1. Для чего и в каких случаях полезны различные варианты усреднения для метрик качества классификации: micro, macro, weighted?

Эти варианты используются в многоклассовых задачах.

macro: арифметическое среднее значение оценки F1 для каждого класса.

micro: вычисляется путем подсчета общего количества истинных положительных, ложных отрицательных значений и ложных срабатываний.

weighted: взвешенное среднее значение частоты класса для оценки F1 для каждого класса.

micro - среднее используется для подсчета по всем классам

macro - среднее используется для подсчета по каждому классу независимо, но не берет в расчет дисбаланс классов.

Weighted – это macro усреднение, но уже с учетом дисбаланса классов.

2. В чём разница между моделями xgboost, lightgbm и catboost или какие их основные особенности?

Xgboost – разбиение дерева осуществляется по параметру похожести (similarity).

Прирост информации считается Gain = LEFTsimilarity + RIGHTsimilarity – ROOTsimilarity.

Так же xgboost стрижет дерево. Задается коэффициент Y ( минимальный прирост информации).

Gain – Y >= 0 – оставляем

Gain – Y < 0 - убираем

LightGBM – оптимизирует xgboost за счет 2ух подходов.

GOSS- gradien-based one-side sampling. Оптимизация по строкам за счет отбора наблюдений с большей ошибкой.

EFB – exclusive feature bundling. Оптимизация происходит по столбцам за счет объединения некоторых признаков.

Так же есть параметр class\_weight для несбалансированных классов.

Работает быстрее, чем xgboost

Catboost – обучает несколько моделей за одну итерацию. Высокий уровень случайности (перемешивание выборки, случайности в начале построения дерева).

Работает с категориальными признаками.