
WHY
DATA
CLEANING
IS
IMPORTANT
@learn.machinelearning

WHY DATA CLEANING??

@learn.machinelearning

- The most used sentence in the data science world "garbage in, garbage out".
- Bad data will lead to bad results, plain and simple.
- It's hard for computers to judge whether data makes sense or not...Computers are not magical boxes they are just machines.
- To get accurate results, you need to remove errors from your data which confuses the algorithms.
- It's a time consuming process but important.

WHAT ARE THE CAUSES??

@learn.machinelearning

- Input Errors: There are plenty of ways a human can enter the wrong information. They may mistype, miscalculate, or misread.
- Malfunctioning Sensors
- Mangled Data: When sensors malfunction, they are likely to generate values outside the acceptable range
- Duplicates: If the initial data set is an amalgamation of multiple sources, there is a high probability of duplicates.
- Lack of Standardization: When using multiple data sources, lack of standardization is common. To achieve true results, all data that is similar in reality must be represented similarly in the input.

IDENTIFYING PROBLEMS

- Range Constraints: typically, numbers or dates should fall within a certain range.
- Data-Type: values in a particular column must be of a particular datatype.
- Categorical Constraints
- Compulsory constraints: certain columns cannot be empty.
- Unique Constraints: a field, or a combination of fields, must be unique across a dataset.
- Cross Field Constraints: certain conditions that span across multiple fields must hold.
- By Visualizations
- Counting the Errors
- Checking Missing Values
- Set-Membership Restrictions

DATA CLEANING TECHNIQUES

@learn.machinelearning

- Removing missing data
- Direct Correction
- Scaling / Transformation
- Normalization
- Syntax errors
- Data Imputation
- Remove Duplicates
- Spell Check
- Remove Irrelevant Values
- Fix Structural Errors
- Filter Unwanted Outliers