algorithms learn to make predictions using historical patterns. An example is a weather forecasting algorithm based on weather characteristics. In this case, a historical database will be used that will allow the algorithm to learn a pattern to make the next predictions.

- Unsupervised algorithms: The algorithms are able, from a database, to identify patterns that are sometimes not obvious. They are used to categorize data. An example of an application is algorithms on sales platforms that classify the customer based on their buying behavior; or the algorithm that detects spam emails based on recurring characteristics of those emails.

- Reinforcement learning algorithms: The algorithms perform a prediction, similar to the supervised algorithm, but here there is a positive or negative response to the prediction made. Considering the example of weather forecasting, mentioned in the supervised algorithms, there is a response from the environment that proves that the prediction was correct or incorrect. It is from this feedback from the environment that the model can improve its predictions.

BUSINESS AND DATA UNDERSTANDING

In the automotive business the prediction of specific consumption of the vehicle based on some variables can be a competitive advantage to take decisions earlier. The nature of the problem at hand is the use of historical vehicle data to predict the specific consumption of a new vehicle. In the present work, the solution is focused on the use of supervised algorithms.

Within the supervised algorithms it is possible to find two subdivisions: the regressions and the classifications. Classifications, as their name suggests, are used when the value of the prediction is discretized into groups. For example, when predicting whether a plant's color is pink or green from the length and width of its leaves, there are only two discrete levels to the result: pink or green. When the result of a prediction has continuous values, then regression algorithms are used.

Thus, the algorithm used is supervised and regressive because the specific consumption values vary continuously.

DATA PREPARATION

The database used as input for the machine learning model was extracted from vehicle registrations from January 2018 until May 2022, considering only combustion vehicles. The base presents some characteristics of the vehicles, such as:

- Vehicle mass
- Fuel (flex or gasoline)
- Cylinder capacity
- Number of cylinders
- Power
- Transmission type
- Number of gears
- Turbo
- Fuel injection type (direct or indirect)
- Start/stop feature
- Variable air conditioning compressor feature
- Gear change indicator feature
- Tire pressure monitoring system

The base presents numerical variables, such as the mass or power of a vehicle, but also categorical variables, which cannot be represented by a number continuously, as they have a limited amount of values that they can assume. Within the categorical variables it is possible to make another distinction, the dichotomous categorical variables, which assume two values, as is the case of the presence of turbo, the polytomous categorical variables, which can assume more than two values. To identify the variable types is very important because the treatment given to them must be different, depending on it.

The database also contains the specific consumption values in mega joules per kilometer, obtained from the validation tests. Emission laws which establish the Route 2030 program, provides a bonus for energy efficiency values from the use of technologies that provide efficiency improvement, but that cannot be measured during the test protocol. These technologies are known as "off-cycles."

Thus, these technologies admittedly do not provide any change in the specific consumption value during the validation tests and therefore will not be considered in the model, since their influence is assumed by the law and not obtained from the measurement. Since machine learning models aim to predict behavior based on the measured data, these variables would harm the model.

So that the variables bellow were not considered in the machine learning model:

- Start/stop feature
- Variable air conditioning compressor feature
- Gear change indicator feature
- Tire pressure monitoring system

CORRELATION AND VARIABLES

Before developing the prediction algorithm, it is necessary to understand whether the predictor variables correlate with the response variable. In addition, it is imperative to understand whether the correlation is statistically significant.

For this, a heat map was performed that calculates the correlation between all the variables of the system, as shown in Figure 2.

The correlation between the variables was established using Pearson's method, which is a measure of the linear association between two variables. Pearson’s correlation coefficient(R) measures the variance shared between two variables, that is, increase or decrement of a unit in variable X generates the same impact in Y. Pearson's correlation ranges from -1 to 1, with the two extremes being the perfect correlation. It is important to know how to interpret the intermediate correlation values, but this interpretation is not unanimous. For Cohen [2], values between 0.10 and 0.29 can be considered small; scores between 0.30 and 0.49 can be considered as average; and values between 0.50 and 1 can be interpreted as large. Dancey and Reidy [3] point to a slightly different classification: r = 0.10 to 0.30 (weak); r = 0.40 to 0.6 (moderate); r = 0.70 to 1 (strong). Analyzing the heat map, it is observed that, in general, the variables present a high correlation, except for the "number of gears", which can be medium or weak, depending on the author.

