Курс "Библиотеки Python для Data Science: продолжение"

Практическое задание к уроку 4

Тема "Оценка и интерпретация полученной модели"

1. Расскажите, как работает регуляризация в решающих деревьях, какие параметры мы штрафуем в данных алгоритмах?

Регуляризация - это способ борьбы с переобучением. Она добавляет модели штраф за слишком большую сложность модели.
Для деревьев решения основные способы борьбы с переобучением:
- накладывает явное ограничение глубины или минимального числа объектов в листе
- стрижка дерева.

В XGBoost регуляризация производится через подсчитывает сумму в зависимости

похожесть.
Если lambda больше, то
похожесть в рамках одной
вершины становится меньше.
Если lambda = 0, то похожесть
будет больше. В
дальнейшем, сравнивается
с затта (гипер
параметр CatBoost,
к-рый отвечает за
стрижку деревьев);
— если полученная инфо
больше
этой затта, то мы
оставляем
это дерево
Т.е. регуляризация в XG
Boost используется при
подсчете похожести

2. По какому принципу рассчитывается "важность признака (feature_importance)" в ансамблях деревьев?

Признаки ранжируются в

соответствии с объясненной дисперсией, которую каждый признак вносит в модель. Признаки отображаются в зависимости от их относительной важности, то есть процентной важности от наиболее важного признака.