Machine learning with dirty tables: encoding, joining and deduplicating

Jovan Stojanovic Inria, dirty_cat maintainer



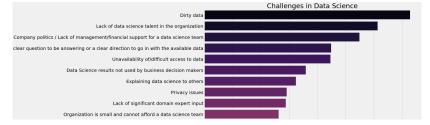


イロト イ団ト イヨト イヨト

11/05/2023

Jovan	





Source: 2017 Kaggle Machine Learning & Data Science Survey

Biggest problem: working with real world datasets.

< □ > < 同 > < 回 > < 回 >



Dirty data

Data	Issue
Paris	Clean
Pqris	Туро
РА	Abbreviation
Paris, France	Alternate
NA	Missing

How to use/represent this data for machine learning?

Jovan Stojanovic



Open-source project started in 2018 at Inria, by Patricio Cerda and Gaël Varoquaux.

- ► 1. Encoding dirty categorical variables
- > 2. Fuzzy joining tables with dirty data
- 3. Cleaning (deduplicate) dirty categorical variables

• • • • • • • • • • •



Jupyter Notebook demo

https://github.com/jovan-stojanovic/ jupytercon2023



Jovan	Stoja	novic
Jovau	Stoja	novic

• • • • • • • • • • •



- ▶ 1. Encoding dirty categorical variables with TableVectorizer
- > 2. Joining on dirty categorical variables with FeatureAugmenter
- ▶ 3. Deduplicating dirty categorical variables with deduplicate

Stay tuned! Exciting future development:

- dirty-cat will evolve into skrub: broadening scope of the project
- Including semantics and other information (word embeddings)
- Working inside databases: identifying potential joins among candidate tables.

< □ > < 同 > < 回 > < Ξ > < Ξ



For more information and examples, check the docs:



Try it! Installation:

pip install dirty-cat

Contribute or support (*) the project on GitHub:



(日)

https://github.com/dirty-cat/dirty_cat