RIA FORECAST

Вероятностная модель событий по гексам H3
на горизонтах 1, 3 и 7 дней

АВТОР
Виктор Веревкин

СТАТУС
Production v1.0

РЕПОЗИТОРИЙ
github.com/torviktor/ria-forecast

ПЕРИОД ДАННЫХ
1412 дней · до 20 апр 2026

— СЕКЦИЯ 01

01 ВВЕДЕНИЕ.

ЧТО ЭТО ЗА ПРОЕКТ

Аналитический инструмент: оценивает вероятность событий целевых категорий в географических ячейках на горизонтах 1, 3 и 7 дней.

01 **Источник данных**

Публичный JSON-эндпоинт ленты РИА Новости. Один источник, четыре года истории.

02 **Пространственная сетка**

H3 resolution 5 — шестиугольники ~252 км², диаметр около 9 км.

01 **Целевые категории**

air_defense, strike_ukraine, strike_russia. Остальные — контекст.

02 **Не предсказание намерений**

Статистические вероятности по зафиксированным паттернам прошлого. Не оперативный прогноз.

ТРОЙНОЕ ПОЗИЦИОНИРОВАНИЕ

01

Прикладной инструмент

Решает реальную задачу профессиональной аналитики автора. Не учебный пример.

02

Портфолио-кейс

Полный цикл на стыке бизнес-анализа, Data Analytics и ML. Код открыт.

03

Методологический эксперимент

Документация избыточна по сравнению с pet-проектом — каждое решение поясняется.

ДЕСКТОП, НЕ ВЕБ

От серверного варианта отказались сознательно — ради компетенций, которых не получить на привычных деплоях.

01 **PyInstaller и упаковка**

Self-contained exe, hidden imports, _internal bundles, frozen-режим.

02 **Проектирование GUI**

customtkinter: тёмная тема, прогресс-бары, не-блокирующий интерфейс.

03 **Распространяемый продукт**

Чужой Windows без Python, без venv, с антивирусом, без интернета.

04 **Полный цикл**

Не «обучил модель и опубликовал ноутбук», а провёл от сырых данных до zip-архива.

— СЕКЦИЯ 02

02

ДАННЫЕ И ПАРСИНГ.

ИСТОЧНИК В ЦИФРАХ

Один публичный JSON-эндпоинт. Четыре года непрерывных машиночитаемых данных, 404 активных ячейки H3-r5.

1412 дн

ПЕРИОД

9 июн 2022 — 20 апр 2026

404

АКТИВНЫХ ЯЧЕЕК

H3 resolution 5

8

КАТЕГОРИЙ СОБЫТИЙ

3 целевые, 5 контекстных

49

ПРИЗНАКОВ

5 групп: A, B, C, D, E

`air_defense`

Целевая · самая плотная категория

`strike_ukraine`

Целевая · средняя плотность

`strike_russia`

Целевая · разрежённая, 2% заполнения

ПАРСЕР

ria_parser_v2.py

Машиночитаемый источник вместо парсинга HTML. Возобновляемый, инкрементальный, не требует БД.

01

JSON-эндпоинт, не HTML

Тот же ответ, что получает фронтенд интерактивной карты. Машиночитаемая схема, стабильна годами.

02

Инкрементальная дозагрузка

Уже обработанные дни не перекачиваются. Ежедневное обновление — секунды вместо часов.

03

Локальный кэш геокодирования

Один топоним резолвится один раз. Воспроизводимость не зависит от внешних геобаз.

04

Plain JSON без БД

Файлы по дням. Легко диффить, бэкапить, точечно перепарсить — без миграций схемы.

ПРОСТРАНСТВЕНН ПЛОТНОСТЬ

Совокупное число событий трёх целевых категорий по гексам H3 за весь период наблюдений.

