Analytics Vidhya
Learn everything about analytics

DATAHACK
SUMMIT
2017

AUTOMATED MACHINE LEARNING

Talk

By

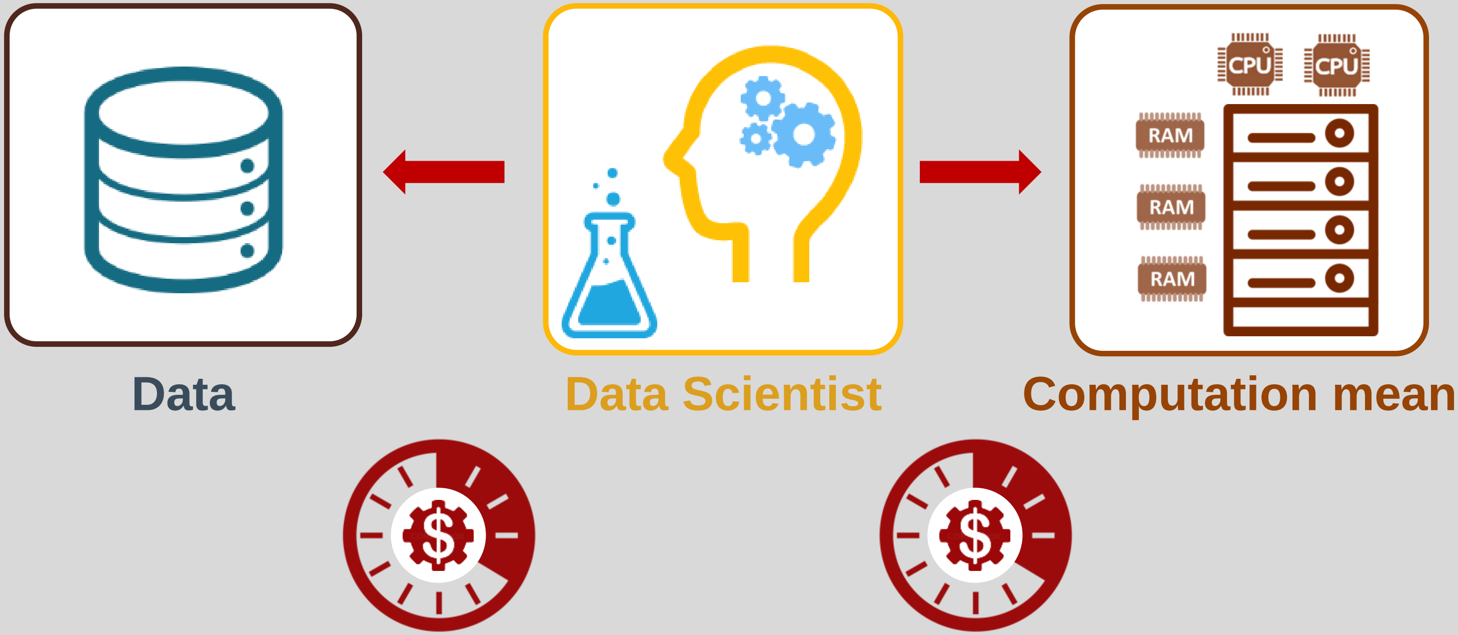
Axel de Romblay

AUTOMATED MACHINE LEARNING

- Why Auto-ML is gaining momentum ?
- Focus on the automation process
- Overview of various solutions
- Conclusion

Machine Learning

Almost an automated process...



