

Дюсенов Дамир Сергеевич, магистрант
Российский новый университет
Dyusenov Damir Sergeevich, Master's student
Russian New University

Трефилова Ольга Леонидовна,
старший преподаватель кафедры информационных систем
в экономике и управлении, Российский новый университет
Trefilova Olga Leonidovna, Senior Lecturer at the Department
of Information Systems in Economics and Management,
Russian New University

ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА КАЧЕСТВА ВЫЯВЛЕНИЯ АНОМАЛИЙ В ЖУРНАЛАХ CI/CD МЕТОДОМ ISOLATION FOREST EXPERIMENTAL EVALUATION OF CI/CD LOG ANOMALY DETECTION USING ISOLATION FOREST

Аннотация. Представлены результаты экспериментальной оценки применения алгоритма Isolation Forest для обнаружения аномальных событий в журналах CI/CD. Исследовано влияние ожидаемой доли аномалий на точность и полноту обнаружения. Для выбранной конфигурации получено значение F1, равное 0,7761, при полноте 0,8667. Определены причины ошибочных решений и условия проведения пилотного внедрения.

Abstract. The paper presents an experimental evaluation of Isolation Forest for detecting anomalous events in CI/CD logs. The effect of the expected anomaly fraction on precision and recall is studied. The selected configuration achieves an F1 score of 0.7761 and a recall of 0.8667. Error causes and conditions for a pilot deployment are identified.

Ключевые слова: Isolation Forest, CI/CD, аномальные события, эксперимент, precision, recall.

Keywords: Isolation Forest, CI/CD, anomalous events, experiment, precision, recall.

Введение

Модель обнаружения аномалий полезна только тогда, когда ее результат можно количественно проверить и соотнести с нагрузкой на специалиста. Высокая полнота уменьшает вероятность пропуска отклонения, однако избыточное число ложных сигналов возвращает процесс к ручному просмотру. Поэтому выбор порога должен учитывать одновременно качество классификации и допустимое количество событий, требующих подтверждения.

Для журналов CI/CD задача осложняется недостатком размеченных промышленных данных и изменчивостью процессов. В учебном исследовании это ограничение компенсируется воспроизводимой синтетической выборкой, содержащей нормальные записи и контролируемые отклонения. Такой эксперимент не доказывает промышленную эффективность, но позволяет проверить корректность реализации, направление изменения метрик и пригодность методики оценки [1-3].

Цель статьи – экспериментально оценить качество Isolation Forest при выявлении аномальных событий CI/CD и выбрать рабочее значение ожидаемой доли аномалий. Проверяемая гипотеза состоит в том, что подбор contamination позволяет обеспечить recall не ниже 0,80 и F1 не ниже 0,70 при приемлемом количестве ложных срабатываний.

Организация эксперимента

Обучающая выборка включает 500 объектов, описывающих нормальное выполнение операций. Контрольная выборка состоит из 300 объектов: 240 нормальных и 60 аномальных. Отклонения формируются за счет увеличения длительности, редких кодов ответа, необычных сочетаний этапа и уровня сообщения. Обучающие и контрольные данные разделены; начальное состояние генератора случайных чисел зафиксировано значением seed 4401.



Основная конфигурация содержит 100 деревьев и размер подвыборки 64. Ожидаемая доля аномалий contamination равна 0,08. Все варианты сравниваются на одной контрольной выборке при неизменных деревьях и seed. Благодаря этому различия в результатах связаны с изменением порога, а не со случайным формированием новых данных.

Таблица 1

Параметры основного эксперимента

Параметр	Значение	Назначение
Обучающая выборка	500	Нормальные события
Контрольная выборка	300	240 нормальных и 60 аномальных
Число деревьев	100	Стабилизация ансамблевой оценки
Размер подвыборки	64	Число объектов для одного дерева
contamination	0,08	Доля объектов выше порога
seed	4401	Воспроизводимость результата

Для оценки используются precision, recall, F1 и accuracy. Precision показывает долю подтвержденных аномалий среди всех сигналов; recall – долю найденных аномалий; F1 является гармоническим средним precision и recall; accuracy отражает долю правильных решений по всей выборке. Из-за дисбаланса классов основными критериями выбраны recall и F1 [3, 4].

