

УДК 004.89

ОЦЕНКА ВЛИЯНИЯ КАЧЕСТВА ДАННЫХ НА ЭФФЕКТИВНОСТЬ МОДЕЛЕЙ ОБНАРУЖЕНИЯ АНОМАЛИЙ

Маргоретов В.И., студент гр. М3205, II курс
Университет ИТМО, г. Санкт-Петербург

Модели обнаружения аномалий являются ключевыми инструментами в различных областях, от финансов до кибербезопасности, где они помогают идентифицировать нестандартные паттерны, которые могут указывать на потенциальные проблемы или возможности. Однако, эффективность этих моделей сильно зависит от качества данных, на которых они обучены и тестируются [1-4].

Значение качества данных. Качество данных определяет надежность выводов, которые могут быть сделаны с помощью моделей машинного обучения. Для моделей обнаружения аномалий, это особенно важно, поскольку низкое качество данных может привести к высокому количеству ложных срабатываний или упущению реальных аномалий. Основные атрибуты качественных данных включают точность, полноту, своевременность, консистентность и достоверность.

Влияние качества данных на производительность моделей. Влияние качества данных на производительность моделей обнаружения аномалий многоаспектно [5-7]:

- Точность и полнота данных. Неполные или неточные данные могут ввести модель в заблуждение, что приведет к неверной классификации обычных паттернов как аномальных.
- Консистентность данных. Отсутствие стандартизации и различия в форматах данных могут затруднить определение истинных аномалий.
- Своевременность данных. Застарелые данные могут не отражать текущее состояние системы, что снижает эффективность модели в реальном времени.

Стратегии улучшения качества данных. Для повышения качества данных и, как следствие, улучшения эффективности моделей обнаружения аномалий, можно применить несколько подходов:

- Очистка данных. Устранение ошибок, дубликатов и неполных записей для повышения точности данных.
- Обогащение данных. Использование дополнительных источников для увеличения полноты данных.

- Регулярное обновление данных. Обеспечение актуальности данных путем их регулярного обновления.

Качество данных играет критическую роль в разработке и эксплуатации моделей обнаружения аномалий. Вложения в улучшение качества данных могут значительно повысить точность и надежность этих моделей, делая их более эффективными в предотвращении потенциальных угроз или выявлении важных возможностей признаков данных [8]. Таким образом, качественные данные не только способствуют лучшему пониманию текущих условий, но и помогают в адаптации моделей к изменениям, что является ключом к их успешному применению.

Список литературы:

1. Chandola V. Anomaly detection: A survey / V. Chandola, A. Banerjee, V. Kumar // ACM Computing Surveys. – 2009. – Vol. 41, № 3. – Article 15. – DOI: 10.1145/1541880.1541882.
2. Aggarwal C. C. Outlier Analysis / C. C. Aggarwal. – 2nd ed. – Cham : Springer, 2017. – DOI: 10.1007/978-3-319-47578-3.
3. Sculley D. Hidden technical debt in machine learning systems / D. Sculley [et al.] // Advances in Neural Information Processing Systems. – 2015. – Vol. 28. – P. 2503–2511. – URL: <https://papers.neurips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems>. (дата обращения: 01.06.2025).
4. Breck E. Data validation for machine learning / E. Breck, N. Polyzotis, S. Roy, S. Whang, M. Zinkevich // Proceedings of Machine Learning and Systems. – 2019. – Vol. 1. – P. 334–347. – URL: https://proceedings.mlsys.org/paper_files/paper/2019/hash/928f1160e52192e3e0017fb63ab65391-Abstract.html. (дата обращения: 11.06.2025).
5. Wang R. Y. Beyond accuracy: What data quality means to data consumers / R. Y. Wang, D. M. Strong // Journal of Management Information Systems. – 1996. – Vol. 12, № 4. – P. 5–33. – DOI: 10.1080/07421222.1996.11518099.
6. Batini C. Data Quality: Concepts, Methodologies and Techniques / C. Batini, M. Scannapieco. – Berlin : Springer, 2006. – 262 p. – DOI: 10.1007/3-540-33173-5.
7. Gudivada V. N. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations / V. N. Gudivada, A. Apon, J. Ding // International Journal on Advances in Software. – 2017. – Vol. 10, № 1–2. – P. 1–20. – URL: https://personales.upv.es/thinkmind/dl/journals/soft/soft_v10_n12_2017/soft_v10_n12_2017_1.pdf. (дата обращения: 12.06.2025).

8. Протодяконов, А. В. Асимптотический анализ поведения прикладных моделей машинного обучения : учебное пособие / А. В. Протодяконов, А. В. Дягилева, П. А. Пылов. – Вологда : Инфра-Инженерия, 2023. – 144 с. – ISBN 978-5-9729-1455-5. – EDN APHQME.