MachineLearnAthon – Feature Engineering

Recorded by Lara Kuhlmann

Learning outcomes of today

After successfully completing this micro-lecture, you are able to….

- Enrich datasets with external information
- Extract informative features from datetime values
- Merge and aggregate multiple data sources
- Apply normalization and logarithmic transformations

Agenda for today

- Data Enrichment
- Datetime Feature Extraction
- Combining datasets – merge and aggregation
- Data normalization and Log-Transformation

Data Enrichment

- Refers to the process of enhancing, refining or organizing data to extract valuable insight from it. [1]
- The purpose of data enrichment is to discern relationships, clusters, semantic ontologies within a collection of data that unveil new insights to make informed decisions. [2]
- Benefits of data enrichment: [1]
  - Diagnosis and mitigation of potential risks
  - Prompt identification of new opportunities
- Example:
  - Weather or temperature data for energy or sales forecasting
  - Currency exchange rates or traffic intensity

[1] Azad, Salahuddin & Wasimi, Saleh & Ali, A B M Shawkat. (2018). Business Data Enrichment: Issues and Challenges.
[2] M. Toussant (2018, Mar. 22). The Importance of Human-Curated Data Enrichment of Big Data Analysis

Data Enrichment Example – Weather APIs

- External weather data can inject real-world signal into predictive models.
- Online sources:
  - Open-Meteo: Global weather model data, hourly or daily resolution, very simple API-Calls, no API key required.
  - OpenWeatherMap: Global coverage, historic data and forecast, granularity down to hourly, API key required

