Question 1

Which hyperparameters are first to tune in sklearn's RandomForest?

Correct answers:

- n_estimators, max_depth, min_samples_split. Yes! These parameters are important. The first one should just be sufficiently large, you do not actually need to tune it.

Incorrect answers:

- n_jobs, random_state, verbose. Some of these parameters can even change the result of the training but only because of randomness involved. They are for sure not the parameters to tune.
- bootstrap, oob_score, warm_start. These parameters are not what you want to tune in the model!

Question 2

Suppose you fit LightGBM to your train data and check performance on the validation set. The train set consists of 500 rows and 1000 different features and validation set consist of 50 objects. You run automatic hyperparameter optimization method overnight and in the morning you select the best parameters, produce results for the test set and submit to the leaderboard. We also know that test set comes from the same distribution as train and validation sets.

Correct answers:

- There is a high chance of overfitting to the validation set. That is, there is a high chance that score on the test set will be bad. This is because we've tried too much hyperparameters while the dataset is small and the number of features is large. *Correct!* This is because of multiple comparisons fallacy