Очистка данных

Андрей Мещеряков

Senior Data Scientist в EPAM Systems

Skillbox

Андрей Мещеряков

Senior Data Scientist в EPAM Systems

5+ лет

опыта в сфере Data Science и Machine Learning

Создавал

решения на основе машинного обучения для ритейла и маркетинга

Цель модуля

Изучить основные техники очистки данных для повышения качества предсказаний моделей машинного обучения.

Очистка данных

Замена пропусков

Skillbox образовательная платформа

Цель видео

Изучить основные стратегии заполнения пропусков в числовых и категориальных данных.

Источники пропусков

Missing Not at Random — пропущенное значение является корректным.

Пример: отсутствие отчества.

Источники пропусков

Missing at Random — пропуски случайны,
но прослеживается закономерность их появления.

Пример: ошибки измерения массы для слишком тяжелых автомобилей.

Источники пропусков

Missing Completely at Random — пропуски полностью случайны, закономерность их появления никак не прослеживается.

Пример: случайные сбои или шумы во время сбора данных.

Заполнение пропусков

Вывод

Вы изучили основные источники пропусков, способы их заполнения и ситуации, в которых лучше применять тот или иной способ.