Data

Data Scientist

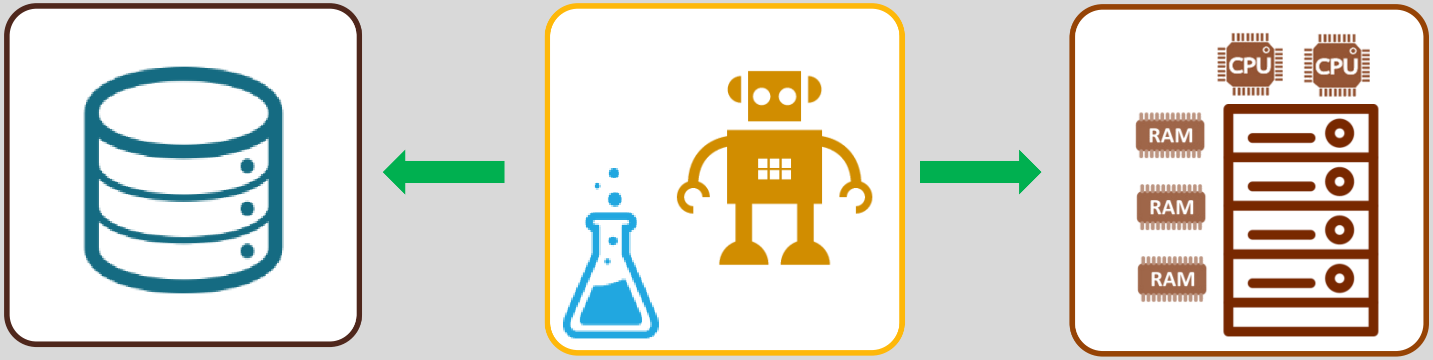
Computation means

Data pre-processing

Model tuning

Auto Machine Learning

A fully automated process



Data

Robot

Computation means

✓

- **Structured data**
  - *csv files*
  - *json files*
  - …

- **Supervised tasks**
  - *classification*
  - *regression*

✗

- **Unstructured data**
  - *images*
  - *texts*
  - …

- **Unsupervised tasks**
  - *outlier detection*
  - *clustering*
  - …

What is auto-ML ?

We want to automate…

…the maximum number of steps in a ML pipeline…

…with minimum human intervention…

…while conserving a high performance !

Some companies struggle to deploy ML models

« I need to recruit **more Data Scientits** ! »

« I have a model ! **Now what** ? »

« Machine Learning **cannot replace a human** »

Others don't…

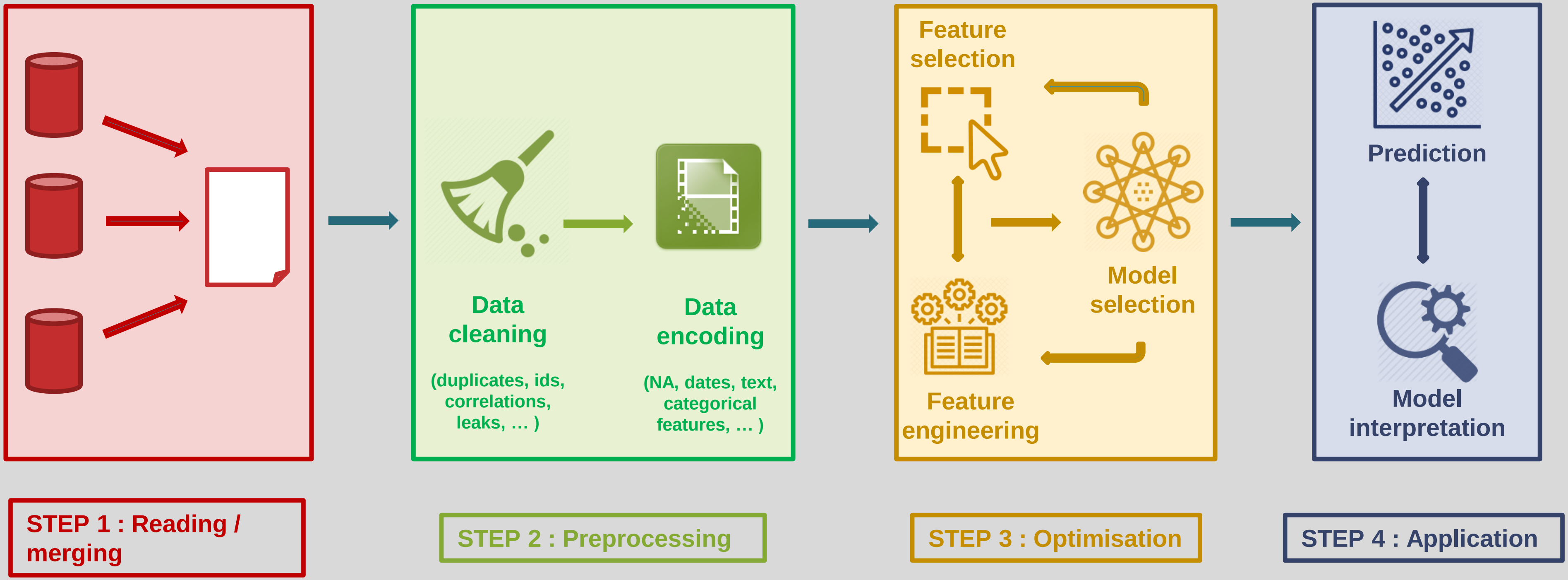
Identify a **generic** and **repetitive/reusable** pipeline