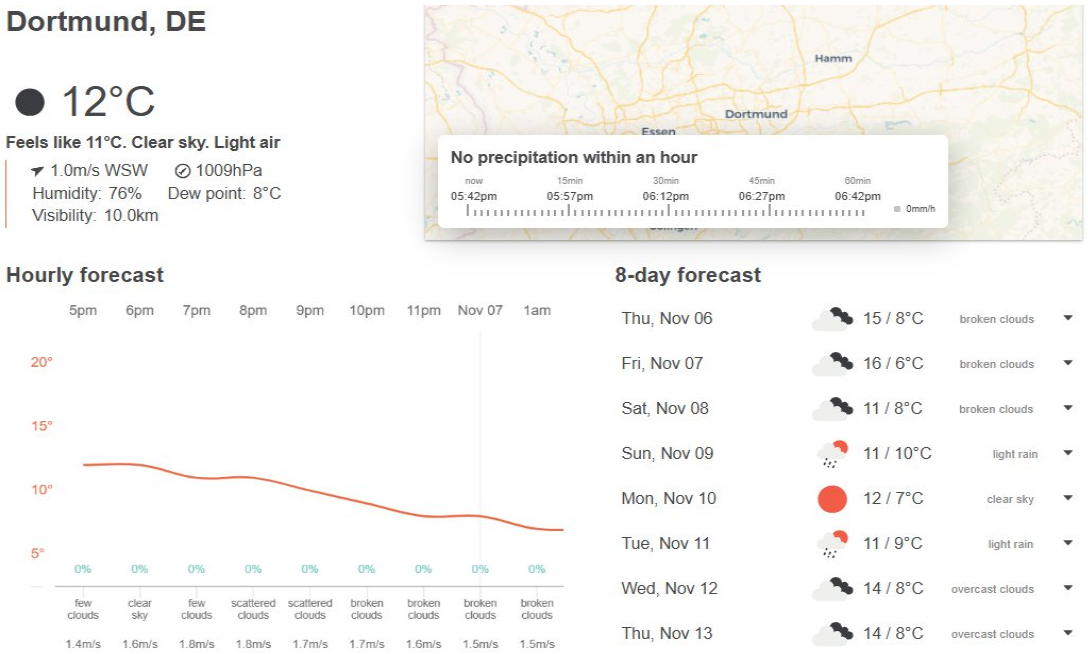


Bild: openweathermap.org/api

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

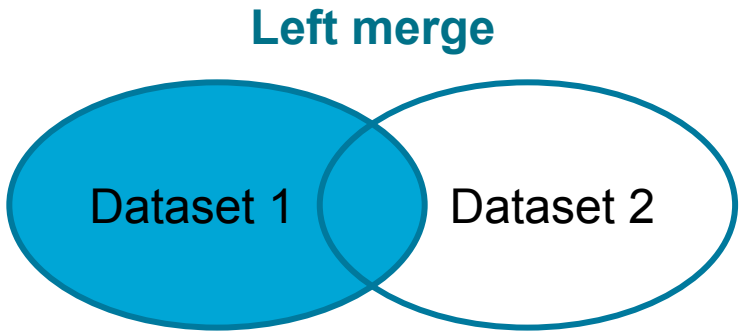
Datetime Feature Extraction

- Datetime features carry seasonal or temporal patterns useful for Machine Learning models.
- DatetimeFeatures() automatically extracts several date and time features from datetime variables.
- DatetimeFeatures() has:
  - Extract time features
  - Extract date and time features
  - Time series
- Derived features can include:
  - Month, quarter, year
  - Weekday/weekend indicator
  - Hour of the day

Python datetime documentation; Featuretools documentation.

Combining datasets – Merge (I)

- If the datasets have at least variable in column, they can be used as keys to join the datasets
- The merge can either be left, right, inner or outer
- In case of a left merge, only the keys from the left dataset are used
- This may result in missing values in the column from the right dataset



left

	key1	key2	A	B
0	K0	K0	A0	B0
1	K0	K1	A1	B1
2	K1	K0	A2	B2
3	K2	K1	A3	B3

right

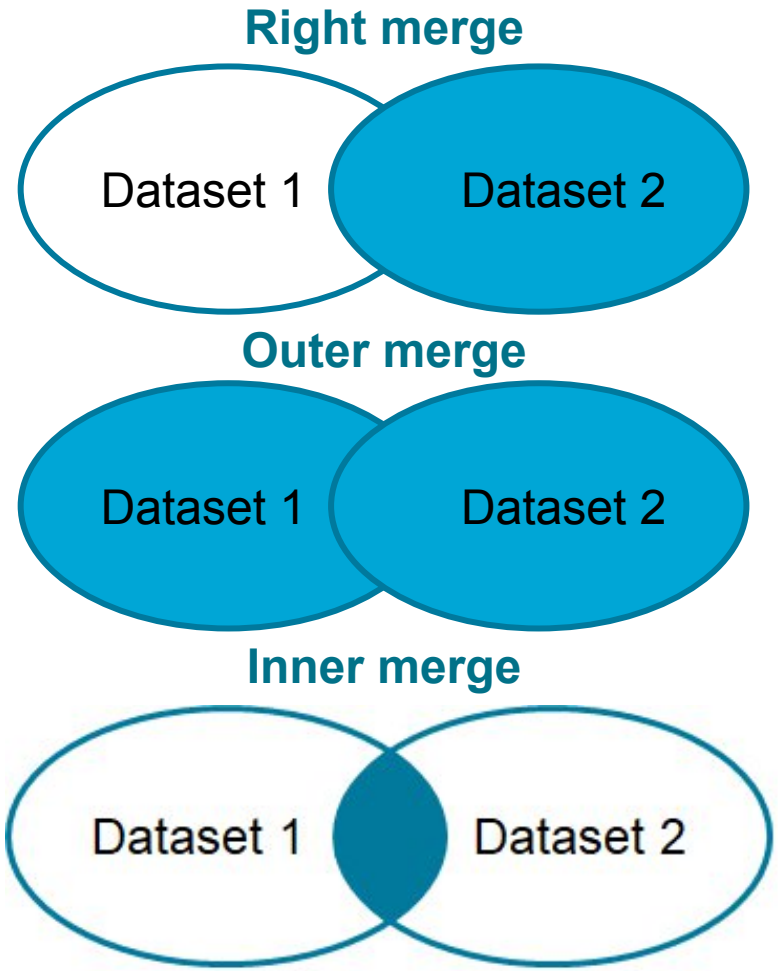
	key1	key2	C	D
0	K0	K0	C0	D0
1	K1	K0	C1	D1
2	K1	K0	C2	D2
3	K2	K0	C3	D3