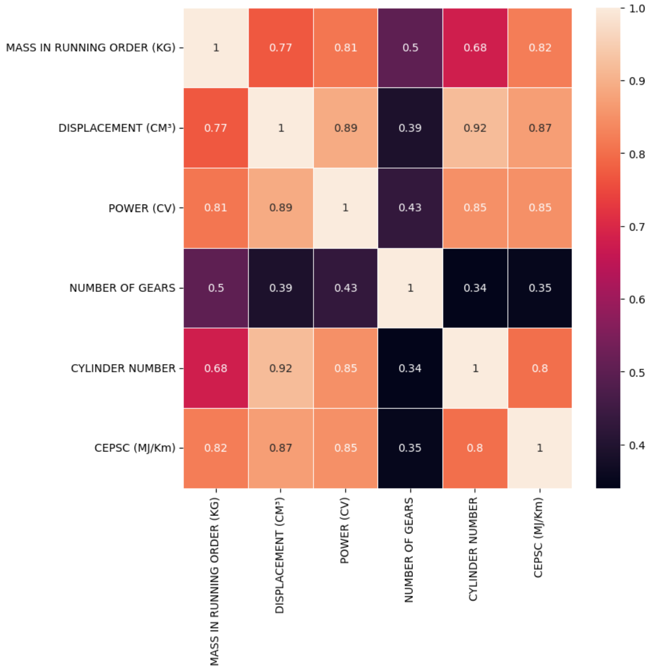

Figure 2. System variables heat map correlation

Another important fact to be observed is the correlation between the predictor variables. It is possible to observe that the variables "displacement" and "cylinder number" have r=0.92 of correlation coefficient, that is, the variables are very close to representing the same trend. Thus, for the purpose of simplifying the model, one of the two was chosen. The variable "displacement" was maintained because it had a higher correlation coefficient with the response variable.

It is important to mention that some of the predictor variables did not undergo a correlation analysis, namely: type of transmission, presence of turbo, presence of direct injection, so that these variables are categorical and therefore it is not possible to establish a linearity relationship between them.

In addition to the correlation coefficient, it is necessary to obtain the significance of this correlation, also known as p-value. This is due to the presence of two hypotheses. Hypothesis 1 states that the correlation happened by coincidence. Hypothesis 2 is contrary to 1. To identify which hypothesis is correct, we evaluate whether hypothesis 1 is correct, evaluating what percentage of chance an event results in the given value of the correlation coefficient. Dancey and Reidy (2005) point out that a value less than 5% is considered significant. Table 1 shows the significance between the variables.

The p-valeu obtain for all variables is 0%. This means that, due to the amount of samples, there is no chance that the correlation happened by coincidence.

Variable	Pearson correlation coefficient in relation to CEPSC	P-value
Mass in running order	0.82	0%
Displacement	0.87	0%
Power	0.85	0%
Number of gears	0.35	0%

Table 1. Significance between the predictor variables and the response variable.

UNDERFITTING AND OVERFITTNG

The goal of supervised algorithms is to perform prediction from new (never-seen) data, based on the pattern of historical data. Most of the time, when developing a model, you have only one database. Therefore, to test and efficiency of the model in predicting data never seen, a division is made in the database. One part of the samples is separated to train the model and the other to test it. In general, the 70/30 rule is used – 70% of the data for training and 30% for testing.

There is great importance in performing tests on the algorithm using data never seen before. The justification of this importance is based on the concepts of underfitting and overfitting. Underfitting occurs when the model cannot learn the patterns from the historical data well enough and therefore has low accuracy in the predictions. Overfitting is the opposite behavior. The model learns too much from the historical data, so it doesn't identify trends, but copies the exact pattern from the training data. A priori this does not seem like a problem, because the accuracy of the model with the training data is very good. The problem occurs when new, never-before-seen data is provided to the model. At that moment its accuracy drops. That is, the model with overfitting presents low bias, but high variance.

When creating a machine learning model, it is important to note the trade-off between bias and variance. It is important to have an accurate model, but one that is also generalist to accept new input data and answer it accurately. The Figure 3 displays a chart with a visual explanation of trade-off between bias and variance.

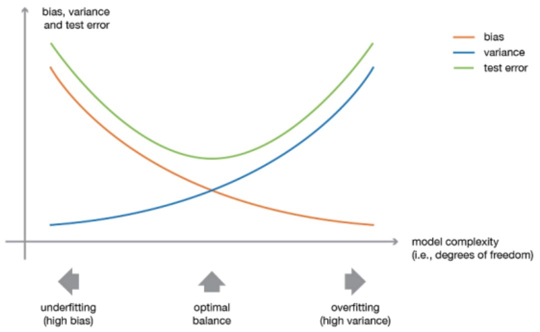

Figure 3. Trade-off between bias and variance [4]

ALGORITMOS DE MACHINE LEARNING

RANDOM FOREST REGRESSOR

Random forest is one of several machine learning algorithms. It is so called because it uses a combination of decision trees to make its prediction. The tree uses some variables to separate into more branches. Variables are chosen from a cost function, which determines which variable and which value maximize the correlation with the response variable. Figure 4 exemplifies in a didactic way a decision tree.

