Question 1

Select true statements

Correct answers:

- We use validation to estimate the quality of our model. Correct! This is the main purpose of validation.
- The logic behind validation split should mimic the logic behind train-test split. Correct! This is the main rule of making a reliable validation.
- Underfitting refers to not capturing enough patterns in the data. Correct! Because a model cant utilize all existing patterns, it has lower quality than it could have.

Incorrect answers:

- The model, that performs best on the validation set is guaranteed to be the best on the test set. Incorrect. Target in the test set can have different distribution and our score estimation can fail.
- Performance increase on a fixed cross-validation split guaranties performance increase on any cross-validation split. Incorrect. You can overfit to the specific CV-split. You should change your split from time to time to reduce the chance of overfitting.

Question 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?

Correct answers:

- 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.

Incorrect answers:

- 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. Almost always 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.

Question 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:

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.

Identify train/test split in a competition.

Correct answer:

- Combined split. For each campaign train and test are divided by a date, and this date can be different for different campaigns. Thus, split is made by id and by time.

Incorrect answers:

- Random split
- Time-based split
- Id-based split

Question 4

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

Correct answers:

- Different scores/optimal parameters between folds. Correct. This can be identified during validation.
- Public leaderboard score will be unreliable because of too little data. Correct. Usually you can estimate variance of Public LB score using validation. You need to train a model and see how its score varies on different folds with the same size as Public LB.
- Train and test data are from different distributions. Correct! Often enough we can find out this during EDA. To refresh your memory about this problem, review the last video in the Validation module.

Incorrect answers:

- Train and test target distribution are from different distributions. Incorrect! To do this, we would need to have test target values, which is not possible in a competition.

Mark as completed