ML
TD1 – feature engineering

An online merchant asks you to make real-time predictions for the customers of his website: when they are ready to buy, tell them the date of receipt of their product, for the various possible transport services. (Chronopost..).

To do this, you have 6 weeks of order history with 3 pieces of information:
- Date and time of the customer's order
- Shipment date (to simplify considered as order shipment date)
- Transport service

In production
- You will have a daily update (every night) of the warehouse status (this same updated file)
- You will need to be able to deliver real-time on-the-fly prediction with each new online order. The speed of your predictive model will be as important as its accuracy

You have 3 hours to build your project strategy and your model

- How do you take the problem?

Do you agree to make this prediction? if so, with which commitment (s)?

Yes, I agree to take the problem. The solution will be presented in the form of a notebook with a technical solution and an indication of the efficiency of the model used. The goal is to give the number of days between the two dates.

- What will be your strategy ?

Target: how to predict a date?

We'll no longer predict a date but a delay

Classification or regression (we are looking for a discrete numerical value ...)

I basically wanted to go towards a regression because we predict numbers (how to know the difference between two dates??) but after reflection the classification is better especially since the target will be in the form of a discrete variable

How to select the test set?

We will take 60% of the data for learning and 40% for the test. This will be selected randomly because the dates are given ordinally on our dataset.

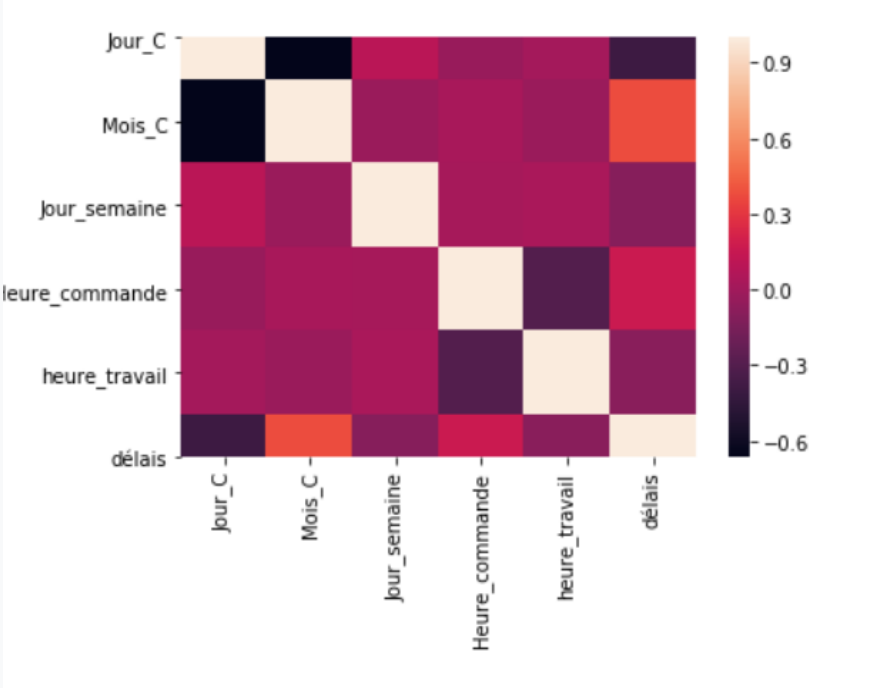
- ➔ Create new features with existing features to perform other type of learning.
- ➔ By creating the variable Delay, we will create our target.
- ➔ Create a weekday variable.
- ➔ Identify Weekdays and Weekends/Holidays

- **Data Discovery:**

Visualize the flows

We can notice that there are shipping dates prior to the order date which is logically impossible so we will perform a data cleanup.

We'll set up a heatmap to study the correlation between the variables.



Make observations

Here are some examples of observation:

We can notice that there are 3 170 orders which have a delay up to 1 week between the order and shipping dates.
And we can notice that the provider service of these orders is 7, 21, 42 et le 48.

We can notice that there are 14 464 orders which have a delay time of 0 days between the order and shipping dates. After that, we try to observe which are the provider service of these orders and it's mainly 48, 21 and 16.

We then try to find out the minimum, maximum and average time delay of each provider service.

- **Feature engineering**

How to enrich the model with new features? Knowing that no other features can be given to you

We separate the order into order date and order hour.
Then we separate the order date into year, month, and day.
We add weekly days in our dataset

- **Model selection**

Test different algorithms

We have built different models such as:

- linear regression
- logistic regression
- random forest
- boosting
- KNN
- NaiveBayes

Visualize your results

We have visualized our results by having accuracies of different models build up and therefore we can say the best accuracy is the accuracy of random forest model.

Analyze your important variables

Deduce in new features?

- **How do you go to production?**

You receive in production only a date and time (real order time) and not from the transport service (not yet chosen by the customer): with this only information and the update of the orders shipments from the warehouse updated during at night you must predict the calendar dates associated with each transport service (the customer will choose)

How to reduce the response time?

-