1

Data Exploration & Preparation: General Intro.

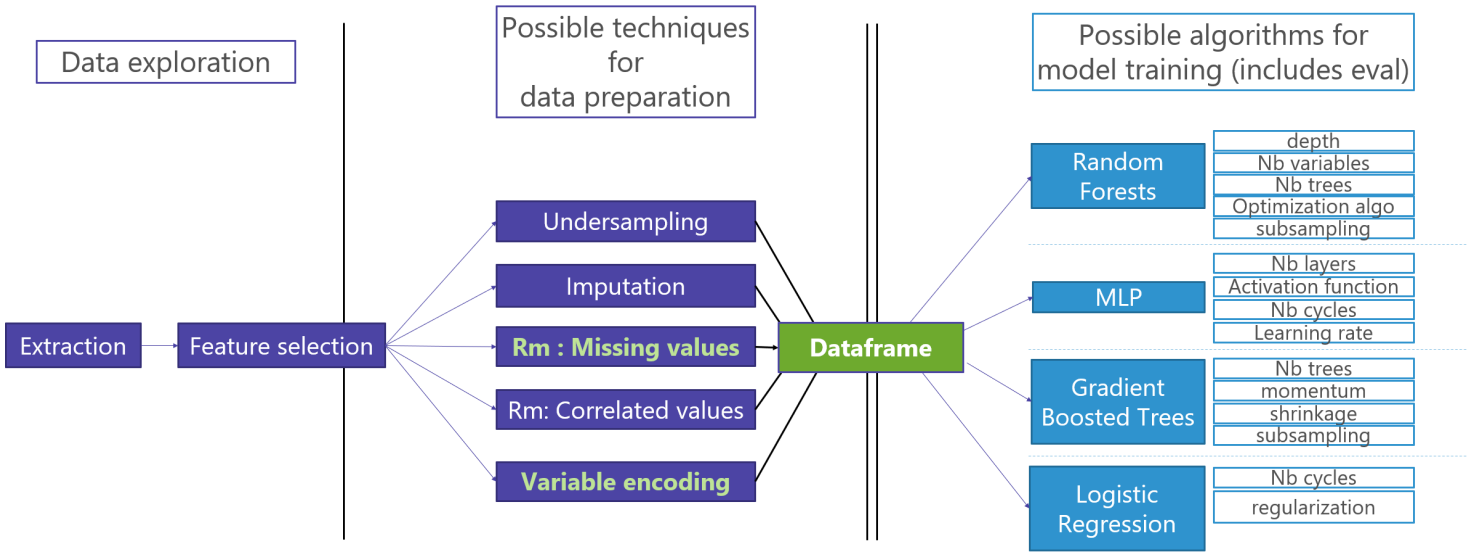
Common problems and solutions when dealing with data science projects?

- Datascience requires and experimental effort
- It has a lot of impacts on the model performances
- Data Preparation requires some knowledge about data format and their constraints
- It can be a mess to organize in an optimal way

Data Explo. & Preparation: common workflow

2

What are the possible steps for data exploration that will lead to a model building?



Problem: too much combination!

3

- Obviously some combinations will be avoided
- Still several transformations tries will be attempted
- The execution is not linear!

4

Preparation? Transformation? For what?

- What are the data types that a statistical algorithm will be able to ingest?
- Remember those main data types
  - Numerical: Integer or Float
  - Categorical: Ordinal, Nominal, or Boolean
- This is the only types of data that an algorithm can take

5

Preparation? Transformation? For what? (2)

- Algorithms want "rectangle" data
- Where data are organized under the form of matrix
- Where **rows** are instances or observations
- Where **columns** are variables or features
- Where one feature will be the target for predictions
- What we want is a **dataframe**

So now: why this course?

- Preparation cannot be made without a bit of software engineering
- Differences between Data Scientist and AI Engineer
- We have to handle different kind of sources

  (Json/XML/CSV/ARFF/Databases/REST Services/Web pages...)

Data exploration: tools and methodology

- Jupyter NoteBook? globally good, but can be tricky! Why?
- what do you think? Is this only a matter of tools?
- What should be the characteristics of the retained methodology?

Data exploration: Orange

- Orange installation and demo

Your turn : Data exploration (from a disaster

- load the following file in orange: dataset.csv