Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

Machine Learning в продакшене: почему модели умирают без MLOps

Почему точная модель тихо деградирует и приносит убытки — и как выстроить надёжный ML-контур

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

О ЧЁМ ЭТОТ РАЗБОР

Суть проблемы

Дата-сайентист обучил модель, метрики на тесте отличные — но модель в Jupyter-ноутбуке ещё не продукт. Без версионирования, мониторинга и переобучения она месяцами не доезжает до эксплуатации, а доехав — тихо деградирует: данные «плывут», точность падает, и вчерашние верные прогнозы превращаются в убыточные решения, о которых бизнес узнаёт по упавшей выручке.

Почему это важно бизнесу

- Модель решает про деньги: цены, скоринг, прогноз спроса. Её тихая деградация — прямые убытки, а не «техническая мелочь»
- Без MLOps вывод одной модели занимает месяцы: вы платите зарплату DS-команде, а бизнес-эффекта нет
- Ошибку модели видно не сразу: сначала падает выручка, и только потом находят причину
- Обучающие данные часто содержат персональные данные — их обработка должна соответствовать 152-ФЗ
- Если модель живёт в ноутбуке одного специалиста, его уход парализует сервис

ЧТО МЫ ЗАКЛАДЫВАЕМ

Ключевые параметры реализации

PSI > 0.2

порог индекса стабильности популяции по признаку, при котором детектор дрифта поднимает алерт…

≤ 5 мин

целевое время отката ML-модели на предыдущую стабильную версию через реестр моделей и алиасы

100%

входных запросов и предсказаний логируем с версией модели — для разбора инцидентов и аудита

24/7

мониторинг качества предсказаний на живом трафике, а не только метрик сервера (CPU, RAM)

p95 ≤ 200 мс

целевая задержка ответа inference-API под нагрузкой с горизонтальным масштабированием реплик

≥ 2

реплики inference-сервиса за балансировщиком для отказоустойчивости и rolling update без прост…

РЕАЛИЗАЦИЯ 1 ИЗ 3

Автоматическое ML-ценообразование без детектора дрифта и стоп-лимитов

Что настраиваем

Технический сценарий: модель оценки активов назначает цену закупки и запускает сделки без ручного контроля

Как мы это делаем

1. Модель оценки автоматически назначает цену закупки актива и инициирует сделки на потоке
2. Рыночные условия разворачиваются, но распределение входных признаков уходит от обучающего — детектора дрифта нет
3. Модель продолжает завышать оценку: активы скупаются дороже реальной стоимости, стоп-лимита на суммарный риск нет
4. Деградацию замечают по убыткам, направление экстренно сворачивают и распродают активы ниже цены покупки

РЕЗУЛЬТАТ

Прямые убытки на масштабе автоматических сделок, экстренная остановка и сворачивание направления. Корень — отсутствие мониторинга дрифта распределения и жёстких стоп-лимитов на решения, доверенные модели.

КЛЮЧЕВОЙ НЮАНС

Чем крупнее решения доверены алгоритму, тем жёстче нужны стоп-лимиты и постоянный контроль качества предсказаний на живых данных. Модель без детектора дрифта не заметит разворот рынка.

РЕАЛИЗАЦИЯ 2 ИЗ 3

Испорченные данные в обучающем пайплайне ломают таргетинг

Что настраиваем

Технический сценарий: рекламная/скоринговая ML-модель без валидации входных данных

Как мы это делаем

1. В обучающий пайплайн модели попадают испорченные данные крупного источника
2. Валидации схемы и диапазонов входных данных нет — мусор проходит в обучение молча
3. Точность таргетинга падает, но техметрики сервера в норме — деградацию по мониторингу не видно
4. Проблему обнаруживают только по снижению выручки, затем недели уходят на пересборку пайплайна данных

РЕЗУЛЬТАТ

Потеря выручки и доверия пользователей инструмента, месяцы на пересборку пайплайна данных. Причина — отсутствие валидации входных данных и мониторинга качества предсказаний.