Result

	key1	key2	A	B	C	D
0	K0	K0	A0	B0	C0	D0
1	K0	K1	A1	B1	NaN	NaN
2	K1	K0	A2	B2	C1	D1
3	K1	K0	A2	B2	C2	D2
4	K2	K1	A3	B3	NaN	NaN

Source: https://pandas.pydata.org/docs/user_guide/merging.html

Combining datasets – Merge (II)



Source: https://pandas.pydata.org/docs/user_guide/merging.html

left

	key1	key2	A	B
0	K0	K0	A0	B0
1	K0	K1	A1	B1
2	K1	K0	A2	B2
3	K2	K1	A3	B3

right

	key1	key2	C	D
0	K0	K0	C0	D0
1	K1	K0	C1	D1
2	K1	K0	C2	D2
3	K2	K0	C3	D3

Result

	key1	key2	A	B	C	D
0	K0	K0	A0	B0	C0	D0
1	K1	K0	A2	B2	C1	D1
2	K1	K0	A2	B2	C2	D2
3	K2	K0	NaN	NaN	C3	D3

left

	key1	key2	A	B
0	K0	K0	A0	B0
1	K0	K1	A1	B1
2	K1	K0	A2	B2
3	K2	K1	A3	B3

right

	key1	key2	C	D
0	K0	K0	C0	D0
1	K1	K0	C1	D1
2	K1	K0	C2	D2
3	K2	K0	C3	D3

Result

	key1	key2	A	B	C	D
0	K0	K0	A0	B0	C0	D0
1	K0	K1	A1	B1	NaN	NaN
2	K1	K0	A2	B2	C1	D1
3	K1	K0	A2	B2	C2	D2
4	K2	K0	NaN	NaN	C3	D3
5	K2	K1	A3	B3	NaN	NaN

left

	key1	key2	A	B
0	K0	K0	A0	B0
1	K0	K1	A1	B1
2	K1	K0	A2	B2
3	K2	K1	A3	B3

right

	key1	key2	C	D
0	K0	K0	C0	D0
1	K1	K0	C1	D1
2	K1	K0	C2	D2
3	K2	K0	C3	D3

Result

	key1	key2	A	B	C	D
0	K0	K0	A0	B0	C0	D0
1	K1	K0	A2	B2	C1	D1
2	K1	K0	A2	B2	C2	D2

Combining datasets – Aggregation (I)

- Aggregation is a GroupBy operation that reduces the dimension of the grouping object.
- The result of an aggregation is a scalar value for each column in a group.
- Example:

	Animal	Height	Weight
0	Cat	9.1	7.9
1	Dog	6.0	7.5
2	Cat	9.5	9.9
3	Dog	34.0	198.0

animals.groupby(“Animal”).sum() has output:

Animal	Height	Weight
Cat	18.6	17.8
Dog	40.0	205.5

Source: https://pandas.pydata.org/docs/user_guide/groupby.html

Combining datasets – Aggregation (II)

- Built-In Aggregation methods include:
  - Any() – compute whether any of the values in the groups are truthy
  - All() – Compute whether all of the values in the groups are truthy
  - First() – Compute the first occuring value in each group
  - Max() – Compute the maximum value in each group

Source: https://pandas.pydata.org/docs/user_guide/groupby.html#built-in-aggregation-methods