Implement or **buy** or **use** automated tools

Deployment and **monitoring** : A&B testing, …

Focus on the automation process

Diagram of a standard ML pipeline



Data cleaning

(duplicates, ids, correlations, leaks, …)

Data encoding

(NA, dates, text, categorical features, …)

Feature selection

Model selection

Feature engineering

Prediction

Model interpretation

STEP 1 : Reading / merging

STEP 2 : Preprocessing

STEP 3 : Optimisation

STEP 4 : Application

Step 1: reading and merging

Maybe the hardest step. Not a priority for auto-ML at the moment.

Some inputs : paths to the data + target name

Difficult to auto-merge different sources

Step 2: preprocessing

Not all packages tackle preprocessing

No inputs : heuristics to detect feature types are easy

Naive encoding for most auto-ML packages

Step 3: optimization

Top priority for auto-ML community

Some inputs : scoring function + a hyper-parameter space

Computation time can be long !

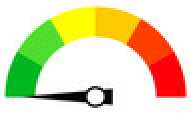
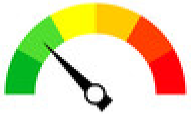
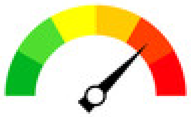
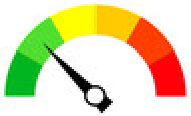
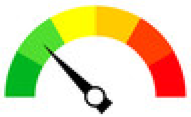
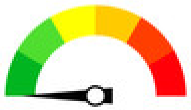
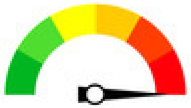
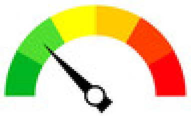
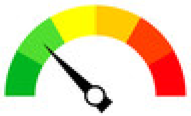
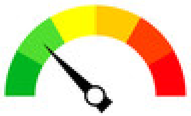
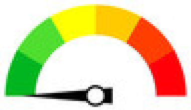
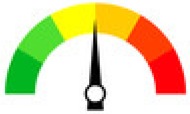
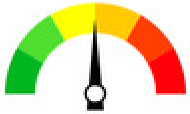
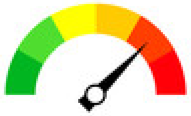
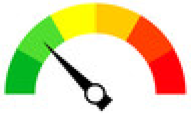
Step 4: application