КЛЮЧЕВОЙ НЮАНС

«Мусор на входе» ломает модель незаметно. Валидация входных данных и мониторинг качества предсказаний должны срабатывать раньше, чем квартальная финансовая отчётность.

РЕАЛИЗАЦИЯ 3 ИЗ 3

Резкий сдвиг спроса обрушивает точность модели

Что настраиваем

Технический сценарий: прогноз наличия/спроса при шоковом изменении поведения пользователей

Как мы это делаем

1. Модель прогнозирует наличие товаров/спрос с высокой точностью на стабильном рынке
2. Резкий шок спроса ломает прежние закономерности — распределение данных смещается за считаные дни
3. Точность прогноза падает в полтора раза и более, заказы собираются с ошибками
4. Команда сокращает цикл переобучения с недель до нескольких дней и восстанавливает качество

РЕЗУЛЬТАТ

Всплеск ошибок в пиковый момент спроса и недовольство клиентов на фоне взрывного роста нагрузки. Спасает то, что деградацию ловят по метрикам модели и быстро ускоряют цикл переобучения.

КЛЮЧЕВОЙ НЮАНС

Дрифт данных — вопрос не «если», а «когда». Выигрывает не самая точная модель, а та, вокруг которой построен быстрый конвейер мониторинга и переобучения.

ЧТО МЫ ВЫУЧИЛИ НА ПРАКТИКЕ

Подводные камни

- ✕ **Модель живёт в ноутбуке дата-сайентиста**
  Деплой — ручное копирование .pkl-файла на сервер. Невоспроизводимо, завязано на одного человека; «у меня работает» — не аргумент.

- ✕ **Нет версионирования моделей и данных**
  Файлы вида model_final_v3_FIXED.pkl: невозможно понять, что крутится в проде, как это воспроизвести и на что откатиться.

- ✕ **Мониторят сервер, а не модель**
  CPU и память в норме, а точность предсказаний упала вдвое — этого никто не видит, пока не просядут бизнес-показатели.

- ✕ **Вера в метрики на тестовой выборке**
  Точность на историческом датасете не гарантирует качества на живых данных: распределение признаков в проде смещается, и модель ошибается на новых сег…

- ✕ **Нет валидации входных данных**
  Испорченные или аномальные данные молча ломают предсказания, а деградацию замечают только по упавшим бизнес-показателям — с задержкой в квартал.

- ✕ **Нет плана отката**
  Новая версия модели оказалась хуже, а быстро вернуть старую нельзя: не сохранены артефакты, окружение и версии фичей.

- ✕ **Игнорирование дрифта данных**
  Со временем деградирует практически любая модель. Без детектора дрифта (PSI/KL по признакам) и регулярного переобучения точность тает незаметно месяц…

- ✕ **Нет владельца модели в эксплуатации**
  Дата-сайентист обучил и ушёл на новый проект; за инциденты, обновления и SLA ML-сервиса не отвечает никто.

УРОВНИ РЕАЛИЗАЦИИ

Как правильно

МИНИМУМ

- Заведите реестр моделей (MLflow Model Registry): версии, данные, параметры, алиасы в…
- Упакуйте модель в контейнер (Docker, multi-stage): зафиксируйте зависимости и окруже…
- Логируйте входные данные и предсказания с версией модели для разбора инцидентов
- Назначьте ответственного (owner) за каждую модель в эксплуатации

НОРМАЛЬНО

- Мониторинг дрифта (Evidently: PSI, KL, Wasserstein) и качества модели с автоалертами…
- CI/CD-пайплайн деплоя модели с тестами и откатом на предыдущий алиас за минуты
- Валидация схемы и диапазонов входных данных до подачи в модель
- Теневой запуск или A/B-тест новой версии до переключения трафика

