

РОЛЬ КАЧЕСТВА ДАННЫХ В ПОВЫШЕНИИ ЭФФЕКТИВНОСТИ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Багауов К.И.

Уфимский университет науки и технологий, Уфа, Россия

Ключевые слова: машинное обучение, качество данных, эффективность моделей, разметка данных, оценка алгоритмов, системы машинного обучения.

Аннотация. В статье рассматривается роль качества данных в повышении эффективности моделей машинного обучения. Показано, что точность, полнота, согласованность, актуальность, репрезентативность и корректность разметки влияют на качество обучения, устойчивость алгоритмов и надежность их практического применения. Проанализированы основные характеристики данных, методы их обеспечения на разных этапах цикла обучения, а также результаты исследований, посвященных обработке пропусков и выявлению ошибочной разметки. На примере практики компаний X5 Digital, Яндекс, Walmart и Apple раскрывается значение качества данных для масштабируемости и результативности решений. Сделан вывод о том, что системная работа с данными является важным условием повышения эффективности моделей.

THE ROLE OF DATA QUALITY IN IMPROVING THE EFFECTIVENESS OF MACHINE LEARNING MODELS

Bagauov K.I.

Ufa University of Science and Technology, Ufa, Russia

Keywords: machine learning, data quality, model effectiveness, data labeling, algorithm evaluation, machine learning systems.

Abstract. This article examines the role of data quality in improving the effectiveness of machine learning models. It is shown that accuracy, completeness, consistency, timeliness, representativeness, and labeling correctness influence training quality, algorithm robustness, and the reliability of practical application. The study analyzes the main characteristics of data quality, methods for ensuring it at different stages of the training cycle, as well as research findings on missing data processing and the detection of mislabeled instances. Drawing on the practices of X5 Digital, Yandex, Walmart, and Apple, the article demonstrates the importance of data quality for the scalability and effectiveness of applied solutions. It is concluded that systematic work with data is an important condition for improving model performance.

Введение

В современных исследованиях в области машинного обучения (МО) все более отчетливо проявляется закономерность, согласно которой предельная эффективность модели определяется не только архитектурой алгоритма и вычислительными ресурсами, но и характеристиками исходных данных, на которых осуществляется обучение, настройка и последующая валидация. Так, в работе Г.В. Соловьева и М.С. Соколова дефицит размеченных данных и сложность аннотирования рассматриваются как критические барьеры для разработки автоматизированных систем контроля, а сама постановка задачи строится вокруг необходимости компенсировать ограничения качества и

доступности данных в промышленной среде [1]. В другом исследовании, посвященном использованию больших языковых моделей с дополненным поиском для повышения интерпретируемости МО, Р.Д. Иванова, А.Б. Менисова, О.А. Мишукова и Т.Р. Сабирова переносят акцент на объяснимость результатов, однако и в этом случае устойчивость выводов и верификация интерпретируемых признаков напрямую связаны с характеристиками информационной базы, на которой строится объяснение модели [2].

Наконец, в исследовании Н.А. Жуковой и В.Э. Ковалевского о метаалгоритме управления процессами синтеза моделей МО показано, что даже автоматизированный выбор и настройка алгоритмов должны опираться на метасвойства данных, поскольку именно они определяют пространство поиска, вычислительную сложность и итоговые показатели качества модели [3]. Тем самым современные исследования в разных направлениях – от промышленного ИИ до интерпретируемого и автоматизированного МО – сходятся в признании данных не вспомогательным, а системообразующим фактором результативности моделей.

Целью настоящей статьи является анализ роли качества данных в повышении эффективности моделей МО с выявлением ключевых характеристик данных, влияющих на качество обучения, устойчивость алгоритмов и надежность их практического применения. Научная новизна работы состоит в обосновании роли качества данных как интегрального фактора, определяющего эффективность моделей МО на разных этапах их разработки, оценки и применения. Гипотеза исследования заключается в том, что повышение эффективности моделей МО достигается прежде всего за счет целенаправленного управления качеством данных, тогда как усложнение алгоритмической архитектуры без устранения дефектов данных обеспечивает ограниченный и нестабильный эффект.