01 **Hotspots вдоль линии фронта**

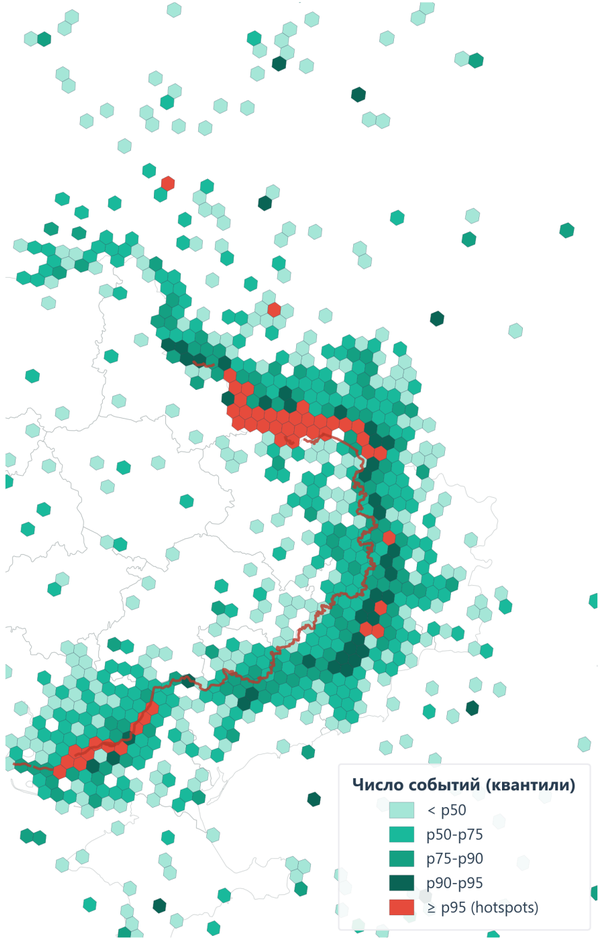
Подавляющее большинство событий сосредоточено в узкой прифронтовой полосе.

02 **Разреженный паттерн**

Большая часть территории — околонулевая плотность. Это влияет на выбор баланса категорий.

Пространственная плотность событий

Совокупное число событий трёх целевых категорий по гексам H3 за весь период наблюдений (июнь 2022 — апрель 2026)



Источник: data/processed/events_final.parquet

— СЕКЦИЯ 03

03

ПРОСТРАНСТВО.

СТРУКТУРА СОСЕДСТВА

У шестиугольника шесть равноудалённых соседей. Любой признак «среднее по соседям» симметричен — без направленного смещения.

01 Ring-1 — 6 соседей, ~9 км

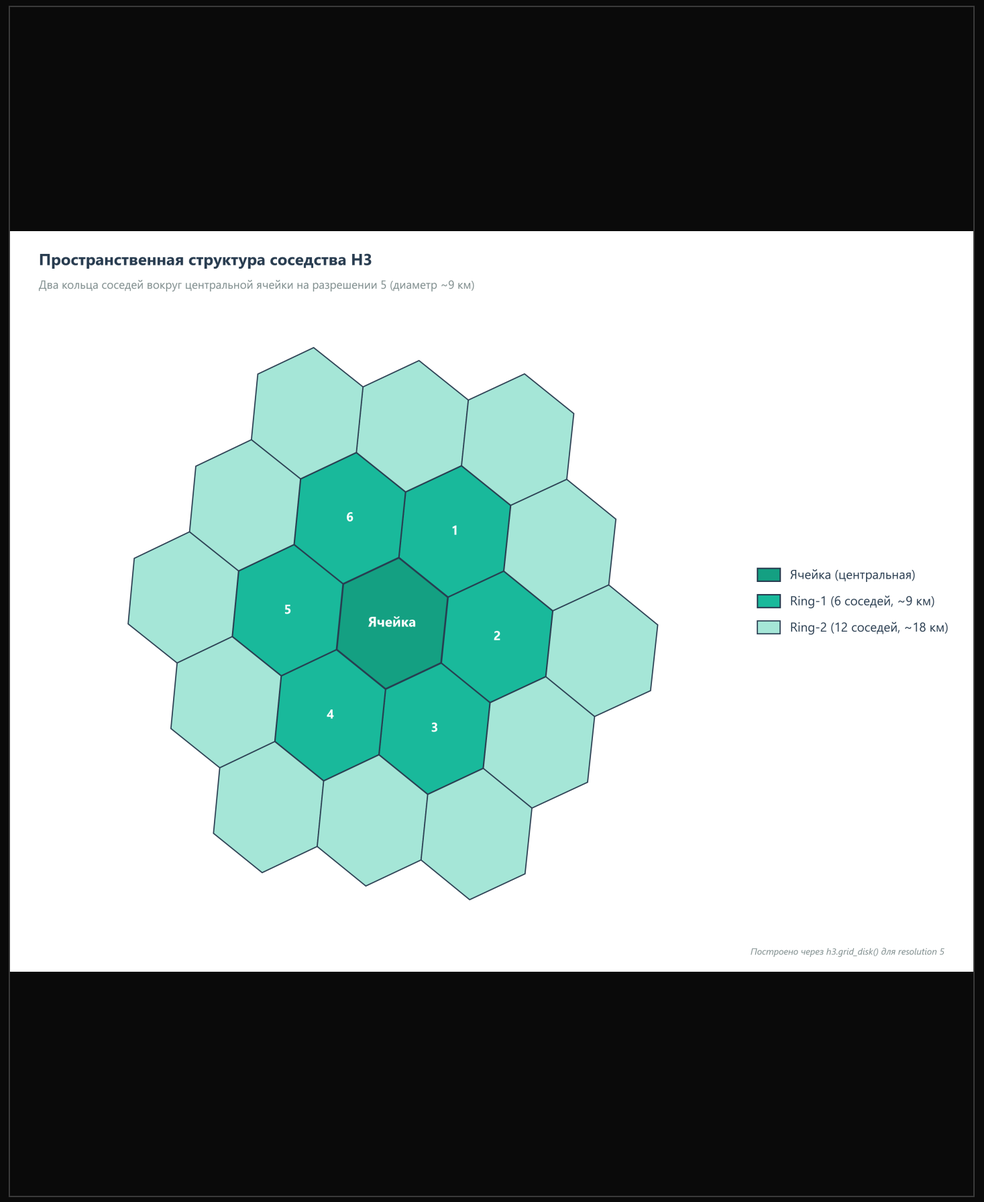
Локальная динамика: один сектор фронта, одна оперативная группа.

