Секция «43.9 Управление рисками и страхование: новые вызовы и возможности»

Оптимальная интенсивность мониторинга ИИ-моделей при ненулевой вероятности критической ошибки

Заявка № 1670096

В настоящее время в различных отраслях экономики активно внедряются технологии искусственного интеллекта (ИИ) в управленческие и операционные процессы, и поэтому решения всё чаще опираются на результаты ИИ-моделей. При этом качество модели в эксплуатации может ухудшаться из-за изменения входных данных и условий применения, накопления ошибок в данных и изменений в поведении пользователей. На уровне управления это проявляется как риск критической ошибки, при котором модель начинает систематически выдавать неверные рекомендации или ошибочные классификации, а организация несёт потери до момента выявления деградации и принятия корректирующих мер. В связи с этим рассматривается прикладная модель выбора оптимальной интенсивности мониторинга ИИ-моделей при ненулевой вероятности критической ошибки и эмпирически демонстрируется существование оптимума. Под интенсивностью мониторинга понимаются частота управленческих проверок качества (например, периодичность по времени или по числу обработанных случаев), а также доля проверяемых кейсов. Данный подход опирается на принципы риск-менеджмента и организационного управления ИИ (ГОСТ Р ИСО 31000-2019; ГОСТ Р ИСО/МЭК 42001-2024) [1, 2], а также на рекомендации по управлению рисками ИИ (ISO/IEC 23894:2023; NIST AI RMF 1.0) [6, 7]. Если сформулировать управленческую задачу как выбор режима мониторинга, минимизирующего суммарные затраты жизненного цикла модели, то необходимо оценить два ключевых компонента. Во-первых, это затраты на контроль (время специалистов, подготовка и проверка контрольных данных, вычислительные ресурсы, аудит). Во-вторых, это ожидаемые потери от критической ошибки, связанные с задержкой обнаружения деградации. Возникает компромисс: частые проверки снижают вероятность длительного пребывания модели в критическом состоянии, но увеличивают текущие расходы на контроль; редкие проверки уменьшают расходы на контроль, но повышают ожидаемые потери из-за позднего обнаружения. Чтобы проверить идею на практике, используются искусственно созданные данные, в которых дважды резко меняются условия работы. Это имитирует ситуацию, когда со временем модель начинает работать хуже, потому что «правила игры» изменяются. В эксперименте применяется простая модель классификации. После запуска модель обновляется только при принятии решения о вмешательстве, то есть заново обучается на наиболее свежих данных. Сбой считается критическим, если доля правильных ответов устойчиво опускается ниже заданного уровня; качество оценивается по последним 400 примерам. Далее сравниваются два подхода к мониторингу: первый предполагает проверки по расписанию через фиксированные интервалы, второй основан на постоянном малозатратном наблюдении за качеством и проведении полной проверки только при появлении признаков ухудшения. Для календарного режима суммарная стоимость рассчитывается как сумма затрат на проверки, затрат на вмешательства (переобучение, перенастройка, изменение регламента) и потерь за время, пока модель находится в критическом состоянии до выявления. При заданных параметрах (в условных единицах) стоимость одной проверки равна 1, стоимость вмешательства равна 40, а потери за единицу времени деградации равны 0,1. В этих условиях минимум суммарной стоимости достигается при проверке каждые 200 шагов: при интервале 100 шагов суммарная стоимость равна 539,2; при интервале 200 шагов — 504,1; при интервале 400 шагов — 677,8. Результат показывает наличие оптимальной интенсивности мониторинга, а не монотонную зависимость по принципу «чем чаще, тем лучше». Для событийного режима предполагается малая сто-

имость «лёгкого» контроля на каждом шаге (0,01 на шаг) и срабатывание триггера при устойчивом ухудшении индикатора качества. В эксперименте событийная политика даёт суммарную стоимость 475,0 и оказывается предпочтительнее календарной при тех же прочих параметрах. Такой подход согласуется с практиками детектирования дрейфа, включая адаптивное окно (ADWIN) [3] и контроль ошибки (DDM) [4]. Полученные результаты можно интерпретировать как управленческую рекомендацию: при ограниченных ресурсах целесообразно сочетать редкие плановые проверки с недорогими триггерами, сокращая задержку обнаружения критических ошибок без непропорционального роста затрат на контроль. Практическая ценность работы состоит в алгоритме настройки мониторинга в организации и включает следующие шаги. 1) Фиксация порога критичности (показатели качества модели и допустимая длительность деградации) в системе менеджмента ИИ [2]. 2) Оценка затрат на контроль, вмешательства и потерь от деградации на основе трудозатрат и классификации инцидентов. 3) Выбор оптимального интервала и режима мониторинга (календарный, событийный или гибридный). 4) Встраивание процедуры в систему менеджмента риска и отчётности [1, 6, 7]. В дальнейшем планируется валидация подхода на открытых наборах данных (например, OpenML Airlines) [8] и в прикладных организационных кейсах.

Источники и литература

1) ГОСТ Р ИСО 31000-2019. Менеджмент риска. Принципы и руководство.
2) ГОСТ Р ИСО/МЭК 42001-2024. Информационные технологии. Искусственный интеллект. Система менеджмента.
3) Bifet A., Gavaldà R. Learning from Time-Changing Data with Adaptive Windowing. 2007.
4) Gama J., Medas P., Castillo G., Rodrigues P. Learning with Drift Detection. 2004.
5) Gama J., Žliobaitė I., Bifet A., Pechenizkiy M., Bouchachia A. A survey on concept drift adaptation. ACM Computing Surveys. 2014.
6) ISO/IEC 23894:2023. Information technology — Artificial intelligence — Guidance on risk management.
7) NIST AI 100-1. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023.
8) OpenML. Airlines dataset (ID 1169). URL: https://www.openml.org/d/1169 (дата обращения: 28.02.2026).