Методы

Методологическую основу исследования составили анализ, сопоставление и обобщение современных научных публикаций, посвященных качеству данных и его влиянию на точность, устойчивость, интерпретируемость и воспроизводимость результатов МО. В работе использован системный подход, позволивший рассмотреть характеристики данных как важный фактор, влияющий на сбор, подготовку, разметку, отбор, валидацию и последующий мониторинг информации при эксплуатации интеллектуальных систем. Для выявления связи между параметрами данных и результативностью МО применялся логико-структурный анализ, а для сопоставления различных дефектов данных и их последствий – сравнительно-аналитический метод. Теоретическую базу исследования составили актуальные работы, посвященные ограниченности обучающих выборок, интерпретируемости алгоритмов и автоматизации их построения, что позволило показать качество данных как одно из ключевых условий эффективности современных интеллектуальных систем.

Качество данных как основа построения эффективных моделей МО

Как было сказано ранее, в условиях быстрого расширения практик ИИ значение качества данных приобретает не вспомогательный, а фундаментальный характер. Именно данные определяют границы обучаемости модели, ее способность к обобщению и устойчивость результатов в реальной среде применения. При этом в рамках ИИ машинное обучение выступает как одно из ключевых технологических направлений, в наибольшей степени зависящих от характеристик обучающих данных.

Актуальность проблематики работы с данными усиливается тем, что МО из экспериментального инструмента все более превращается в инфраструктурный элемент цифровой экономики. Согласно глобальному отчету The Business Research Company, размер рынка МО достиг \$93,73 млрд в 2025 году (рис. 1).

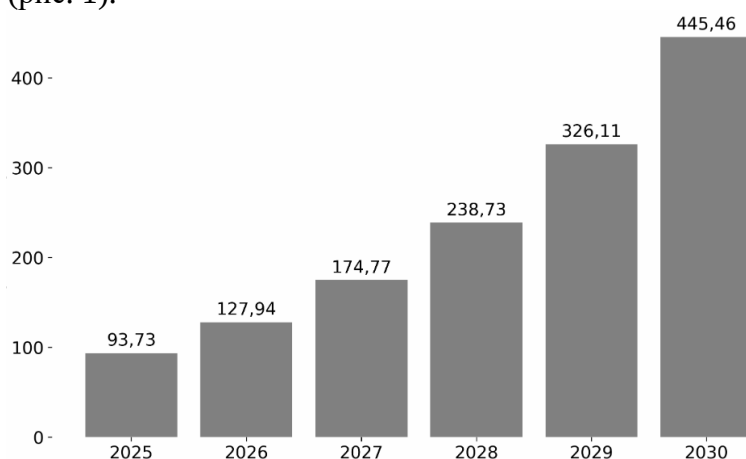


Рис. 1. Размер рынка МО во всем мире, согласно исследованию The Business Research Company, млрд. долларов [4]

При такой динамике даже незначительные дефекты в данных перестают быть частной технической проблемой и начинают непосредственно влиять на экономическую результативность, надежность цифровых сервисов и воспроизводимость аналитических решений.

Теоретически качество данных в МО должно рассматриваться как совокупность характеристик, обеспечивающих пригодность выборки для обучения, валидации и последующей эксплуатации модели (табл. 1).

Значимость перечисленных характеристик подтверждается на практике. В исследовании Sedir Mohammed и соавт. (2025) была проанализирована связь шести параметров качества данных с результативностью нескольких популярных алгоритмов МО в задачах классификации, регрессии и кластеризации [5]. Авторы отдельно рассмотрели случаи, когда дефекты присутствовали в обучающей, тестовой выборке или сразу в обеих. Результаты показали, что состояние данных на каждом этапе напрямую влияет на итоговые показатели работы алгоритмов.

Табл. 1. Основные характеристики качества данных в системах МО

Характеристика	Краткое содержание
Точность	Степень соответствия данных реальным значениям и объектам наблюдения.
Полнота	Наличие достаточного объема данных и отсутствие критически значимых пропусков.
Согласованность	Логическая и структурная непротиворечивость данных в пределах набора.
Актуальность	Соответствие данных текущему состоянию объекта, процесса или среды.
Репрезентативность	Способность выборки адекватно отражать свойства генеральной совокупности.
Отсутствие систематических искажений	Минимизация смещений, способных приводить к устойчиво ошибочным выводам модели.

Таким образом, качество данных следует рассматривать как базовое условие построения эффективных систем МО. Именно характеристики данных определяют не только возможность корректного обучения алгоритма, но и устойчивость его работы, достоверность получаемых результатов и применимость модели в реальных условиях.

