Report - XPorters Data challenge

Florian Bertelli, Gaspard Donada--Vidal, Gérémy Hutin
Ghassen Chaabane, Moez Ezzeddine, Ziheng Li

Website - Repository - Video

***Abstract* - This report presents the supervised machine learning challenge that we designed and published on Codalab.lri.fr. This challenge falls under the theme of the future of transportation. The competitors are given labeled data that they must use to train a regression model.**
In this report, we present the context and the material we will provide. We show that the data we chose is meaningful. We propose a model to solve the problem, and, we discuss our results to make sure that they are adequate.

1. Background

With globalization, our world tends to be more and more connected, so this implies more and more transport.

There are various types of transports: transports of goods, energy, information, or people.

This century will be marked by a revolution in the transport of people. With the development of autonomous cars, more and more data will be collected: speed, location, population, etc…

The challenge is to get value from this. As the number of cars, and particularly autonomous cars tends to grow, we'll need to deal with an increasing traffic flow to avoid huge traffic jams.

Indeed, some experts pretend that autonomous vehicles will be able to reduce travelling time up to 30%, even if the world traffic increases of 10%.

But how is it possible to reduce travelling time and traffic jams if the number of vehicles increases ? Thanks to prediction.

In fact, prediction will be the key to determine the fastest way to get you from your home to your work, without getting in traffic jams. Predictions may also be used to determine which transportation infrastructures to build.

Xporters challenge is a small standard multivariable regression data set from the UCI Machine Learning Repository, formatted in the AutoML format. It uses a data set concerning the traffic volume off a highway in the USA from 2012 to 2018, the date, and some informations about the weather. The aim of this challenge is to predict the traffic volume.

2. Material and method

Our data was taken from : https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume

It is a table composed of 48 204 examples and 9 features.

The data has been recorded between 2012 and 2018, in the Interstate 94, which is a highway between Minneapolis and St Paul, in the United States.

Our aim is to predict the traffic volume using the date and the weather.

Some preprocessing is needed to be able to use the date (which is presented as “29/08/2012”), or the hour.

This preprocessing is useful so we can better learn from the correlation between each interesting feature (Day, Month, Year, Hour) and the traffic volume.

For example, the traffic volume tends to vary a lot during the holidays (the densest days being the first and last days).

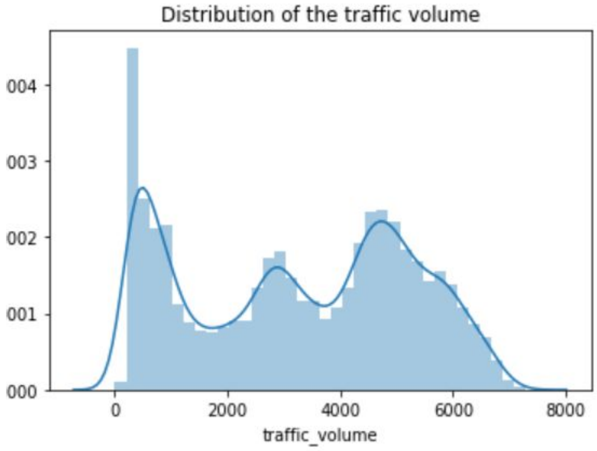
Furthermore some preprocessing on the weather is also needed to transform the features “weather main” and “weather description” into categorical features using the one hot encoding.

The one hot encoding creates a binary column for each category of “weather main” and “weather description”.

Since there are 11 categories in “weather main” and 36 categories in “weather description”, the one hot encoding will add 11+36=47 columns to our dataset and if the “weather main” is for example “Clear”, we will set 1 to the column “Clear” and 0 to all the others.

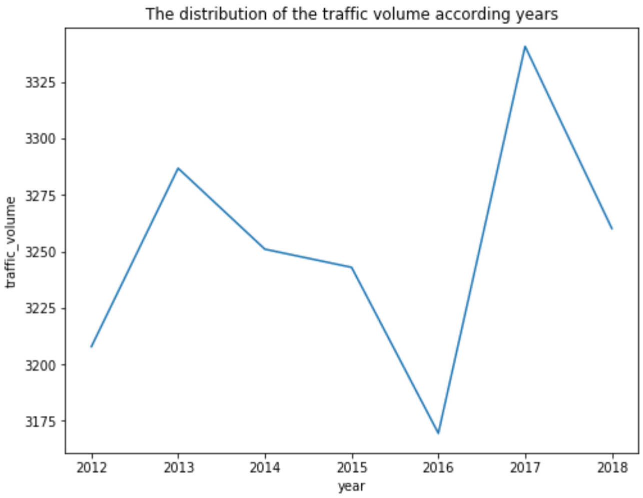
This one hot encoding technique is crucial and is needed for feeding categorical data to many estimators, such as the Random Forest Regressor.

As a first step, we need to visualize our data so we can have a better interpretation of the features.

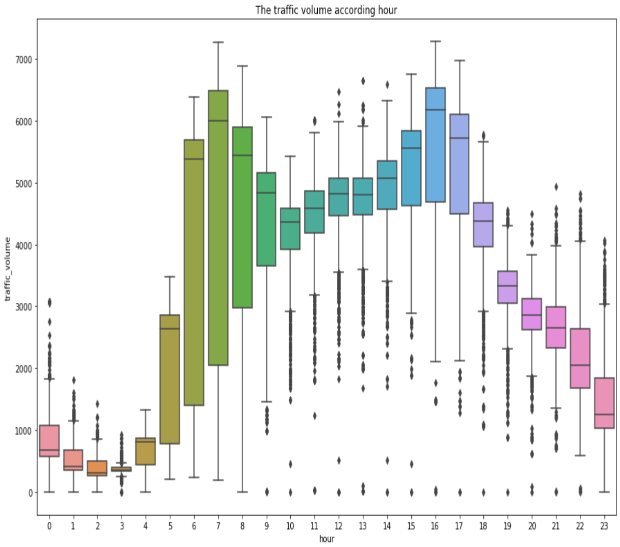


The distribution of the traffic volume confirms the fact that we will be able to do a regression (as the traffic volume is a continuous variable).

