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

Разные усреднения метрик помогают, когда классов множество, и по-разному отображают дисбаланс классов.

Micro-усреднение представляет собой сумму верных срабатываний классификатора для всех классов, делённое на общее количество объектов. Micro-усреднение метрик precision, recall, f-score всегда равно и принимает значение accuracy. Micro-усреднение подходит, когда в данных баланс классов близок к идеальному. Например, оно может быть эффективно для классификации цифр на картинках.

Macro-усреднение представляет собой среднее арифметическое от суммы метрик для каждого из классов. Например, $\text{precision}_{macro} = \frac{\text{precision}_0 + \text{precision}_2}{2}$. Macro-усреднение следует применять, когда данные несбалансированы и нужно дать всем классам одинаковый вес, вне зависимости от количества объектов класса в датасете. Например, это усреднение можно применять при выявлении мошеннических транзакций.

Weighted -усреднение представляет собой взвешенное среднее от суммы метрик для каждого из классов. То есть для каждого класса его вес (количество объектов этого класса делённое на общее количество объектов) умножается на метрику для этого класса, а затем все эти значения суммируются. Weighted -усреднение больше берет во внимание метрику большего класса и меньше берет во внимание метрику меньшего класса. Weighted-усреднение следует применять, когда есть дисбаланс и нужно дать классам вес согласно количеству. Например, нужно хорошо распознавать часто встречающиеся объекты и не столь важно классифицировать редкие.

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

XGBoost - основана на фреймворке GBM, в основе данной модели лежит градиентный бустинг. XGBoost посчитывает похожесть между элементами и производит стрижку результирующих деревьев, основываясь на гиперпараметр гамма. Прирост информации в XGBoost подсчитывается, как разница между суммой веток и вершины. XGBoost может работать только с числовыми данными и хуже, чем две другие модели, работает на больших датасетах.

LightGBM - это облегченная версия градиентного бустинга, которая заостряет внимание на ошибках и не использует всю выборку. Также, эта модель группирует разреженные признаки, типичные для бинарного формата машинного обучения, а также признаки с близкими диапазонами значений. При построении модели LightGBM для выбора критерия разбиения используется GOSS - техника выделения объектов с наибольшим значением градиента.

Catboost - это быстрая модель градиентного бустинга, построенная на симметричных разветвлениях. Выдерживая симметрию, модель решает проблемы классификации быстрее аналогов. Catboost хорошо заточена на работу с категориальными признаками, эта модель сама переводит признаки в нужный формат без использования функции get_dummies и имеет встроенные алгоритмы, препятствующие переобучению.