Prediction is also a priority for auto-ML

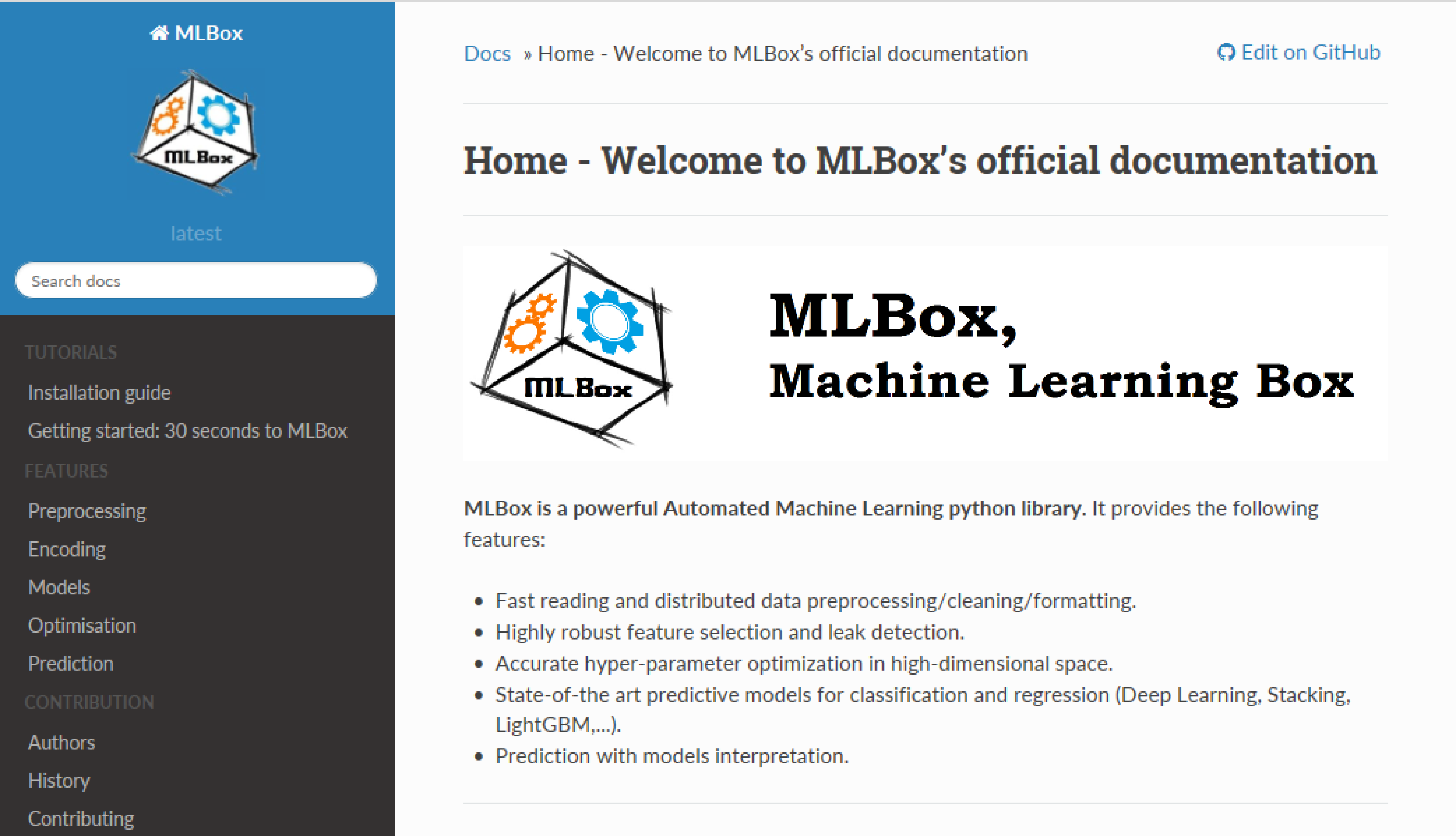
No auto-monitoring after model deployment

Fast and accurate step

Overview of various solutions

	DataRobot	Auto-Sklearn	ClimbsRocks / auto_ml	TPOT	MLBox
Quality of automation					
Automated steps	- Reading - Encoding - Optimisation	- Encoding - Optimisation	- Encoding - Optimisation	- Optimisation	- Reading/merging - Cleaning - Encoding - Optimisation
Maintenance					
Ease of setup					
Open source ?	✗	✓	✓	✓	✓

