Дедупликация (Deduplication)

> Синонимы: Дедубликация, Устранение дубликатов
> Разделы: Бизнес-задачи
> Решения: Loginom Data Quality

В информационных технологиях дедупликацией называют процесс исключения из наборов данных идентичных записей, называемых дубликатами.

Дедупликация является важной и неотъемлемой частью процесса предобработки и очистки данных, поэтому соответствующие инструменты входят в состав большинства аналитических платформ. Она производится на всех этапах работы с данными, начиная от ETL и заканчивая аналитическими модулями непосредственно перед анализом.

Необходимость многоступенчатой процедуры дедупликации обусловлена тем, что интеграция данных из различных источников, операции слияния, соединения и другие виды трансформации данных могут порождать новые дубликаты. Кроме этого, дубликаты могут появляться при неудачно выбранном алгоритме выборки (например, когда одно и то же наблюдение извлекается несколько раз при использовании выборки с возвратом).

Причинами, по которым дубликаты в большинстве случаев следует удалять, являются:

1. Большое количество дубликатов увеличивает объем хранимых данных и перегружает сеть при их передаче, особенно если объем данных значителен.
2. Дубликаты ухудшают репрезентативность данных.
3. Дубликаты могут искажать форму распределения данных и вызвать смещение статистических оценок.

Исключение дубликатов может производиться тремя основными способами:

1. **Удаление дублирующих записей** — следует использовать, если известно, что дублирование является следствием технического сбоя, человеческого фактора или некорректной интеграции данных (например, две или более записи в клиентской базе, содержащие идентичные персональные данные). Очевидно, что в этом случае дублирующие записи (кроме оригинальной) отражают реально не существующие объекты и события, что нарушает бизнес-логику. Поэтому такие дубликаты должны быть удалены, т.е. фактически произойдет их слияние в одну запись.
2. **Агрегирование дубликатов** — объединение дубликатов в одну запись, если дубликаты отражают два или более реальных, но идентичных события или объекта. Например, товар отгружался со склада по накладной, но из-за большого объема отгрузка была произведена за два рейса в одинаковых объемах в пределах одной даты. В результате были зафиксированы две отгрузки в одинаковом объеме по

одной накладной, что противоречит бизнес-логике. В этом случае нужно объединить две записи в одну, а количества и суммы сложить.

3. Обогащение данных — заключается в добавлении к набору данных дополнительных полей, которые позволяют сделать записи уникальными, даже если ранее они были дубликатами. Например, для анализа потребовалась выборка, содержащая фамилию, инициалы, год рождения и стаж работы клиента. Очевидно, что в такой выборке может оказаться значительное количество дубликатов. Однако если добавить в выборку поле с номером паспорта, то все записи станут уникальными.

Следует отметить, что существуют ситуации, когда аналитиком может быть принято решение не обрабатывать дубликаты. Это может быть сделано, если:

- дубликатов немного и они не нарушают бизнес-логику анализа;
- объем выборки после удаления дубликатов становится недостаточным для анализа.

Кроме этого, в некоторых случаях дубликаты специально вводятся в данные, например, для увеличения числа примеров обучающего множества, если их критически мало для построения модели, или для балансировки классов в задачах бинарной классификации.