02 Ring-2 — 12 соседей, ~18 км

Операционная связность: давление распространяется на смежные сектора с лагом.

03 Ring-3 не используется

На радиусе ~27 км пространственная связность ослабевает до уровня шума.



РАЗРЕШЕНИЕ 5 — ПОЧЕМУ

Баланс двух противоположных давлений: слишком крупная ячейка размывает сигнал, слишком мелкая — обнуляет статистику.

r=4

Слишком крупно

Площадь в 7 раз больше. События размазаны, локальная детализация теряется. Сотни ячеек на всей территории.

r=5

Выбрано

252 км² · диаметр ~9 км · ~1000 ячеек на интересующей территории. Содержательное разнообразие при управляемом объёме.

r=6

Слишком мелко

Площадь в 7 раз меньше. Большинство ячеек пустые на средних горизонтах. Модель деградирует на разреженности.

— СЕКЦИЯ 04

04

ПРИЗНАКИ.

49 ПРИЗНАКОВ, 5 ГРУПП

От локальных календарных к глобальным режимным. Все строятся со сдвигом shift(1) — защита от утечки будущего.

A	**Календарные**	dow, month, day_of_month, is_weekend, days_since_war_start, sin/cos dow и month	9 шт.
B	**Авторегрессионные**	Лаги 1–28, скользящие sum/mean/std/max за 7/14/30/60/90, EWMA 3/7/14, флаги, метки времени	35 шт.
C	**Пространственные**	nbr1 / nbr2 sum и ewma по соседям ring-1 и ring-2 — добавлены в Phase 2.5	5 шт.
D	**Кросс-категориальные**	battle, troops, sabotage, infrastructure_ever, had_terrorist_attack_ever	9 шт.
E	**Глобальные режимные**	global_sum_7d / 14d, global_trend_7d, russia_share_14d	4 шт.