MLBox: a fully automated package



MLBox

latest

Search docs

TUTORIALS

- Installation guide
- Getting started: 30 seconds to MLBox

FEATURES

- Preprocessing
- Encoding
- Models
- Optimisation
- Prediction

CONTRIBUTION

- Authors
- History
- Contributing

Docs » Home - Welcome to MLBox's official documentation

Edit on GitHub

Home - Welcome to MLBox's official documentation

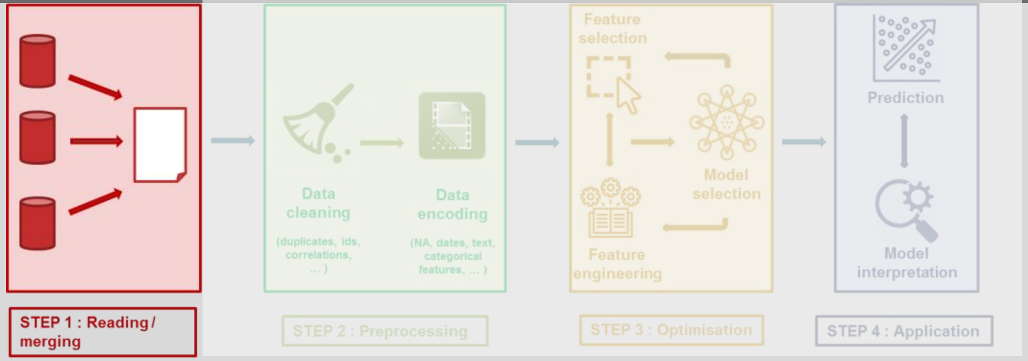
MLBox,
Machine Learning Box

MLBox is a powerful Automated Machine Learning python library. It provides the following features:

- Fast reading and distributed data preprocessing/cleaning/formatting.
- Highly robust feature selection and leak detection.
- Accurate hyper-parameter optimization in high-dimensional space.
- State-of-the art predictive models for classification and regression (Deep Learning, Stacking, LightGBM,...).
- Prediction with models interpretation.

Features

what is Automated by MLBox ?



STEP 1 : Reading / merging

From a several **raw datasets** to one **structured dataset.**

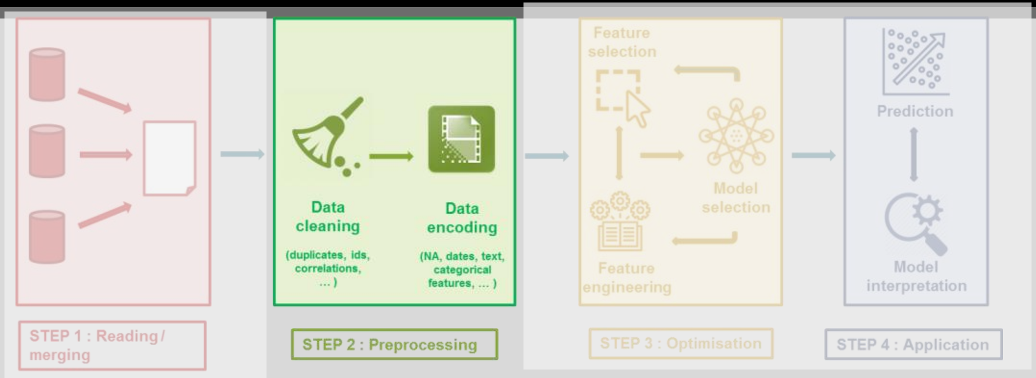
- List of paths to all the datasets
- Target name

- **Reading** of several files (csv, xls, json and hdf5)
- **Auto-merging** of all the sources – information crunching
- **Task detection** (binary/multiclass classification or regression)
- **Split** between **train** and **test** sets
- Basic **information display**

- Dense structured train and test files (without duplicates)

Features

what is
Automated by MLBox ?



STEP 2 : Preprocessing

From a dirty dataset to a **cleaned numerical one**

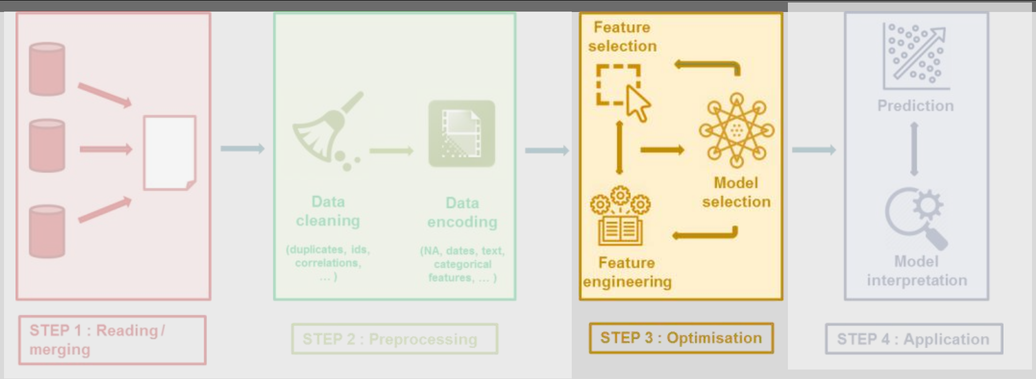
- A dataset

- **Auto-cleaning/dropping** : duplicates, drifts / covariate shifts (*), high correlations, highly sparse features / samples, constants, …
- **Feature encoding** : missing values, lists, dates, categorical features, text, … - *SEVERAL STRATEGIES*