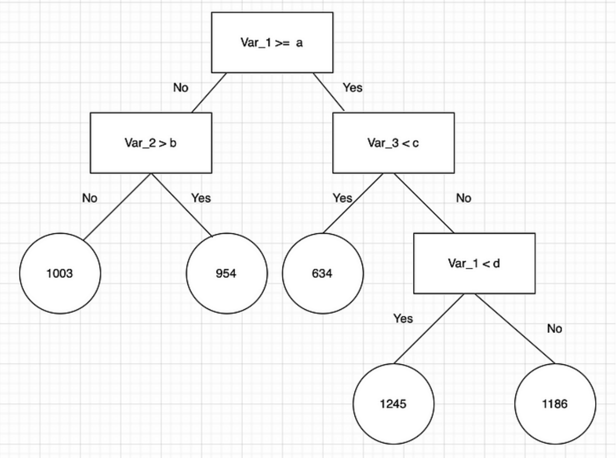

Figure 4. Decision Tree - example.[5]

Random Forest aggregates several decision trees, which are trained differently but using the same database. This concept is called Ensemble Learning and consists of aggregating several weak learners to form a robust predictor. Random Forest uses the bootstrap aggregating method, a form of ensemble learning that creates parallel and independent predictors. In short, it creates decision trees that are specialized in a specific part of the prediction, so that together they can handle all the variability of the database. Figure 5 exemplifies the bootstrap aggregating method, also called Bagging.

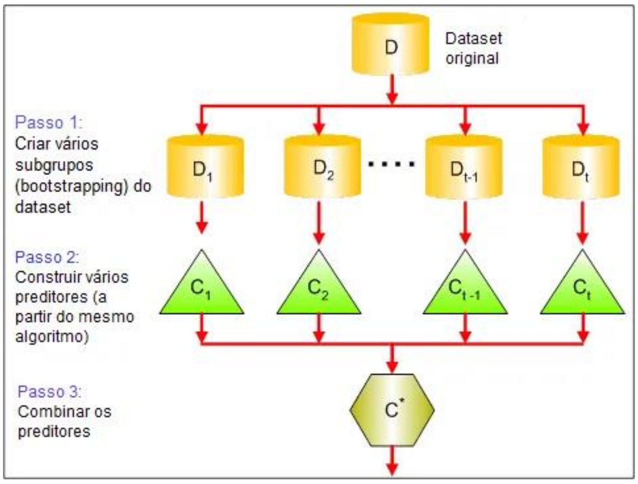

Figure 5. Bootstrap aggregating method [6]

Some parameters are important when creating a Random Forest model, such as:

- n_estimators: the number of decision trees the model will have.
- max_depth: the maximum possible depth of each tree.
- max_features: the maximum number of features the model will consider when determining a split.
- min_samples_split: the minimum number of samples needed to create another branch in the tree.

GRADIENT BOOSTING REGRESSOR

The Gradient Boosting regression method also uses decision trees and the concept of ensemble learning to create strong predictors. The crucial difference from it to the random forest is in the ensemble learning method used. While Random Forest uses the bagging method, which creates independent and parallel trees, the gradient boosting creates sequential predictors, which improve the performance of its previous one, using a method known as Boosting.

The method boosting initialization considering the prediction as the mean of the response variable. From the mean the residuals are generated, which is the difference between the real value and the mean. The method creates decision trees from the residuals obtained. With the tree created, the value of the perdition is updated: what was once a simple average of the values, now considers the added mean of the residual multiplied by the learning factor. This iteration between residual value and predicted value is performed several times, ensuring that at each iteration, the model takes small steps towards the reduction of the residual and consequent accuracy in the prediction. Figure 4 illustrate the gradient boosting model.

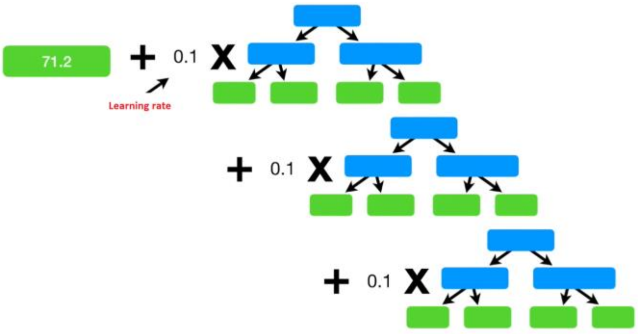

Figure 6. Gradient boosting method [7]

Some parameters are important when creating a Gradient Boosting model, such as:

- Learning rate: changes the contribution of each tree in the model. Varies between 0 and 1.
- n_estimators: the number of decision trees the model will have.
- max_depth: the maximum possible depth of each tree.
- max_features: the maximum number of features the model will consider when determining a split.

