Question 1

Suppose that you have a credit scoring task, where you have to create a ML model that approximates expert evaluation of an individual's creditworthiness. Which of the following can potentially be a data leakage? Select all that apply.

Correct answers:

- An ID of a data point (row) in the train set correlates with target variable. Data was not shuffled, this information can not be used in real-world scenario.
- First half of the data points in the train set has a score of 0, while the second half has scores > 0. Same as above, data was not shuffled, this information can not be used in real-world scenario..

Incorrect answers:

- Among the features you have a company_id, an identifier of a company where this person works. It turns out that this feature is very important and adding it to the model significantly improves your score. This is a perfectly fine categorical feature, don't mix it up with and ID of a data point.

Question 2

What is the most foolproof way to set up a time series competition?

Correct answers:

- Split train, public and private parts of data by time. Remove all features except IDs (e.g. timestamp) from test set so that participants will generate all the features based on past and join them themselves. Correct! Only complete removal of all features from test set can guarantee that there is no data leakage.

Incorrect answers:

- Make a time based split for train/test and a random split for public/private. Vulnerable to leaderboard probing.
- Split train, public and private parts of data by time. Remove time variable from test set, keep the features. Participants can try to reverse engineer time order and exploit future peeking.

Question 3

Suppose that you have a binary classification task being evaluated by logloss metric. You know that there are 10000 rows in public chunk of test set and that constant 0.3 prediction gives the public score of 1.01. Mean of target variable in train is 0.44. What is the mean of target variable in public part of test data (up to 4 decimal places)?

Correct answers:

- ~0.771 Use logloss formula.

Question 4

Suppose that you are solving image classification task. What is the label of this picture?

Correct answer is 3. Check image name!

Mark as completed