Результаты основного запуска

В основной конфигурации модель сформировала 74 сигнала. Из них 52 соответствовали аномальным объектам, а 22 оказались ложными срабатываниями. Восемь аномальных объектов не превысили порог. Матрица ошибок позволяет проверить опубликованные метрики: precision равно 52 / 74, recall – 52 / 60, а accuracy – 270 / 300.

Таблица 2

Результаты основного эксперимента

Показатель	Значение	Интерпретация
TP	52	Обнаруженные аномалии
FP	22	Ложные сигналы
FN	8	Пропущенные аномалии
TN	218	Верно распознанная норма
Precision	0,7027	Доля полезных сигналов
Recall	0,8667	Доля найденных аномалий
F1	0,7761	Баланс precision и recall
Accuracy	0,9000	Доля правильных решений

Результат удовлетворяет заданным критериям: recall превышает 0,80, а F1 – 0,70. Значение accuracy 0,90 не используется как единственное доказательство качества, поскольку преобладание нормальных объектов способно завышать этот показатель. Для процесса эксплуатации более важны число пропусков и количество сигналов, которые специалист должен проверить.



Распределение оценок в контрольной выборке представлено на рисунке 1. Вертикальная линия соответствует порогу 0,529. Большинство нормальных объектов находится левее порога, а большая часть аномалий – правее. Область пересечения классов объясняет наличие одновременно ложных срабатываний и пропусков.

