Data leakages
TOTAL POINTS 6

1.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.

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

() 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.

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

2 points

2.Question 2
What is the most foolproof way to set up a time series competition?

(**X**) 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.

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

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

1 point

3.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)?

Enter answer here 0.7711

2 points

4.Question 4
Suppose that you are solving image classification task. What is the label of this picture?

Enter answer here 3

1 point

I, Chun-Min Jen, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account. Learn more about Coursera's Honor Code