Plotting the distribution of the traffic volume according to the years shows us that the traffic volume drastically decreased in 2016. This low traffic volume in 2016 is actually due to an undergoing maintenance in the highway.



Intuitively the most important feature that seems to affect the traffic volume is the “hour”. We proceed by visualizing it using boxplots.



The boxplot confirms our intuition and shows that the traffic volume is highly correlated with the traffic volume.

Running a data correlation shows us the most important features with respect to the traffic volume.

Feature	**Correlation**
hour	0.3505
temp	0.1318
weather_main_Clouds	0.1195
weather_description_scattered_clouds	0.0839
weather_description_broken_clouds	0.0656
clouds_all	0.0642
weather_description_few_clouds	0.0443
weather_description_proximity_shower_rain	0.0340
weather_main_Haze	0.0193

Using a Random Forest Regressor seemed very appropriate for this data. It’s hard to overfit and provides a small variance. Moreover, as random forests are very flexible and don’t require a lot of treatment on the input, they seemed to be a good choice for testing our challenge. Finally, the fact that they are time-consuming was not really a problem in our case.

We split the data into three parts; 80% for training, 10% for validation and 10% for testing.

3. Results

We evaluated the performance of our model using the R2 metric, which is defined as:

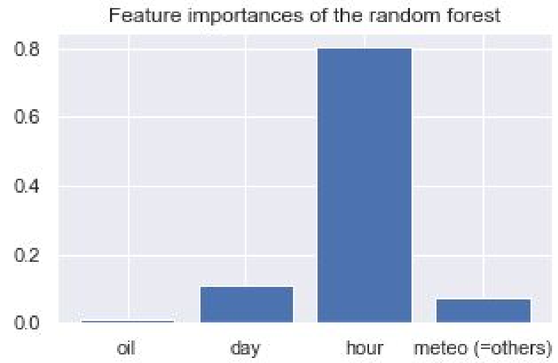
$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

We chose this metric because it is recognized as a very simple and good method to evaluate the goodness of fit of a model, as the goal is simply to minimize the upper term of the division. In other words, in order to beat our challenge, the model will have to perform a least-squares regression.

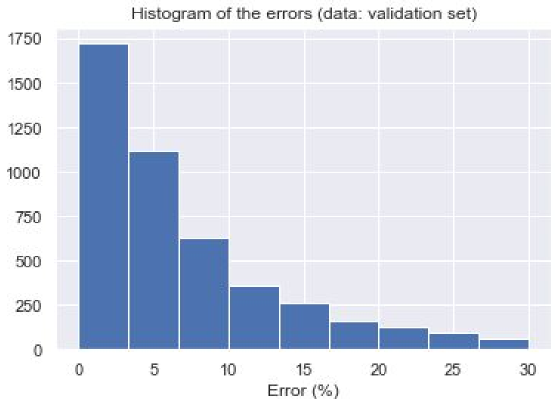
The training was performed using 38563 rows. We then cross-validated with 4820 different rows.

The score of the 5-folds cross validation is 0.950, with a standard deviation of 0.0014.

The importance of each feature for the model is given below:



Furthermore, we calculate the relative error for each prediction of the validation set. We obtain the following repartition:



4. Discussion

In terms of performance, the model worked quite well. 0.95 is a very good score that definitely proves that the data is meaningful. Hopefully, our tests seem to demonstrate that different models would yield different results (ranging usually from 0,80 to at least 0,95), which is important because the challenge would be pointless if everybody was to get the same score.

Moreover, the feature importances chart shows clearly that, although the hour of the day is by far the most important feature, others are important too, especially the day and the week and the meteorological data.

It's also reassuring to see that, as planned, our "fake" oil price column doesn't have any impact.

Finally, we believe that a neural network could yield better results than what we found. Indeed, our dataset is huge and usually, the score of random forests and other standard models converge fast, while a neural network could benefit immensely from all of these samples.

5. Conclusion

For many of us, this project was our first time designing a machine learning competition.

Throughout the whole process, we discovered and learned a lot about data visualization.

In fact, visualization helped us understand the raw data. We then proceeded to cleaning it by removing irrelevant information and formatting it using techniques such as the one hot encoding, in order to feed it to our regression estimator.

We also studied Regression algorithms in order to pick the best one that fits to our problem and scores the best with respect to the R2 metric.

Besides the acquired technical knowledge, we also learned a lot while working in group.

In summary, the amount of new concepts learned while working on the Xporters challenge was definitely beyond our expectations.

Acknowledgments

This project was developed for the M2 AIC class.

We thank our colleagues from the Medichal team who tested the competition and provided insight that greatly assisted this project.

We would also like to show our gratitude to Isabelle Guyon for guiding us and sharing her pearls of wisdom with us during the course of this project.

A special acknowledgement also goes to Adrien Pavão who assisted us and was always present to give feedbacks that helped us improve the competition.

References

[1] Sandro Tosi, “Matplotlib for Python Developers”.

[2] Chris Mofitt, “Choosing a Python Visualization Tool”.

[3] Jessica Hamrick, “Creating Reproducible, Publication– Quality Plots with Matplotlib and Seaborn”

[4] L. Breiman, “Random forests” Machine Learning, vol. 45, no. 1, pp.5–32, 2001.