Transformation: Data normalization

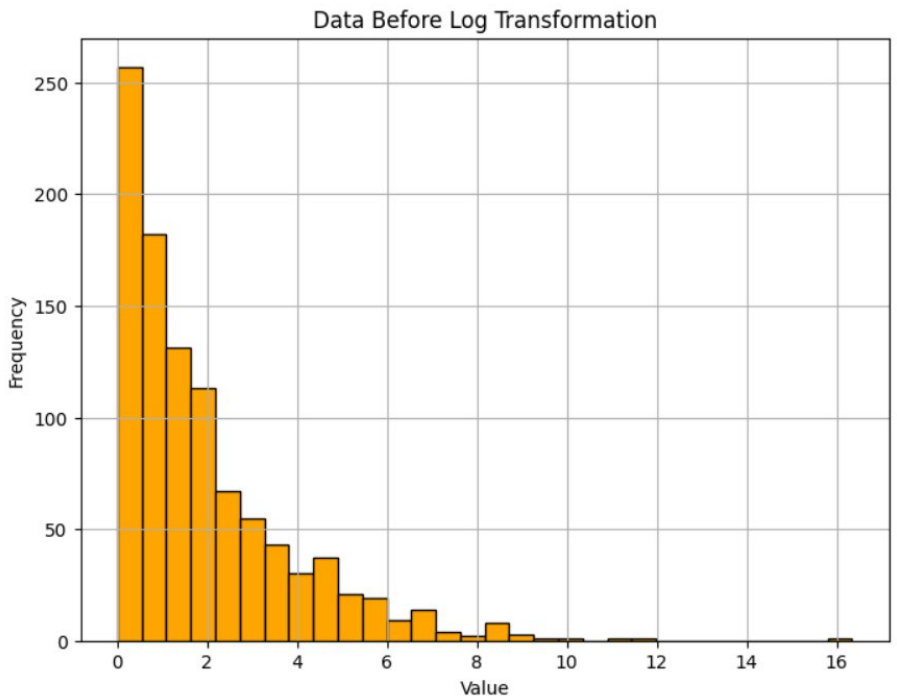
- Data normalization is a technique used to rescale numerical data into a standardized range, typically between 0 and 1 or -1 and 1
- This ensures that all features contribute equally to analysis or modeling, regardless of their original scale
- The most commonly used normalization technique is the Min-Max Scaling

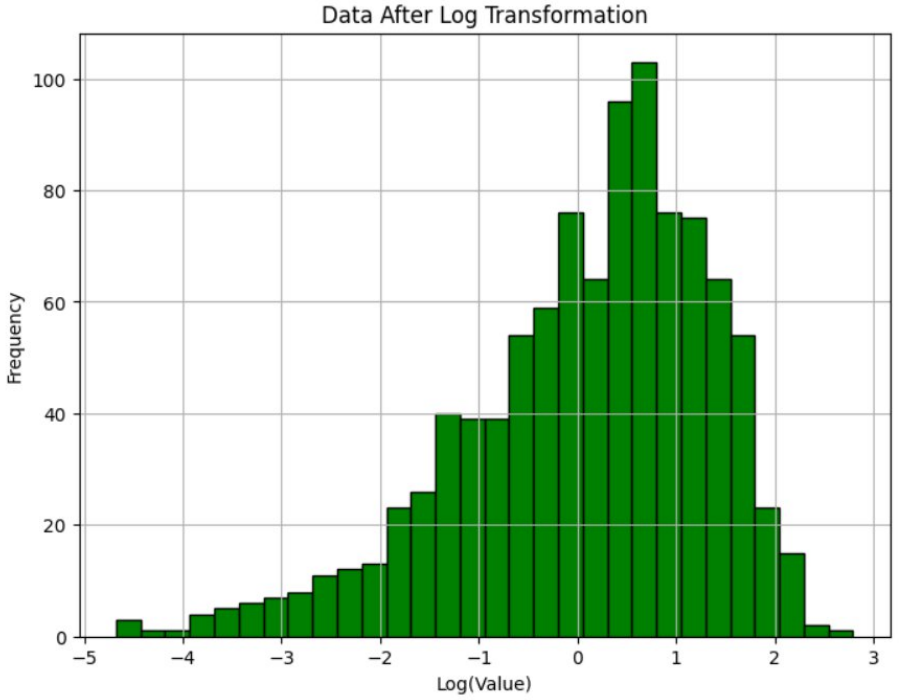
$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

García et al. (2015): Data Preprocessing in Data Mining, p.46 f.

Transformation: Data logarithmic transformation (I)

- The log-transformation is widely used to deal with skewed data.





- After log-transformation, the data is much more closer to a normal distribution.

Source: https://www.geeksforgeeks.org/data-science/log-transformation/

Recap this lecture

After successfully completing this lecture, you are able to….

- Enrich datasets with external information
- Extract informative features from datetime values
- Merge and aggregate multiple data sources
- Apply normalization and logarithmic transformations

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.