Мусор на входе — мусор на выходе (Garbage in, garbage out)

> Синонимы: GIGO, Rubbish in — rubbish out, RIRO

В науке о данных и машинном обучении принцип, согласно которому качество результатов работы любой аналитической модели полностью зависит от качества данных, на которых она построена. И если оно низкое, то и результаты будет соответствующими даже в том случае, если сам алгоритм обучения модели полностью корректен.

Говоря простыми словами, как на основе плохого сырья или материалов невозможно произвести качественную продукцию, так и на основе некачественных данных нельзя получить корректные результаты их обработки.

Сам принцип появился в 1950-х годах, еще на заре компьютерной эры, когда вычислительные средства стали все более широко использоваться в бизнесе и управлении. Это был период перехода от ручной обработки данных к автоматизированным процессам. И принцип отражал основную идею о том, что качество выходных данных информационной системы напрямую зависело от качества входных. Термин приобретал популярность по мере того, как все больше организаций стали использовать обработку данных для поддержки принятия решений.

Со временем принцип GIGO вышел за рамки IT и начал использоваться в широком спектре дисциплин, включая бизнес, финансы, страхование и другие. Он служит общим принципом, согласно которому решения могут быть настолько хороши, насколько хороша информация, на основе которой они принимаются.

В эпоху больших данных и искусственного интеллекта концепция GIGO остается актуальной и подчеркивает важность правильной организации управления данными, необходимость в репрезентативных и непредвзятых обучающих выборках, а также обращает внимание на потенциальные последствия пренебрежения их качеством.

Принцип GIGO может проявляться в разных формах и отраслях. Например, в финансах, использование некорректных финансовых показателей из-за ошибок в отчетности или устаревших экономических данных, может привести к неверной оценке ситуации на рынке и плохим инвестиционным решениям, что приведет к потерям.

Виды входных мусорных данных, которые могут производить мусорные же результаты на выходе информационной системы или аналитической модели, весьма разнообразны, и отражают проблемы с их качеством, целостностью и релевантностью. Рассмотрим основные из них.

Неточные данные. Данные с ошибками, возникающими из-за ручного ввода, в ходе измерения при сборе или в процессе передачи информации. Это приводит к прямому искажению выходных значений, и система обрабатывает ложную информацию как истинную, что приводит к некорректным выводам и плохим решениям.

Неполные данные. Наборы данных, в которых отсутствует часть наблюдений или признаков. Они могут возникать из-за системных ошибок, плохих методов сбора или сбоев при передаче информации. Модели могут неправильно интерпретировать такие данные, что приводит к предвзятым или искаженным результатом анализа. Неполные данные также могут привести к переобучению или недообучению моделей, что влияет на их производительность в ходе практического применения.

Устаревшие данные. Информация, которая не является актуальной и не отражает текущего состояния бизнес-процессов компании, ситуацию на рынке и т.д. Эта проблема особенно характерна для бизнес-сред с высокой динамикой. Решения, основанные на устаревших данных, могут быть неактуальными или неподходящими для текущих условий, что приводит к неэффективным или контрпродуктивным результатам.

Предвзятые данные. Данные, которые систематически благоприятствуют получению определенных результатов, выгодных определенным лицам или группам.

Нерелевантные данные. Данные, которые не относятся к решаемой задаче и не способствуют достижению целей анализа.

Вводящие в заблуждение данные. Данные, которые, будучи точными и актуальными в ограниченном контексте, приводят к неверным предположениям или выводам из-за их неверного представления или интерпретации.

Дубликаты. Повторяющиеся записи в наборе данных, которые могут возникать на этапах сбора или агрегации. Они приводят к получению смещенных результатов анализа и избыточным вычислениям.

Плохо структурированные данные. Данные, которые плохо организованы или отформатированы, распределены по нескольким источникам и не интегрированы должным образом. Это усложняет обработку и анализ данных, увеличивая риск появления ошибок при обработке и интерпретации результатов.

Чтобы избежать замусоривания данных перед аналитической обработкой необходимо производить их профайлинг и очистку.