Data Leakage

Data Leakage

- When data is improperly and unrealistically used for training our models.
- Gives an unfair advantage to our model.
- Typical cases:
  - Data not available in production.
  - Target leakage.
  - Train-validation-test contamination.

Unavailable Features Leakage

- Using external data that won’t be available in production.
- Examples:
  - In stock prediction, using GDP that is calculated after the time of our logs.
  - The same if using future news releases.
  - In house price prediction, using economic indicators calculated post-hoc.

Target Leakage

- Deriving features from the target variable.
- Incorrectly helps the model to make better predictions.
- These features will not be available in production, as we won’t know the target.
- New feature example:
  - €/m^2 for house price prediction.
  - Patient medications that are only prescribed after the diagnosis.

Train-Validation-Test Contamination

- Including information from the validation/test sets in the training features.
- Train and validation/test datasets need to be properly isolated!
- Examples when using the whole dataset:
    - Feature scaling.
    - Feature selection.
    - New feature: Percentage of hotel cancellations per country.