ABSTRACT

According to embodiments of the disclosure, methods, apparatuses, devices and medium for model performance evaluation are provided. A method includes: applying, at a client node, a plurality of data samples to a prediction model respectively to obtain a plurality of predicted scores output by the prediction model, the plurality of predicted scores indicating respectively predicted probabilities that the plurality of data samples belong to a first category or a second category; determining values of a plurality of metric parameters related to a predetermined performance indicator of the prediction model based on the plurality of predicted scores and a plurality of ground-truth labels of the plurality of data samples; performing perturbation on the values of the plurality of metric parameters to obtain perturbed values of the plurality of metric parameters; and sending the perturbed values of the plurality of metric parameters to a server node.