Методы обеспечения и повышения качества данных на разных этапах цикла МО

Обеспечение качества данных в системах МО представляет собой не разовую процедуру предобработки, а непрерывный контур управленческих и технологических действий, встроенный во все этапы жизненного цикла модели [6]. На практике эффективность систем МО определяется не только характеристиками исходного массива наблюдений, но и способностью команды поддерживать согласованность, полноту, достоверность и актуальность данных при их сборе, разметке, подготовке, обучении модели и последующей эксплуатации. В связи с этим методы повышения качества данных целесообразно рассматривать как совокупность взаимосвязанных процедур, направленных на предупреждение дефектов, своевременное выявление искажений и поддержание пригодности данных для устойчивого обучения и надежного применения моделей в реальных условиях (табл. 2).

Представленные в таблице 2 методы образуют взаимосвязанную систему процедур, обеспечивающих надежность обучения и последующего применения алгоритмов. Их практическая значимость подтверждается результатами исследований, посвященных обработке неполных данных. Так, в статье Miao и соавторов (2023) на основе экспериментального сравнения различных методов заполнения пропусков показано, что выбор способа обработки неполных данных существенно влияет на обучение и итоговую результативность алгоритмов [7]. Авторы подчеркивают, что универсального

метода, одинаково эффективного для всех типов данных и задач, не существует, поэтому предобработка должна рассматриваться как значимое условие надежной работы алгоритмов. Это позволяет сделать вывод о том, что качество данных определяется не только исходным состоянием выборки, но и корректностью применяемых процедур подготовки.

Табл. 2. Основные методы обеспечения качества данных на этапах жизненного цикла модели МО

Этап цикла обучения	Метод обеспечения качества данных	Практический результат
Сбор данных	Проверка обязательных полей, типов данных и допустимых диапазонов значений.	Исключение грубых ошибок до попадания данных в обучающую выборку.
Подготовка данных	Удаление дублей, исправление несоответствий и обработка пропусков.	Снижение искажений в структуре данных и сохранение информативности выборки.
Разметка данных	Выборочный аудит меток и повторная проверка части объектов.	Снижение шума в целевой переменной и повышение точности обучения.
Формирование выборки	Балансировка классов.	Улучшение распознавания редких классов и повышение устойчивости алгоритма.
Обучение и тестирование	Исключение утечки данных при разделении выборок.	Получение более реалистичной оценки качества алгоритма.
Эксплуатация модели	Контроль изменений в данных после внедрения модели.	Своевременное выявление деградации системы и необходимость обновления обучения.

Если обработка пропусков показывает значимость предобработки данных в целом, то исследования ошибочной разметки позволяют точнее оценить роль контроля качества на этапе разметки. Научный интерес представляет статья Nazaretyan и соавторов (2025) [8]. Авторы рассмотрели 34 подхода и подробно протестировали 20 из них на 10 наборах данных при уровне шумовой разметки от 5 до 50%. Было показано, что даже наиболее эффективные методы при шуме около 20-30% позволяют обнаружить примерно 80% ошибочных меток, однако точность их работы остается умеренной – около 0,58-0,65. Это означает, что контроль качества разметки на ранних этапах является более надежной стратегией, чем последующее исправление уже искаженной выборки.

Таким образом, обеспечение качества данных на разных этапах цикла МО представляет собой не набор отдельных технических действий, а целостную систему процедур, от которой зависит надежность обучения,

корректность оценки и устойчивость алгоритмов в практической эксплуатации [9]. Рассмотренные исследования показывают, что своевременная проверка данных, корректная предобработка и контроль разметки позволяют снизить влияние дефектов выборки и повысить итоговую результативность системы.

Корпоративные подходы к управлению качеством данных в проектах МО

Практика крупных компаний показывает, что работа с данными в системах МО становится не только техническим, но и организационным фактором эффективности. На уровне прикладных решений это проявляется в стандартизации исходной информации, усилении контроля разметки, разработке специальных процедур валидации и постоянной проверке качества результатов уже после внедрения алгоритма.

В российской практике показательным примером является **проект X5 Digital** по автоматическому сопоставлению товаров (2025), целью которого была обработка до 20 млн. товарных позиций за трое суток. При этом система должна была не просто находить совпадения, но и делать это достаточно точно. Для повышения качества данных были унифицированы единицы измерения, приведены к единому виду текстовые описания, а также учтены сокращения и различия в написании. Согласно корпоративному техническому материалу на Habr, в результате реализации проекта уровень распознавания оказался высоким: модель CatBoost показала ROC-AUC 97%, а BERT – 98%. Кроме того, предварительная проверка данных позволила сократить число обращений к модели примерно на 50%, а производительность системы составила 13 запросов в секунду на одном вычислительном узле с последующим увеличением до целевого уровня 77 запросов в секунду. Этот пример показывает, что тщательная подготовка данных позволяет одновременно повысить и точность работы алгоритма, и эффективность его практического использования.