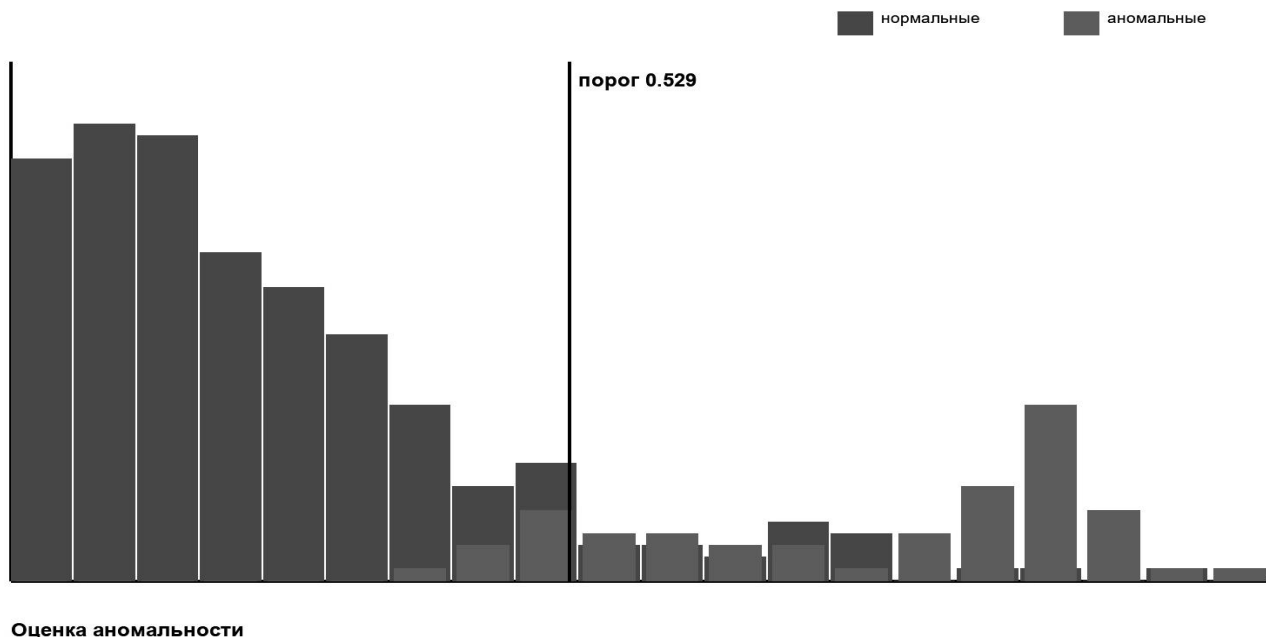


Рисунок 1. Распределение оценок аномальности в контрольной выборке

Смещение порога влево увеличивает recall, но одновременно расширяет очередь проверки. Смещение вправо уменьшает число ложных сигналов, однако повышает риск пропуска. Следовательно, выбор порога является управленческим компромиссом и зависит от критичности стенда, стоимости пропуска и доступного времени DevOps-команды.

Исследование параметра contamination

Для проверки устойчивости параметр contamination изменялся от 0,04 до 0,12 при неизменных данных, числе деревьев, размере подвыборки и seed. Сравнение показывает ожидаемую закономерность: при увеличении доли объектов выше порога recall растет, а precision снижается. Рабочее значение следует выбирать по заранее установленному критерию, а не по одной максимальной метрике.

Таблица 3

Сравнение вариантов ожидаемой доли аномалий

contamination	Precision	Recall	F1	Решение
0,04	0,833	0,667	0,741	Много пропусков
0,06	0,758	0,783	0,770	Недостаточный recall
0,08	0,703	0,867	0,776	Рабочий вариант
0,10	0,644	0,900	0,751	Больше сигналов
0,12	0,578	0,917	0,709	Высокая нагрузка

Максимальное F1 получено при contamination 0,08. Вариант 0,10 обеспечивает более высокий recall, но увеличивает число ложных сигналов. Для критичного выпуска такое соотношение может быть оправдано, тогда как для массовых некритичных сборок предпочтителен более строгий порог. Поэтому параметр следует задавать по классу систем и периодически пересматривать по подтвержденным событиям.

Сравнение метрик показано на рисунке 2. График используется не для повторения таблицы, а для отображения компромисса: линия recall возрастает, precision убывает, а F1 достигает локального максимума в средней части исследованного диапазона.

