

Explicability

In this TD (1.5h) you will have to do a small project and it will be submitted by the end of next week.

Question 1.

Form groups of 4-5 people.

Question 2.

Select a dataset on Kaggle that seems interesting for the group. The dataset must contain explanatory variables that can be interpreted.

Question 3.

Let's begin with some visualizations.

- Based on your intuition, what are the most important variables for predicting your target variable?
- Create visualizations that take into account the type of variable (categorical, numerical) that shows the correlation of the features you selected.
- Don't forget to clean your dataset as needed in order to deal with missing values and outliers.

Question 4.

Let us do a linear regression.

- Do the variable transformations that are needed (on hot encoding, log etc...)
- Fit the linear regression
- Interpret the coefficients of your regression
- Does any coefficient contradict your intuition?

Question 5.

Let's build a more complex model

- Build the "best" model you can using traditional data science (no deep learning)
- Don't hesitate to fine tune it, try different combination of variables etc
- What is the best metric you can reach? Can you consider it "good"
- Given your results what use case can be done with this model?

Question 6.

Using shapley values, explain your model.

- Start by giving some forceplot examples
- Use the feature importance and identify the most important variables
- Compare them to the variables identified in the linear regression
- Using the beeswarm plot explain the model behaviour for the most important variables
- Use dependance plots in order to understand the behaviour of the most important variables