LINEAR REGRESSOR

Linear regression is one of the simplest and most well-known methods of regression in existence. It attempts to find a linear relationship between the predictor variables and the response variable. Because there are several predictor variables, the algorithm becomes a multiple linear regression because it identifies the linear relationship between each predictor variable and the response variable.

Precisely because it is a linear regression, it presents limitations for prediction when the relationship between the variables does not occur in a linear way. Still, it is a quick and simple to run model, which can offer good answers depending on the problem analyzed.

RESULTS

The same training and test data were used to evaluate the three models, Linear Regression, Random Forest, and Gradient Boosting.

The metric used to evaluate the accuracy of the models were the root mean square error and coefficient of determination, known as R^2 score. The R^2 score ranges from 0 to 1 and can be understood as the percentage of the response variable that can be determined from the predictor variable. The closer to 1 (which is equivalent to 100%), the more "parts" of the response variable can be obtained by the predictors.

Table 2 shows the comparison of results between three methods.

	Multiple Linear Regression	**Random Forest**	**Gradient Boosting**
Mean square error in train	0.176752	0.151589	0.099710
Mean square error in test	0.182307	0.177481	0.154472
R^2 score in train	85.4598%	89.3051%	95.3727%
R^2 score in test	83.2295%	84.1055%	87.9595%

Table 2. Comparison among Multiple Linear Regression, Gradient Boosting and Random Forest

Analyzing the coefficient of determination of the test data, it is possible to notice that the gradient boosting obtained the best performance. But it can be observed that there is a difference greater than 7% between the training and test coefficient, which tests an overfitting of the model.

Random Forest also showed a bit of overfitting, with a 5% difference in training and test performance.

The Multiple Linear Regression had the lowest test determination coefficient, but also has the lowest overfitting, about 2 percentage points of difference.

It is expected that the values of R2 and mean square errors are higher in the test than in the error phase and the values obtained are satisfactory for the objectives of trend investigation that motivated this work.

CONCLUSION

The results show that machine learning algorithms have a great potential for application in various types of problems.

The algorithms that are supervised need historical data processed so that they can make the correct interpretation of trends.

It is imperative to understand the trade-off between bias and variance to obtain a model that is accurate and at the same time generalist, accepting input data never seen before.

In the prediction of energy consumption from the vehicle parameters, all three algorithms showed a small percentage of overfitting, even after fine-tuning the parameters. It is understood that the gradient boosting presented a level above 5 percentage points and therefore should be discarded. Thus, both Random Forest and Multiple Linear Regression presented results that are acceptable.

It is considered that this work, in addition to achieving it's initial goal of providing a tool for predicting the energy consumption of a vehicle based on a few characteristics, also evidenced the potential of machine learning applications as a tool that can be used as guidance for project decision-making at initial levels, especially in a comparative context.

The continuity of this work leads to the introduction of other parameters of fundamental contribution to energy consumption can be inserted, such as aerodynamic drag coefficients, tire characteristics, among others. The evaluation and adoption of special strategies to include the effects of categorical variables should be done.

REFERENCES

[1] Source: https://www.the-modeling-agency.com/crisp-dm.pdf

[2] Cohen, J: Statistical Power Analysis for the Behavioral Sciences,2nd Edition, Imprint Routledge ,June 1988,DOI https://doi.org/10.4324/9780203771587

[3] Dancey, C.P;Reidy,J: Statistics Without Math for Psycology, 2nd Edition, Pearson Education, 2005

[4] Roca, J. Ensemble methods: bagging, boosting and stacking. Available at: <https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205>

[5] Beheshti, N. Random Forest Regression. Available at: <https://towardsdatascience.com/random-forest-regression-5f605132d19d>

[6] Neves, E. C. Modelos de Predição | Ensemble Learning. Available at: <https://medium.com/turing-talks/turing-talks-24-modelos-de-predi%C3%A7%C3%A3o-ensemble-learning-aa02ce01afda >

[7] Adapted from Stromer, J. Gradient Boost: Regression Main Ideas. Available at : < https://www.youtube.com/watch?v=3CC4N4z3GJc >