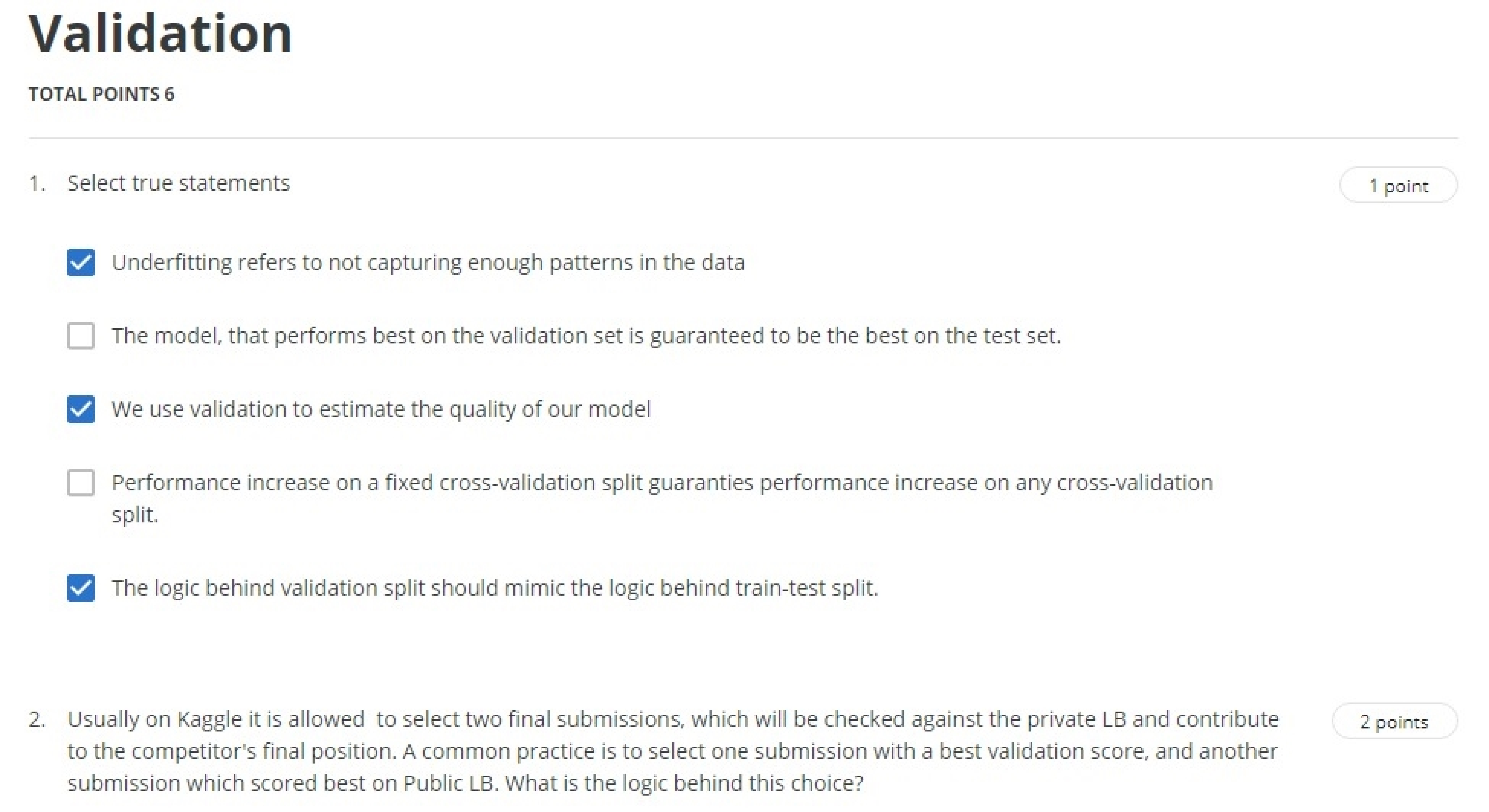


Validation

TOTAL POINTS 6

1. Select true statements

1 point

☑ Underfitting refers to not capturing enough patterns in the data

□ The model, that performs best on the validation set is guaranteed to be the best on the test set.

