Data leakage

A moment to reflect

- How bad it is?

- Whats the public opinion?

- To exploit or not to exploit

Contents

- Leakage types and examples

- Competition specific. Leaderboard probing

- Concrete walkthroughs

Leaks in time series

- Split should be done on time.

  - In real life we don't have information from future

  - In competitions first thing to look: train/public/private

split, is it on time?

- Even when split by time, features may contain information

  about future.

  - User history in CTR tasks

  - Weather

Unexpected information

- Meta data

- Information in IDs

- Row order