

Table S1 The parameter descriptions of the training parameters of XGBoost, CatBoost, LightGBM, GBDT, and RF

Classifier	Training parameters	Parameter description
XGBoost	n_estimators	The maximum number of trees that can be built when solving ML problems
	learning_rate	Step size shrinkage used in update to prevent overfitting. After each boosting step, we can directly get the weights of new features, and learning_rate shrinks the feature weights to make the boosting process more conservative
	max_depth	Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit
	subsample	Subsample ratio of the training instances. Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees. This will prevent overfitting. Subsampling will occur once in every boosting iteration
	colsample_bytree	colsample_bytree is the subsample ratio of columns when constructing each tree. Subsampling occurs once for every tree constructed
	colsample_bylevel	colsample_bylevel is the subsample ratio of columns for each level. Subsampling occurs once for every new depth level reached in a tree. Columns are subsampled from the set of columns chosen for the current tree
CatBoost	n_estimators	Same as n_estimators in XGBoost
	learning_rate	Same as learning_rate in XGBoost
	max_depth	Same as max_depth in XGBoost
	reg_lambda	Coefficient at the L2 regularization term of the cost function. Increasing this value will make the model more conservative. Normalized to the number of training examples
LightGBM	n_estimators	Same as n_estimators in XGBoost
	learning_rate	Same as learning_rate in XGBoost
	max_depth	Same as max_depth in XGBoost
	num_leaves	Maximum tree leaves in the resulting tree
	lambda_l1	L1 regularization term on weights. Increasing this value will make the model more conservative. Normalized to the number of training examples
	lambda_l2	Same as reg_lambda in CatBoost
GBDT	n_estimators	Same as n_estimators in XGBoost
	learning_rate	Same as learning_rate in XGBoost
	max_depth	Same as max_depth in XGBoost
	max_leaf_nodes	Same as num_leaves in LightGBM
	subsample	Same as subsample in XGBoost
RF	n_estimators	Same as n_estimators in XGBoost
	max_depth	Same as max_depth in XGBoost
	min_samples_split	The minimum number of samples required to split an internal node
	min_samples_leaf	The minimum number of samples required to be at a leaf node

GBDT, gradient boosting decision trees; RF, random forest; ML, machine learning.