Среди компаний, непосредственно связанных с разработкой и внедрением решений в области МО, показателен опыт **Яндекса**. В материалах компании отмечается, что при оценке генеративных моделей недостаточно использовать только открытые стандартные бенчмарки, поскольку их результаты могут искажаться из-за утечки данных и ориентации на слишком упрощенные форматы ответов. Для повышения надежности проверки компания дополняет стандартные бенчмарки экспертной оценкой, что позволяет точнее судить о прикладном значении модели и ее работе в реальных сценариях. Такой подход показывает, что в практике МО точность данных – это не только чистота обучающей выборки, но и качество самих оценочных контуров, от которых зависит корректность выводов о реальной эффективности модели.

Среди зарубежных компаний интересен пример **Walmart**, где работа с данными встроена в процессы управления товарным каталогом. Компания

отмечает, что ранние версии моделей МО допускали ошибки при анализе данных, поэтому автоматическая загрузка результатов стала возможной только после достижения необходимого уровня надежности. В материалах Walmart указано, что модели с точностью выше 90-95% могут использоваться автоматически для работы с каталогами, тогда как в остальных случаях применяется отдельная модель проверки с участием человека. Кроме того, Walmart сообщает, что платформа МО Element используется примерно в 20 сервисах компании [10]. Это свидетельствует о значении качественно подготовленных и проверенных данных как основы для масштабируемого применения моделей МО в крупных корпоративных системах.

Среди зарубежных компаний, развивающих собственные технологии МО, рассмотрим пример **Apple**. В исследовании, посвященном автоматической расшифровке речи в Apple Podcasts (2024), компания показала, что традиционная метрика WER не всегда точно отражает результат с точки зрения пользователя [11]. Поэтому была предложена дополнительная метрика HEWER, которая лучше учитывает смысловые и читаемые ошибки. При анализе 800 фрагментов из 61 эпизода подкастов среднее значение WER составило 9,2%, тогда как HEWER – 1,4%. Это показывает, что при оценке систем МО важно учитывать не только качество самих данных, но и корректность выбранных критериев оценки. Даже достаточно точный алгоритм может давать впечатление более слабого или, напротив, более успешного результата, если для его анализа используются недостаточно точные метрики.

Таким образом, корпоративная практика показывает, что качество данных в проектах МО имеет многослойный характер. В одних случаях решающее значение показывает стандартизация и предварительная очистка исходных массивов, в других – контроль качества разметки, настройка проверочных моделей или совершенствование самих метрик оценки. Российские и зарубежные кейсы демонстрируют, что инвестиции в работу с данными позволяют повышать точность моделей, снижать нагрузку на ручную проверку, улучшать масштабируемость решений и делать результаты МО более надежными в реальной эксплуатации.

Заключение

Проведенный анализ показал, что качество данных является одним из ключевых факторов эффективности моделей МО. Рассмотренные материалы подтверждают, что точность, полнота, согласованность, актуальность, репрезентативность и корректность разметки напрямую влияют на качество обучения, устойчивость работы алгоритмов и надежность их применения в практических задачах.

Гипотеза исследования подтвердилась: повышение эффективности моделей МО во многом достигается за счет целенаправленного управления качеством данных. Рассмотренные исследования и корпоративные примеры показали, что проверка данных, корректная предобработка, контроль

разметки и выбор адекватных критериев оценки позволяют повысить точность моделей, снизить влияние дефектов выборки и сделать применение алгоритмов более надежным в реальных условиях.

