MLOps для ИБ

Искусственный интеллект все чаще используется для обеспечения информационной безопасности. Причин для этого несколько – от ускорения реакции на инциденты до увеличения нагрузки на ИБ-службу. Дефицит ИБ-специалистов и возрастание количества и сложности атак требуют от предприятий во всем мире применения тех или иных методов автоматизации принимаемых решений, анализа событий и реагирования на инциденты. Наиболее популярным методом повышения производительности служб ИБ является использование методов искусственного интеллекта в помощь специалистам по информационной безопасности. Строящиеся SOC также присматриваются к инструментам, использующим искусственный интеллект.

Однако с увеличением количества ML-моделей, которые заступили на ИБ-службу, возрастает и вероятность реализации нового типа рисков – так называемого модельного. Он возникает в том случае, когда искусственный интеллект начинает принимать неправильные решения – иногда это может привести к еще более опасным последствиям, чем пропущенная атака. Именно поэтому разработчики технологий искусственного интеллекта создали защиту от модельного риска, которая получила название MLOps – платформа эксплуатации моделей искусственного интеллекта с оценкой эффективности принимаемых ими решений.

Названная технология развивается прежде всего в банковской среде, где ИИ активно используют для повышения рентабельности. Первые ее внедрения есть и в российских банках. Однако для ИБ-компаний и SOC потребность в верификации корректности работы конвейера машинного обучения (ML-пайплайна) может оказаться даже более критичной, поскольку уязвимости в моделях ИИ могут использоваться и для нападения на тех клиентов, кто внедрил современные инструменты ИБ с применением искусственного интеллекта. Именно поэтому отрасли ИБ стоит присмотреться к MLOps, а возможно, и разработать свой MLSecOps.

Стандарт доверия к ИИ

«Росстандарт» уже озаботился снижением модельного риска и даже разработал и принял в 2019 г. специальный стандарт ГОСТ Р 59276-2020, который введен в действие с 1 марта 2021 г. Стандарт «Системы искусственного интеллекта: способы обеспечения доверия» определяет понятие доверия к системам искусственного интеллекта и методы его повышения. В стандарте приведена классификация факторов, влияющих на качество и способность систем искусственного интеллекта гарантированно реализовать свои основные функции на стадиях жизненного цикла, формализована взаимосвязь качества и способности систем искусственного интеллекта вызывать доверие, а также определена классификация основных способов обеспечения доверия к системам искусственного интеллекта. Именно потому перед рассказом о MLOps стоит обратиться к материалам указанного документа.

Одним из важных понятий, которые вводит стандарт, является термин «предвзятость» или «необъективность искусственного интеллекта». В глоссарии стандарта эта характеристика ИИ определяется как свойство системы искусственного интеллекта, заключающееся в принятии ошибочных решений, которые связаны со статистической смещенностью обучающей выборки исходных данных. То есть предвзятость искусственного интеллекта очень сложно определить по уже построенной модели, поскольку она размазана по всем коэффициентам и сложно понять, в каком направлении в ней сформирована предвзятость.

Например, если мы используем модель для определения самолетов, но взяли предобученную библиотеку, где в обучающей выборке все изображения стелс-самолетов размечены как птицы, то научить такую модель распознавать стелс-самолеты будет довольно сложно, поскольку у нее уже выработана к ним предвзятость – модель считает их птицами. Сейчас редко кто самостоятельно с нуля обучает модели распознавания изображений или готовит для них гарантированную обучающую последовательность – обычно для этого используют уже разработанные кем-то решения, доказавшие свою эффективность. Однако и у таких моделей могут быть свои предвзятости. Выяснить их и помогает MLOps за счет постоянной оценки качества работы моделей.

В стандарте также определены факторы, которые влияют на доверие к системам искусственного интеллекта. Они привязаны к стадиям жизненного цикла модели. Рассмотрим каждую из них в отдельности с перечислением факторов, которые могли бы повлиять на эффективность работы ML-модели.

- **Концепция**. На стадии проектирования может быть

недостаточная полнота выбранного набора функциональных характеристик – прикладных, безопасности, надежности и др. Это не позволяет считать выбранный набор характеристик представительным.

- **Разработка**. В процессе разработки ML-модели могут возникнуть: недостаточная представительность обучающей выборки, использованной при создании ML-модели; смещенность обучающей выборки, способная привести к предвзятости (необъективности) результатов работы; неоптимальность используемой модели данных; недостаточный уровень унификации и низкая интероперабельность разрабатываемой системы.
- **Производство**. Среди проблем, которые могут возникнуть на этой стадии, стандарт определяет недостаточную надежность создаваемой ML-модели, чрезмерную стоимость владения решением, недостаточную понятность, объяснимость, предсказуемость и др., а также недостаточную защищенность информации о модели данных.
- **Эксплуатация** (применение по назначению). На этой стадии ML-модель может работать неэффективно в случае применения ее не по назначению, недостаточной представительности выборки, используемой при тестировании модели; редких тестирований характеристик модели; отсутствия средств автоматического самотестирования после каждого обучения или дообучения системы ИИ; недостаточной защищенности информации о функционировании и модели данных, используемых в модели. Не стоит забывать и о защищенности обрабатываемых персональных данных.
- **Поддержка**. На этапе обслуживания модели может произойти утрата актуальности модели данных, например, в связи с изменившейся рыночной ситуацией.
- **Прекращение применения**. При выводе решения из эксплуатации может быть нарушена конфиденциальность персональных данных.