ХОРОШО

- Автопереобучение по сигналу дрифта с гейтом качества — не выкатывать модель хуже тек…
- Feature store (Feast) и единый MLOps-контур: реестр, serving, мониторинг
- Отказоустойчивый inference (Kubernetes): ≥2 реплики за балансировщиком, батчинг, SLA…
- Регулярный аудит бизнес-эффекта модели, а не только техметрик

ПРОВЕРЬТЕ У СЕБЯ

Чек-лист самопроверки

- [ ] Знаете ли вы, сколько ML-моделей работает в ваших процессах и кто за каждую отвечает?
- [ ] Сможете ли воспроизвести любую продакшен-модель: код, данные, параметры обучения?
- [ ] Узнаете ли вы о падении точности модели раньше, чем о падении выручки?
- [ ] Есть ли автоматический контроль дрифта входных данных и предсказаний?
- [ ] Сможете ли откатить модель на предыдущую версию за минуты, а не дни?
- [ ] Проверяются ли входные данные на качество и аномалии перед подачей в модель?
- [ ] Соответствует ли обработка обучающих данных 152-ФЗ: обезличивание, согласия, локализация?
- [ ] Есть ли SLA на время ответа и доступность ML-сервиса?
- [ ] Тестируете ли новую модель на части трафика перед полным переключением?
- [ ] Переживёт ли ML-сервис уход дата-сайентиста, который его создал?

Если хотя бы на два вопроса ответ «нет» или «не знаю» — тема требует внимания.

ЧТО МЫ ДЕЛАЕМ

Как поможет ITFresh

ITFresh — ИТ-аутсорсинг для юридических лиц до 50 рабочих мест в Москве и области. 15+ лет практики, собственная инфраструктура в дата-центре МТС (8 серверов Dell Xeon Platinum).

- Аудит ML-контура: воспроизводимость, мониторинг, риски деградации, соответствие 152-ФЗ
- Внедрение MLOps на открытых и российских инструментах: реестр моделей, CI/CD, мониторинг дрифта
- Построение отказоустойчивого ML API: контейнеры, реплики, алерты, план отката
- Сопровождение ML-сервисов: обновления, переобучение, SLA и реагирование на инциденты

15+
лет в ИТ-поддержке

50
рабочих мест — наш профиль

МТС
дата-центр, Москва

КОНТАКТЫ

Обсудить вашу задачу

Сайт	**itfresh.ru**
Телефон	**+7 903 729-62-41**
Telegram	**@ITfresh_Boss**

Бесплатно посмотрим вашу инфраструктуру по этому чек-листу и скажем, где тонко — без обязательств.

itfresh.ru

ДОКУМЕНТАЦИЯ И НАШ СТЕК

Техническая база

01 MLflow Model Registry — версии, алиасы и жизненный цикл модели (mlflow.org — 2026 (ML...)

02 Docker — сборка воспроизводимых образов (Dockerfile, multi-stage) (docs.docker.com — 2026)

03 Kubernetes — Deployments, реплики и rolling update inference-сервиса (kubernetes.io — 2026 (v1...)

04 Prometheus — сбор метрик качества модели и правила алертинга (prometheus.io — 2025 (v3...)

05 Grafana — дашборды мониторинга ML-сервиса и дрифта (grafana.com — 2026 (v1...)

06 Evidently — детекция дрифта данных (PSI, KL, Wasserstein) (docs.evidentlyai.com — 2026)

07 Feast — feature store для онлайн/офлайн-признаков (docs.feast.dev — 2026)

08 FastAPI — построение inference-API для модели (fastapi.tiangolo.com — 2026)

09 Федеральный закон 152-ФЗ «О персональных данных» — обработка и локализация (pravo.gov.ru — 2025)

10 ITfresh — внутренний шаблон MLOps-контура: реестр, CI/CD, мониторинг дрифт... (itfresh.ru — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.