Список литературы

1. Соловьев Г.В., Соколов М.С. Применение гибридных методов искусственного интеллекта для практических производственных задач в условиях труднодоступности обучающих данных // Научно-технический вестник информационных технологий, механики и оптики. – 2026. – Т. 26, №2. – С. 420-427. – doi.org/10.17586/2226-1494-2026-26-2-420-427.
2. Иванов Р.Д., Менисов А.Б., Мишуков О.А., Сабиров Т.Р. Подход к применению больших языковых моделей с дополненным поиском для повышения интерпретируемости моделей машинного обучения // Научно-технический вестник информационных технологий, механики и оптики. – 2026. – Т. 26, №1. – С. 69-76. – doi.org/10.17586/2226-1494-2026-26-1-69-76.
3. Жукова Н.А., Ковалевский В.Э. Метаалгоритм управления процессами синтеза моделей машинного обучения // Информационно-управляющие системы. – 2025. – №6(139). – С. 28-41. – doi.org/10.31799/1684-8853-2025-6-28-41.
4. Machine Learning Market Report 2026 / The Business Research Company [Электронный ресурс]. – Режим доступа: <https://www.thebusinessresearchcompany.com/report/machine-learning-global-market-report>.
5. Mohammed S., Budach L., Feuerpfeil M., Ihde N., Nathansen A., Noack N. The effects of data quality on machine learning performance on tabular data // Information Systems. 2025, vol. 132, p. 102549.
6. Нурматова Е.В. Методы преобразования данных для аналитики // Известия Тульского государственного университета. Технические науки. – 2024. – №3. – С. 8-13.
7. Miao X., Wu Y., Chen L., Gao Y., Yin J. An Experimental Survey of Missing Data Imputation Algorithms // IEEE Transactions on Knowledge and Data Engineering. 2023, vol. 35, no. 7, pp. 6630-6650. doi.org/10.1109/TKDE.2022.3186498.
8. Nazaretyan L., Leser U., Kircher M. Benchmarking machine learning methods for the identification of mislabeled data // Artificial Intelligence Review. 2025, vol. 58, no. 10, p. 310.
9. Яковлев Н.Б. Модели управления качеством на основе больших данных и машинного обучения // Известия Тульского государственного университета. Технические науки. – 2025. – №8. – С. 74-80.
10. Walmart's Element: A machine learning platform like no other / Walmart [Электронный ресурс]. – Режим доступа: https://tech.walmart.com/content/walmart-global-tech/en_us/blog/post/walmarts-element-a-machine-learning-platform-like-no-other.html.
11. Humanizing Word Error Rate for ASR Transcript Readability and Accessibility / Apple [Электронный ресурс]. – Режим доступа: <https://machinelearning.apple.com/research/humanizing-wer>.

References

1. Solovev G.V., Sokolov M.S. Application of hybrid artificial intelligence methods to practical industrial tasks under conditions of scarce training data // Scientific and

- Technical Journal of Information Technologies, Mechanics and Optics. 2026, vol. 26, no. 2, pp. 420-427. doi.org/10.17586/2226-1494-2026-26-2-420-427.
2. Ivanov R.D., Menisov A.B., Mishukov O.A., Sabirov T.R. An approach to using large language models with augmented search to improve the interpretability of machine learning models // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2026, vol. 26, no. 1, pp. 69-76. doi.org/10.17586/2226-1494-2026-26-1-69-76.
 3. Zhukova N.A., Kovalevsky V.E. Meta-algorithm for the process control of complex machine learning model synthesis // Information and Control Systems. 2025, no. 6, pp. 28-41. doi.org/10.31799/1684-8853-2025-6-28-41.
 4. Machine Learning Market Report 2026 / The Business Research Company [Electronic resource]. – Access mode: <https://www.thebusinessresearchcompany.com/report/machine-learning-global-market-report>.
 5. Mohammed S., Budach L., Feuerpfeil M., Ihde N., Nathansen A., Noack N. The effects of data quality on machine learning performance on tabular data // Information Systems. 2025, vol. 132, p. 102549.
 6. Nurmatova E.V. Data transformation methods for analytics // News of the Tula State University. Technical Sciences. 2024, no. 3, pp. 8-13.
 7. Miao X., Wu Y., Chen L., Gao Y., Yin J. An Experimental Survey of Missing Data Imputation Algorithms // IEEE Transactions on Knowledge and Data Engineering. 2023, vol. 35, no. 7, pp. 6630-6650. doi.org/10.1109/TKDE.2022.3186498.
 8. Nazaretyan L., Leser U., Kircher M. Benchmarking machine learning methods for the identification of mislabeled data // Artificial Intelligence Review. 2025, vol. 58, no. 10, p. 310.
 9. Yakovlev N.B. Quality management models based on big data and machine learning // News of the Tula State University. Technical Sciences. 2025, no. 8, pp. 74-80.
 10. Walmart's Element: A machine learning platform like no other / Walmart [Electronic resource]. – Access mode: https://tech.walmart.com/content/walmart-global-tech/en_us/blog/post/walmarts-element-a-machine-learning-platform-like-no-other.html.
 11. Humanizing Word Error Rate for ASR Transcript Readability and Accessibility / Apple [Electronic resource]. – Access mode: <https://machinelearning.apple.com/research/humanizing-wer>.

Багауов Карим Ильдарович – студент	Bagauov Karim Ildarovich – student
bagauovkarim@ro.ru	

Received 28.04.2026