WOMEN WHO CODE®
/data-science

Build a Data Science Project From Scratch - SESSION 3

JUNE 23
10 PM ET

Victoria Ubaldo
Data Analyst @ Labentana
Innovation Lab - Interbank

womenwhocode.com/datascience/events

Data Wrangling & Feature Engineering

1. Why data cleaning in big data is important?

2. What is data wrangling?

3. What is feature engineering?

Big Data & Data Cleaning

- In scenarios where datasets are exceptionally large, automated data cleaning becomes a necessity.

- When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled : **raw data**

Data Wrangling (Data Cleaning)

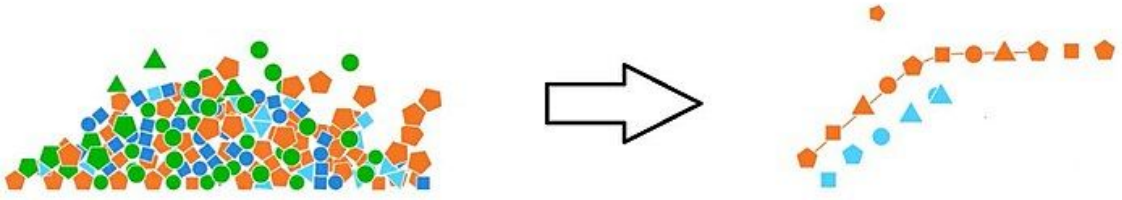
- A processes designed to transform raw data into more readily used formats.
- The methods depending on the project, the data you're leveraging and the goal you're trying to achieve.
- Tools used to clean data include:
  - **Python**
    - Pandas - data cleaning including handling missing values
    - NumPy - large, multidimensional arrays & matrices. Used with Pandas for numerical analysis and manipulation
    - Regular Expression (regex) - pattern matching & text manipulation. Can identify and remove or replace specific patterns of text.
    - Beautiful Soup - web extraction
    - SciPy - clustering, filtering, smoothing data
    - Openrefine: - user-friendly interface for tasks
  - **SQL**
    - Query specific data from a database and organize it into a useable dataset.

Why is data wrangling important?

- Garbage data in is garbage analysis out

- Incomplete and inaccurate data affects business operations.

- As data becomes more unstructured, diverse, and distributed, data wrangling becomes a common practice in organizations.



How to clean data?

1. Handle missing values
    - impute rows or columns vs drop?
2. Remove duplicate or irrelevant observations from your dataset.
3. Address inconsistencies in naming conventions, formatting & data types
4. Filter unwanted outliers
5. Fix structural errors or incorrect encoding
    - (For example, you may find “N/A”)
6. Data validation - ensure clean data is accurate and reliable
    - Hypothesis testing
    - External comparison

Example of data cleaning with Python

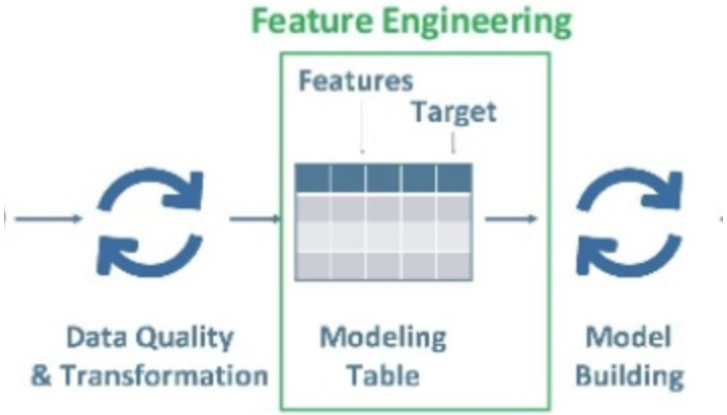
time_signature	valence	target	song_title
4	0.286	1	Mask Off
4	0.588	1	Redbone
4	0.173	1	N/A
4	0.23	1	N/A
4	0.904	1	Parallel Lines
4	0.264	1	N/A
4	0.308	1	Childs Play

```
df["song_title"].fillna("no title",
inplace = True)
```

Feature Engineering

Process of identifying & extracting relevant features from raw data for a machine learning algorithm.

The goal is simplifying and speeding up data transformations while also enhancing model accuracy.



How use Feature Engineering ?

- Brainstorm features.
- Create new features by combining or transforming existing features.
- Transform features
  - Encapsulates various data engineering techniques such as; **encoding the data, normalizing it, dummy variables.**

Example: One-hot Encoding

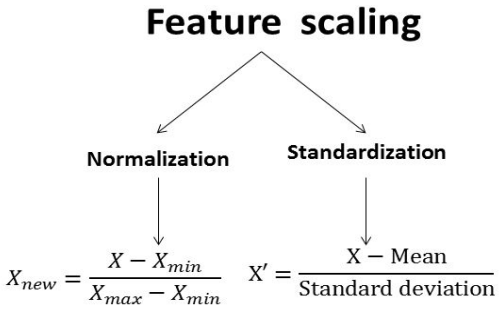
artist
Future
Childish Gambino
Future
Beach House
Drake
Drake

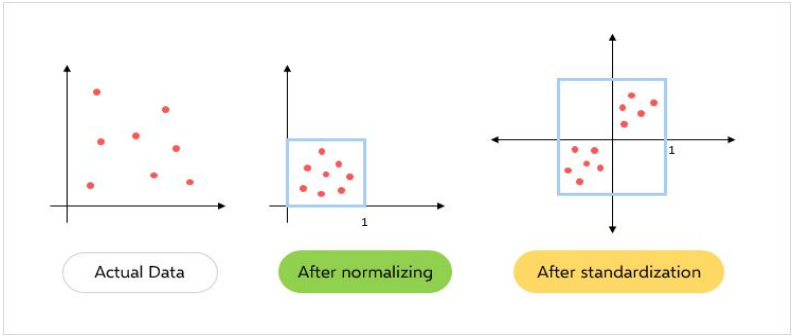
Future	Childish Gambino	Beach House	Drake
1	0	0	0
0	1	0	0
1	0	0	0
0	0	1	0
0	0	0	1
0	0	0	1

Normalize the data Vs. Standardize

Normalizing the values of the numerical features to a fixed range between 0 and 1 using the MinMaxScaler .

Standardization is a technique to transform numerical features to have zero mean and unit variance, standard deviation.





Create new features

Feature engineering can include tasks such as creating new features based on domain knowledge, combining or transforming existing features, and selecting the most relevant features for the analysis.

- Create a new column by assigning the output to the DataFrame.
- Use **rename** with a dictionary or function to rename row labels or column names.

Next Steps

- After data wrangling we can now use the dataset to train the model to make the desired predictions.

- The feature engineering mindset is very experimental. Data quality comes into play when we deal with feature feature engineering.

- Construct as many relevant features as possible from your data and follow it up with a feature selection process to drop out bad features.

Join us on Slack to ask questions and keep the discussion going!

Use the channel:

#build-a-ds-project

resources

https://online.hbs.edu/blog/post/data-wrangling

https://www.datacamp.com/cheat-sheet/pandas-cheat-sheet-data-wrangling-in-python

https://www.tableau.com/learn/articles/what-is-data-cleaning