

Data leakages

测验, 4 个问题

6/6 分 (100%)

**恭喜！您通过了！**[下一项](#)2 / 2
分数

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.



First half of the data points in the train set has a score of 0, while the second half has scores > 0 .

正确

Data was not shuffled, this information can not be used in real-world scenario



An ID of a data point (row) in the train set correlates with target variable.

正确

Data was not shuffled, this information can not be used in real-world scenario



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.

未选择的是正确的1 / 1
分数

2.

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



Make a time based split for train/test and a random split for public/private.





Data leakages

测验, 4 个问题

Split train, public and private parts of data by time. Remove all features except IDs (e.g. target comp) from test set so that participants will generate all the features based on past and join them themselves.

6/6 分 (100%)

正确

Correct!



Split train, public and private parts of data by time. Remove time variable from test set, keep the features.



2 / 2
分数

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)?

0.7712

正确回答



1 / 1
分数

4.

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

Data leakages

测验, 4 个问题

6/6 分 (100%)



3

正确答案



Data leakages

测验, 4 个问题

6/6 分 (100%)