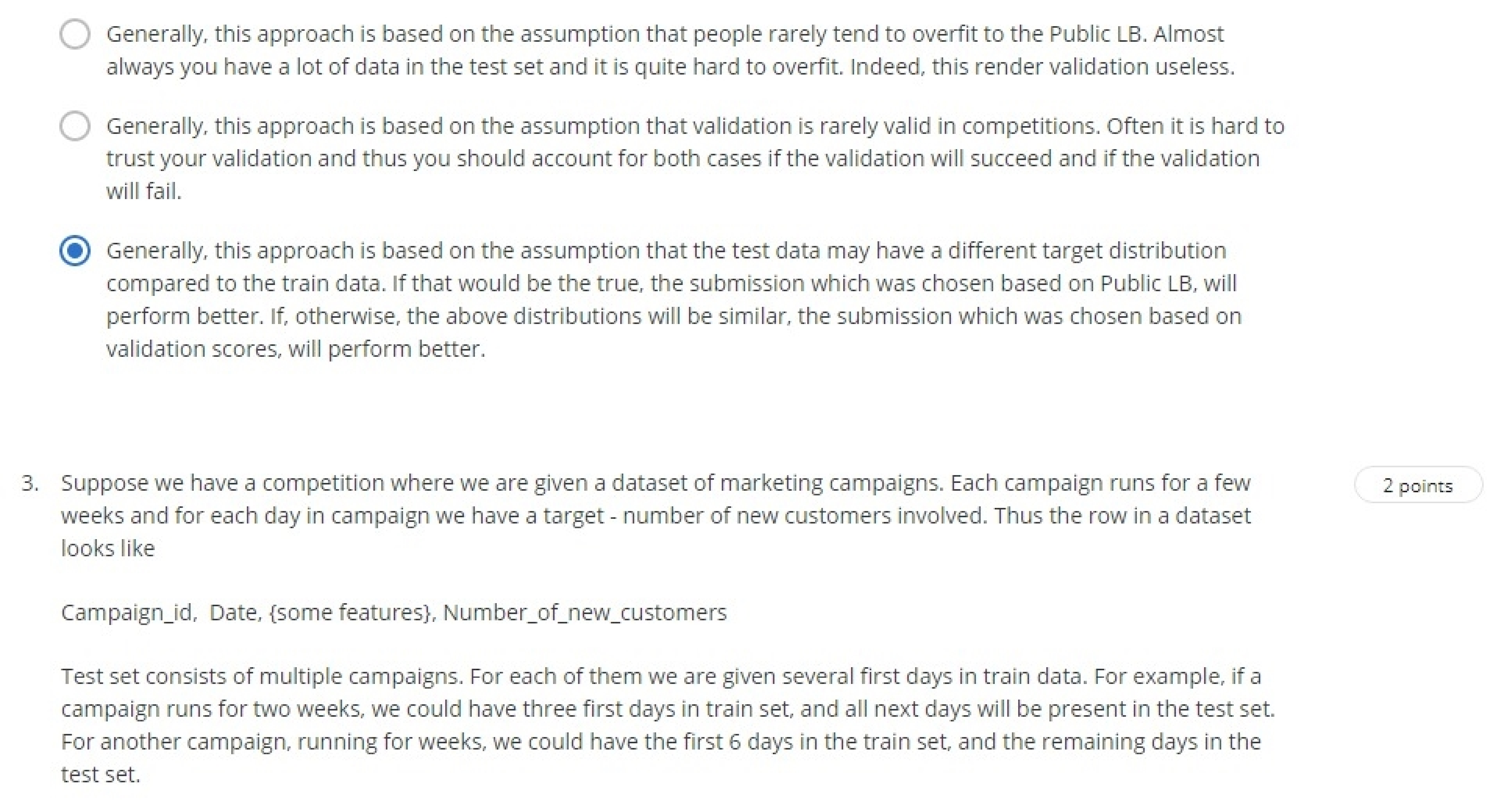
☑ We use validation to estimate the quality of our model

□ Performance increase on a fixed cross-validation split guaranties performance increase on any cross-validation split.

☑ The logic behind validation split should mimic the logic behind train-test split.

2. Usually on Kaggle it is allowed to select two final submissions, which will be checked against the private LB and contribute to the competitor's final position. A common practice is to select one submission with a best validation score, and another submission which scored best on Public LB. What is the logic behind this choice?

2 points



○ Generally, this approach is based on the assumption that people rarely tend to overfit to the Public LB. Almost always you have a lot of data in the test set and it is quite hard to overfit. Indeed, this render validation useless.

○ Generally, this approach is based on the assumption that validation is rarely valid in competitions. Often it is hard to trust your validation and thus you should account for both cases if the validation will succeed and if the validation will fail.