Таким образом, наиболее сложные проблемы по обеспечению эффективности работы системы искусственного интеллекта возникают на стадиях эксплуатации и поддержки. Именно на автоматизацию этих процессов и направлена технология MLOps – она предназначена для контроля качества моделей не только во время разработки, но и в процессе эксплуатации и переобучения. В стандарте перечислены методы, с помощью которых можно снизить влияние каждого фактора и воспользоваться рекомендациями стандарта при проектировании собственной системы ИИ. Причем для факторов недостаточной периодичности тестирования характеристик модели и отсутствия средств автоматического самотестирования после каждого обучения или дообучения системы ИИ в стандарте нет отдельных рекомендаций – понятно, что для минимизации этих факторов нужно создавать специальную автоматизированную систему тестирования и эксплуатации, которая строится именно по технологи MLOps.

> MLOps предназначена для контроля качества моделей не только во время разработки, но и в процессе эксплуатации и переобучения.

Эксплуатация ИИ

В соответствии с определением Wikipedia MLOps (MLOps) – это набор методов, направленных на надежное и эффективное развертывание и поддержку моделей машинного обучения в производственной среде. MLOps подразумевает следующий жизненный цикл моделей искусственного интеллекта.

1. **Определение задач ML**. Разработка моделей искусственного интеллекта традиционно начинается с формулировки задачи с четко определенными показателями продуктивности – KPI. Для ИБ это может быть снижение ложных срабатываний первого и второго рода до приемлемых показателей, а для антифрод-решений – процент мошеннических транзакций. В качестве целей можно использовать и соблюдение законодательных требований. В частности, в стандарте указано три источника требований: разработчики ML-моделей, их потребители и регуляторы соответствующей предметной области. В ИБ достаточно много регламентных документов, требования которых необходимо соблюдать, поэтому логично именно на их основе и строить контрольные метрики для оценки качества работы моделей искусственного интеллекта.
2. **Проектирование ИИ**. После того как цели функционирования ИИ и его KPI определены, построены все модели угроз, нарушителей и объекта защиты, можно приступать к проектированию самих моделей. Для этого необходимо провести поиск подходящих входных данных и типов моделей, которые можно построить на таких данных. Конечно, для производителей средств защиты сбор первичных данных для обучения упрощается тем, что они, как правило, уже достаточно давно собирают сведения по атакам на всех своих клиентов и могут выполнить

требования по полноте набора данных. Клиентам, которые хотят контролировать качество детектирования построенной сторонним разработчиком модели, придется собирать данные самостоятельно. Если данных не очень много – не отчаивайтесь. Можно использовать различные инструменты симуляции атак, проверка на которых гарантирует, как минимум, отсутствие предвзятости от наиболее популярных методов атаки. Такие генераторы и сканеры уязвимостей стоит предусмотреть при проектировании системы MLOps для ИБ.

3. **Дата-инжиниринг**. Построение собственно ML-моделей включает в себя такие задачи, как конструирование признаков, очистка данных (форматирование, проверка на выпадающие значения, ребалансировка и т. д.), а затем выбор того набора признаков, который будет максимально релевантным основной задаче. На этом этапе определяется полный вид конвейера, т. е. всего ML-пайплайна, а затем необходимо наполнить его полученными на предыдущем этапе чистыми и размеченными данными. Важной частью развертывания таких конвейеров является выбор правильной комбинации облачных сервисов и архитектуры, сочетающей производительность и рентабельность. Как правило, процессы обучения и переобучения моделей требуют достаточно высоких вычислительных затрат, которые лучше закупать у поставщиков PaaS-услуг.
4. **Создание ML-моделей**. Как только определится эскиз всего конвейера обработки данных с помощью методов искусственного интеллекта, можно переходить к следующему этапу – собственно обучению ML-моделей. Вряд ли хороших результатов можно будет добиться с первого раза – скорее всего, придется проверить работу нескольких типов моделей для выбора лучшей модели, максимально полно соответствующей метрикам, определенным на первом этапе разработки MLOps-системы. Еще одна проблема, с которой придется столкнуться на этом этапе, – воспроизводимость результатов. Эту проблему можно решить с помощью инструмента контроля версий как самих ML-моделей, так и наборов данных, используемых для обучения. Сейчас разработано достаточно много инструментов контроля версий, причем есть и проекты с открытыми кодами, такие как DVC и CML. Кроме самих моделей необходимо разработать тесты для оценки характеристик ML-моделей, соответствующие установленным на первом этапе KPI, а также аналогичные характеристики других моделей, чтобы вовремя обнаружить как ухудшение качества текущей модели, так и улучшение характеристик альтернативных.
5. **Построение ML-пайплайнов**. При создании собственно конвейеров машинного обучения (ML-пайплайнов) следует учитывать системные требования для их функционирования, правильно настроить элементы локального и облачного компонентов, провести собственно обучение ML-моделей и выбранных на предыдущем этапе тестов, проверить достижение с помощью работы построенного конвейера заранее определенных характеристик отдельных моделей и системы целиком, провести проверки данных и элементов систем защиты.
6. **Эксплуатация ML-моделей**. Существует два способа эксплуатации ML-моделей: статическая, когда модель встраивается в устанавливаемое ПО, а затем эксплуатируется, и динамическая – модель встраивается в веб-фреймворк типа FastAPI или Flask и к ней организуется доступ через API. Примером статической модели являются, например, песочницы, которые устанавливаются у клиента и пытаются