- A cleaned dataset with numerical features

Features

what is Automated by MLBox ?



STEP 3 : Optimisation

A **wide range** of models are **tested** and **cross-validated**

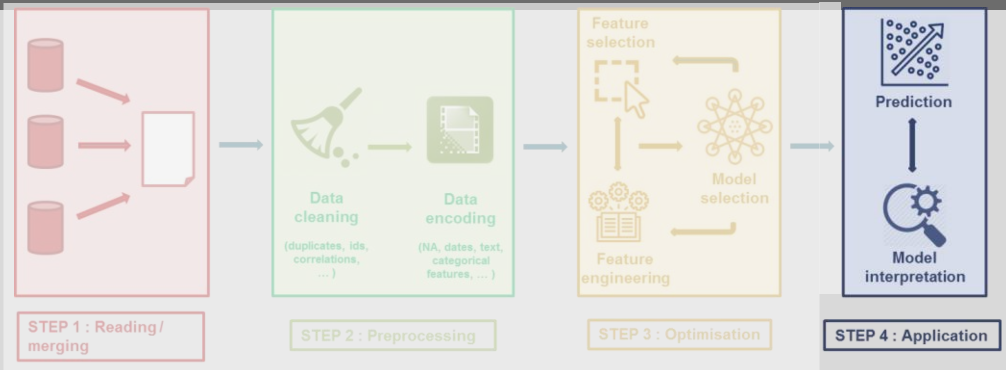
- A metric (a wide choice, otherwise can be implemented) - *OPTIONAL*
- A hyper-parameter space – *OPTIONAL*
- The train set

- **Feature engineering** : using neural networks (*)
- **Feature selection** : filter methods, wrapper methods, embedded methods
- **Model selection** : a wide range of accurate models (LightGBM, XGBoost, Random Forest, Linear, …)
- **Hyper-parameter optimisation** : TPE (Bayesian optimisation method) – dumping of fitted pipelines
- **Ensembling** : multi-layer stacking, boosting, bagging, …

- The optimal pipeline configuration

Features

what is Automated by MLBox ?



STEP 4 : Prediction

The **best model is fitted** and **predicts** on the test set

- Train and test datasets
- The optimal pipeline configuration - *OPTIONAL*

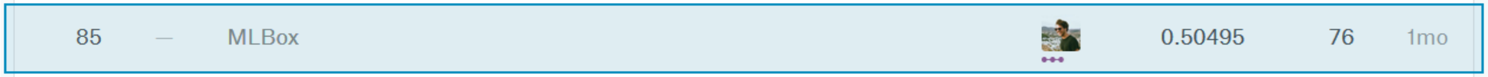
- **Target prediction** : classification and regression - dumping of predictions + optimal fitted pipeline
- **Model interpretation** : feature importances (saved)
- **Leak detection** : warning

- The predictions on the test set

MLBox: about the package

- **Compatibility**: Python 2.7-3.6, Linux OS
- **Quick setup**: $ pip install mlbox
- **User friendly**: tutorials, docs, examples…

- **Quality**: functional code : tested on Kaggle



85 — MLBox 0.50495 76 1mo

- **Performance**: fully distributed and optimised
- **AI**: dumping and automatic reading of computations
- **Updates**: latest algorithms

Conclusion: the benefits of auto-ML

Increases productivity

Repetitive tasks are **automated** and **accelerated** ! A Data Scientist can focus more on non-traditional issues !

Avoids errors

A robot never makes **mistakes**…

Democratizes Machine Learning

Machine Learning for everybody **(no coding)**

Thank you !

Questions ?