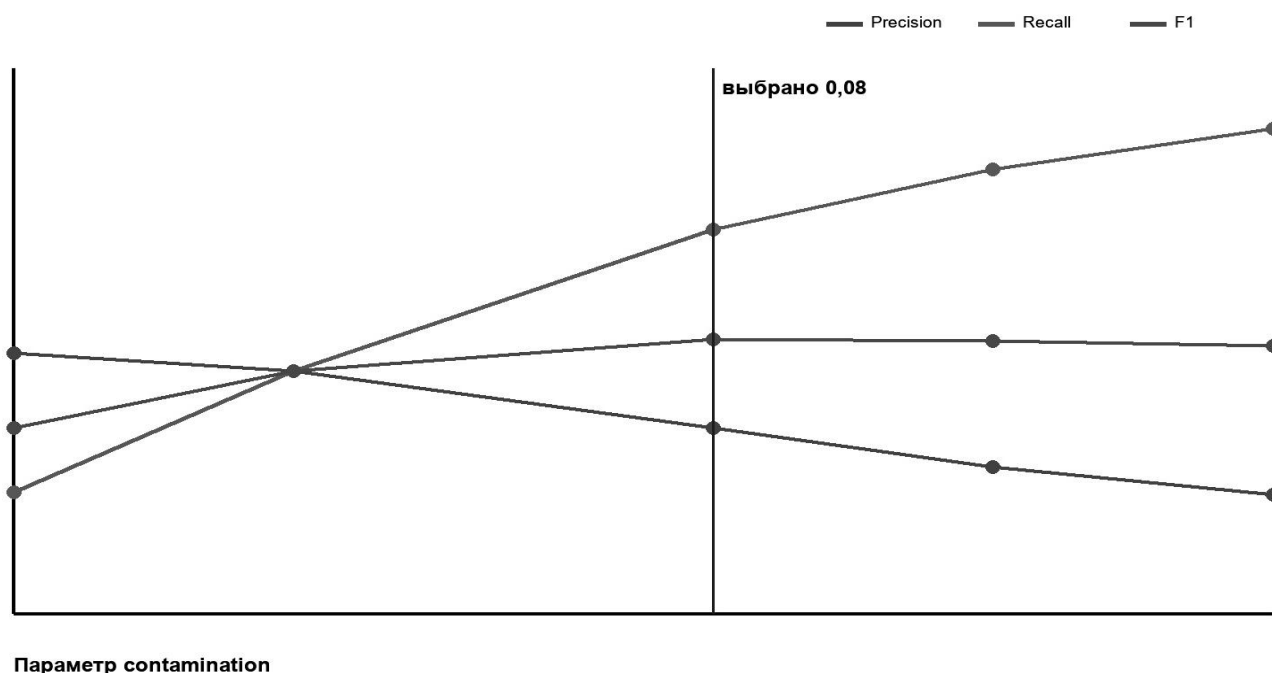


Рисунок 2. Изменение качества при выборе параметра contamination

Анализ ограничений и план пилота

Главным ограничением является синтетическая природа выборки. Внедренные аномалии сформированы по известным правилам и могут отделяться от нормы лучше, чем неизвестные промышленные случаи. Кроме того, используемые признаки описывают отдельную запись и слабо учитывают последовательности: одиночное событие может быть нормальным, тогда как серия повторов указывает на деградацию.

Следующий этап должен включать теневой пилот на обезличенных журналах одной системы. Модель рассчитывает оценки, но не блокирует выпуск. Специалисты подтверждают сигналы, после чего метрики вычисляются на временной контрольной выборке. Одновременно измеряются число сигналов на одну установку, время до подтверждения и доля событий, для которых сохранены версия модели и признаки.

Для приемки предлагаются четыре условия: recall не ниже 0,80; F1 не ниже 0,70; отсутствие секретов в сохраняемых данных; воспроизводимость каждой оценки. При невыполнении критериев расширение на промышленный стенд не проводится. Сначала уточняются признаки, сравниваются отдельные модели по стендам и проверяется влияние дрейфа данных [5-7].

Заключение

Эксперимент подтвердил работоспособность реализации Isolation Forest на контролируемой выборке журналов CI/CD. При 100 деревьях, подвыборке 64 и contamination 0,08 получены precision 0,7027, recall 0,8667, F1 0,7761 и accuracy 0,9000. Заданные критерии качества выполнены, а лучший баланс precision и recall в исследованном диапазоне достигнут при contamination 0,08.

Полученные значения следует рассматривать как результат учебного эксперимента, а не как оценку промышленной эффективности. Для обоснования внедрения необходимы реальные обезличенные данные, временное разделение выборок, сравнение с базовыми правилами и проверка качества после изменения инфраструктуры. Практическим результатом исследования является воспроизводимая методика, связывающая выбор порога с нагрузкой на специалиста и риском пропуска аномалии.

Список литературы:

1. Liu F. T., Ting K. M., Zhou Z.-H. Isolation Forest // 2008 IEEE International Conference on Data Mining. Pisa, 2008. P. 413-422. DOI: 10.1109/ICDM.2008.17.
2. Liu F. T., Ting K. M., Zhou Z.-H. Isolation-Based Anomaly Detection // ACM Transactions on Knowledge Discovery from Data. 2012. Vol. 6, No. 1. Article 3.
3. Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey // ACM Computing Surveys. 2009. Vol. 41, No. 3. Article 15.
4. Bishop C. M. Pattern Recognition and Machine Learning. New York: Springer, 2006. 738 p.
5. Scikit-learn developers. IsolationForest: API documentation. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html> (дата обращения: 17.09.2026).
6. Sculley D. et al. Hidden Technical Debt in Machine Learning Systems // Advances in Neural Information Processing Systems. 2015. Vol. 28.
7. NIST. Guide to Computer Security Log Management: Special Publication 800-92. Gaithersburg, 2006

