УДК 004.89

АВТОМАТИЧЕСКОЕ ОБНАРУЖЕНИЕ РИСКОВ РАЗГЛАШЕНИЯ ДАННЫХ В ПАЙПЛАЙНЕ МАШИННОГО ОБУЧЕНИЯ

Каплан М., студент направления MSc Computer Science, II курс
Королевский технологический институт, г. Стокгольм, Швеция

Автоматическое обнаружение рисков разглашения данных в пайплайне машинного обучения становится всё более актуальной задачей в условиях масштабной интеграции ИИ в сферы, где обрабатываются чувствительные или персональные данные [1].

Такие риски возникают не только на этапе финального использования модели, но и на всех стадиях жизненного цикла машинного обучения – от подготовки данных до деплоя и мониторинга модели в продуктивной среде.

Это делает необходимым развитие автоматизированных инструментов и протоколов, способных отслеживать потенциальные утечки конфиденциальной информации в реальном времени, предупреждать разработчиков и администраторов, а также формировать доказательную базу для последующего аудита.

Под пайплайном машинного обучения подразумевается последовательность этапов, включающая сбор и валидацию данных, их препроцессинг (очистку, нормализацию, трансформации), разбиение на обучающую и тестовую выборки, обучение и валидацию модели, гиперпараметрическую оптимизацию, сериализацию и деплой, а также постэксплуатационный мониторинг. Каждый из этих этапов может стать вектором атаки или причиной утечки данных – как по технической (например, уязвимости в реализации), так и по логической (например, недостаточная анонимизация) причине [2].

Автоматическое обнаружение рисков в таком пайплайне должно основываться на сочетании формальных критериев, статического и динамического анализа, а также эвристических и машинных методов оценки конфиденциальности. Одним из ключевых подходов является анализ потока данных: отслеживание перемещения, трансформации и использования каждого поля или признака, начиная с исходной базы данных и заканчивая входами модели. Это позволяет фиксировать, какие именно приватные поля сохраняются, агрегируются или участвуют в обучении, и как они представлены в параметрах модели.

Особое внимание в автоматизированных системах мониторинга уделяется идентификации полей с высоким риском приватности – таких как идентификаторы, даты рождения, диагнозы, GPS-координаты и т. п. Система должна быть способна распознавать такие поля независимо от их наименования и

структуры, используя методы семантического анализа схемы данных, включая машинное обучение для классификации атрибутов по приватности. Далее, важным компонентом становится оценка воздействия этих признаков на поведение модели: если они оказывают существенное влияние на результат, то даже косвенные утечки через поведение модели могут представлять угрозу.

На стадии обучения модели автоматическое обнаружение рисков может включать анализ градиентов, отслеживание чувствительности выходов модели к приватным полям, а также метрики запоминания данных– например, проверку вероятности воспроизведения уникальных примеров из обучающей выборки. В генеративных моделях (таких как языковые или образные) это дополнительно сопровождается тестами на меморизацию и инференцию: система может автоматически запускать атаки восстановления данных на случайно выбранных шаблонах, чтобы проверить степень уязвимости [3].

В продуктивных условиях особую роль играет анализ логов запросов к модели: автоматические детекторы должны идентифицировать аномальные шаблоны использования API, подозрительную частоту запросов, избыточную последовательность вариаций одного и того же примера (что характерно для атак на извлечение данных). Также эффективны механизмы контекстной сигнализации, когда система сопоставляет входные данные и поведение модели с накопленными профилями нормального использования и генерирует тревожные сигналы при отклонениях.

Существуют также методы символьного и формального анализа пайплайна, позволяющие не просто регистрировать уже произошедшие утечки, но и заранее находить потенциальные уязвимости в конфигурации и логике операций. Например, если приватное поле не было должным образом исключено из модели, или если использовался компрометированный механизм нормализации, система может сигнализировать об этом на стадии проектирования. Инструменты типа DataLint, PrivacyTracer или TensorFlow Data Validation уже сегодня реализуют часть этой функциональности.

Для полной интеграции в реальные бизнес-процессы автоматические системы анализа рисков должны обладать модульной архитектурой, позволяющей встраиваться в существующие CI/CD-пайплайны машинного обучения.

Они должны взаимодействовать с системами контроля версий данных (Data Version Control), трекерами экспериментов (MLflow, Weights & Biases), системами мониторинга (Prometheus, Grafana), а также с политиками доступа и классификации данных в корпоративных хранилищах [4].

Таким образом, автоматическое обнаружение рисков разглашения данных в пайплайне машинного обучения – это не изолированная техническая задача, а стратегическая функция в архитектуре доверенного ИИ.

Она требует объединения знаний из области инженерии данных, кибербезопасности, теории информации и правового регулирования.

Только на базе такой интеграции возможно создание полноценных систем, способных не только обнаруживать утечки, но и предупреждать их, обеспечивая соответствие требованиям приватности, надёжности и прозрачности как в академических, так и в индустриальных приложениях [5].

Список литературы:

1. Пылов, П. А. Teamlead как разработчик и юридический лидер команды в одном лице / П. А. Пылов, А. В. Протодьяконов, А. И. Бобровских // Россия молодая : сб. материалов XIV Всероссийской научно-практической конференции с международным участием, Кемерово, 19–21 апреля 2022 года. – Кемерово : Кузбасский государственный технический университет имени Т. Ф. Горбачева, 2022. – С. 94406.1–94406.7. – EDN TOAHDH.

2. European Data Protection Board. Guidelines 4/2019 on Article 25 Data Protection by Design and by Default. – Brussels : EDPB, 2020. – 30 p.

3. Article 29 Data Protection Working Party. Guidelines on Data Protection Impact Assessment (DPIA). – Brussels : European Commission, 2017. – 21 p.

4. ENISA. Data Protection Engineering: From Theory to Practice. – Heraklion : European Union Agency for Cybersecurity, 2022. – 62 p.

5. ENISA. Recommendations on Shaping Technology According to GDPR Provisions. – Heraklion : European Union Agency for Cybersecurity, 2018. – 50 p.