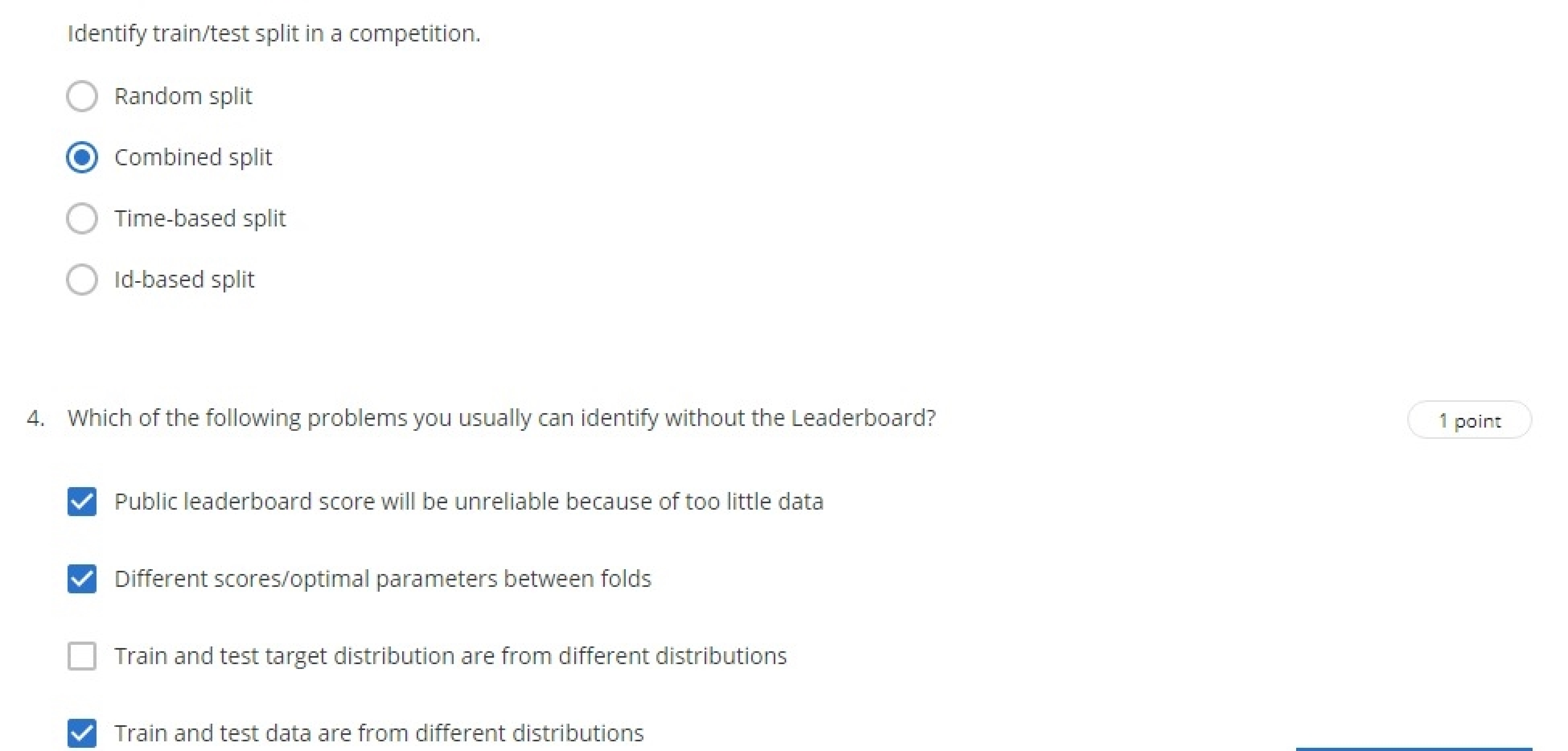
● Generally, this approach is based on the assumption that the test data may have a different target distribution compared to the train data. If that would be the true, the submission which was chosen based on Public LB, will perform better. If, otherwise, the above distributions will be similar, the submission which was chosen based on validation scores, will perform better.

3. Suppose we have a competition where we are given a dataset of marketing campaigns. Each campaign runs for a few weeks and for each day in campaign we have a target - number of new customers involved. Thus the row in a dataset looks like

2 points

Campaign_id, Date, {some features}, Number_of_new_customers

Test set consists of multiple campaigns. For each of them we are given several first days in train data. For example, if a campaign runs for two weeks, we could have three first days in train set, and all next days will be present in the test set. For another campaign, running for weeks, we could have the first 6 days in the train set, and the remaining days in the test set.



Identify train/test split in a competition.

○ Random split

● Combined split

○ Time-based split

○ Id-based split

4. Which of the following problems you usually can identify without the Leaderboard?

☑ Public leaderboard score will be unreliable because of too little data

☑ Different scores/optimal parameters between folds

□ Train and test target distribution are from different distributions

☑ Train and test data are from different distributions