с помощью обученной модели определить опасность того или иного файла. Пример динамической модели развертывания – всевозможные сервисы разработчиков СЗИ, в частности репутационные спам-фильтры, – им отправляются параметры письма, а в ответ приходит вердикт, является письмо спамом или нет. У каждого типа эксплуатации есть свои преимущества и недостатки, и компоненты системы MLOps тоже от этого зависят. Так, для установленной у заказчика статической модели необходимо разработать специальные инструменты тестирования и переобучения. Динамическую модель контролировать и тестировать значительно проще. Однако статическая модель может более точно соответствовать потребностям заказчика, поскольку будет учитывать локальную специфику.

7. **Мониторинг, оптимизация, переобучение**. В процессе эксплуатации качество работы ML-моделей обычно ухудшается – ситуация меняется, появляются новые факторы, и отработанные решения перестают приносить результаты, особенно в сфере информационной безопасности, где постоянно существует противостояние между защитниками и атакующими. Поэтому в процессе эксплуатации решений искусственного интеллекта приходится регулярно проверять соответствие текущих результатов работы критериям, определенным в начале проекта. Важно не допустить резкой деградации качества работы искусственного интеллекта и провести его дообучение на текущих данных, в которых будут содержаться сведения о текущей обстановке. Для этого необходимо выполнять следующие задачи: отслеживать производительность и KPI моделей; контролировать установленные метрики непрерывной оценки; выявлять сбои в работе моделей; оптимизировать производительность их работы и обучения; контролировать воспроизводимость результатов и работы моделей у всех клиентов и во всех локациях. Именно решение перечисленных и целого ряда других задач и будет направлено на выявление и локализацию проблем, чтобы своевременно и с минимальными затратами обеспечить качественное функционирование системы искусственного интеллекта.

Для реализации жизненного цикла ML-моделей используются различные инструменты автоматизации – в совокупности они и создают систему MLOps, которая позволяет гарантировать постоянную работоспособность всех моделей искусственного интеллекта в соответствии с заранее определенными метриками. Она дает возможность строить и эксплуатировать доверенный искусственный интеллект в соответствии с описанным выше ГОСТом. Для сферы информационной безопасности соблюдение регуляторных требований, хотя и таких рекомендательных, как ГОСТ, является определенным подспорьем в реализации лучших мировых практик.

> Для сферы информационной безопасности соблюдение регуляторных требований, является определенным подспорьем в реализации лучших мировых практик.

Пример «Открытия»

Банковская сфера сейчас находится на передовой по использованию искусственного интеллекта. Некоторые банки уже построили системы MLOps и даже успешно эксплуатируют решения. В частности, банк «Открытие» развернул MLOps-платформу для разработки и эксплуатации моделей машинного обучения IRIS от российской компании Neoflex (источник https://globalcio.ru/projects/18692/). Она базируется на продуктах с открытыми кодами компании Databricks и других лидеров интеграции данных, моделей и процессов. Архитектура платформы IRIS включает в себя развернутый контур разработки и контур применения моделей: разработка ведется в отдельном окружении, при этом прошедшие тестирование модели могут быть переданы в эксплуатацию в любой момент, практически без ручных операций.

Внедрение платформы было завершено в октябре 2021 г., что и позволило банку успешно пройти финансовые пертурбации. Платформа MLOps обеспечила возможность вовремя заметить изменение пользовательского поведения и оперативно подстроить конвейер принятия решений в банке. Индустрия информационной безопасности тоже начинает внедрять технологии искусственного интеллекта. Кроме того, крупные заказчики, которым приходится активно автоматизировать системы принятия решений, в том числе в области информационной безопасности – в составе SOC, начинают использовать ML-продукты для выявления новых угроз и атак. У них также может возникнуть потребность построения подобных MLOps-платформ. Следует отметить: в России есть производители и интеграторы, которые помогут построить подобные ML-конвейеры с контролем их характеристик в соответствии со ГОСТом доверия к искусственному интеллекту. ■

Валерий Коржов,
Connect