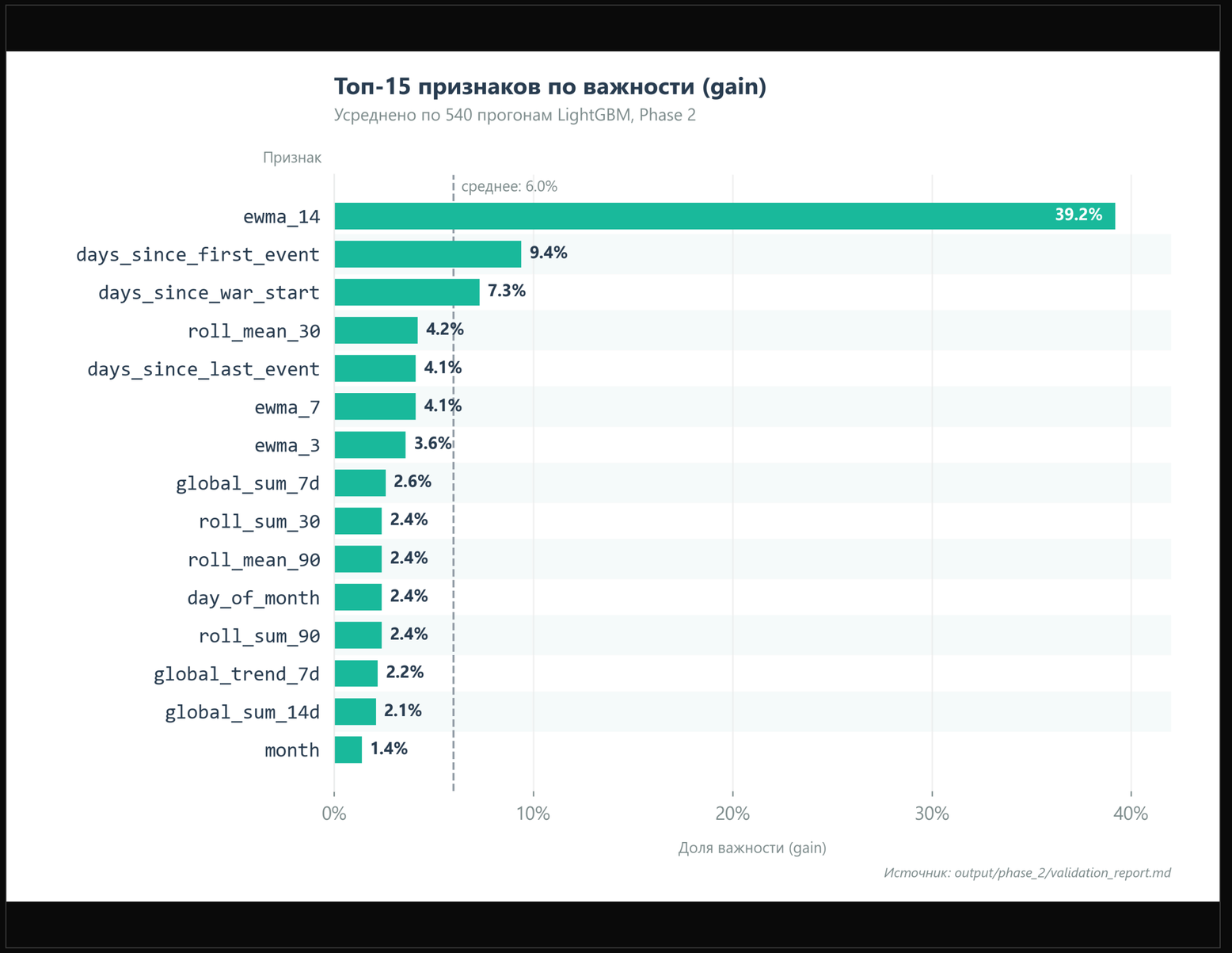
ЧТО ОКАЗАЛОСЬ ВАЖНЫМ

Топ-15 признаков по нормализованной важности (gain), усреднённой по 540 прогонам LightGBM.

ВЫВОД

ewma_14 даёт 39.2% важности — почти в 4 раза больше следующего признака.

Экспоненциально взвешенная свежая история ячейки — самый сильный одиночный предиктор. Это объясняет, почему RecencyWeightedRateBaseline конкурирует с LightGBM.



— СЕКЦИЯ 05

05

ВАЛИДАЦИЯ.

EXPANDING-WINDOW CV

Случайный K-fold для временных данных — утечка будущего. Здесь обучающее окно растёт вперёд, проверочное смещается, между ними gap.

01 **12 фолдов**

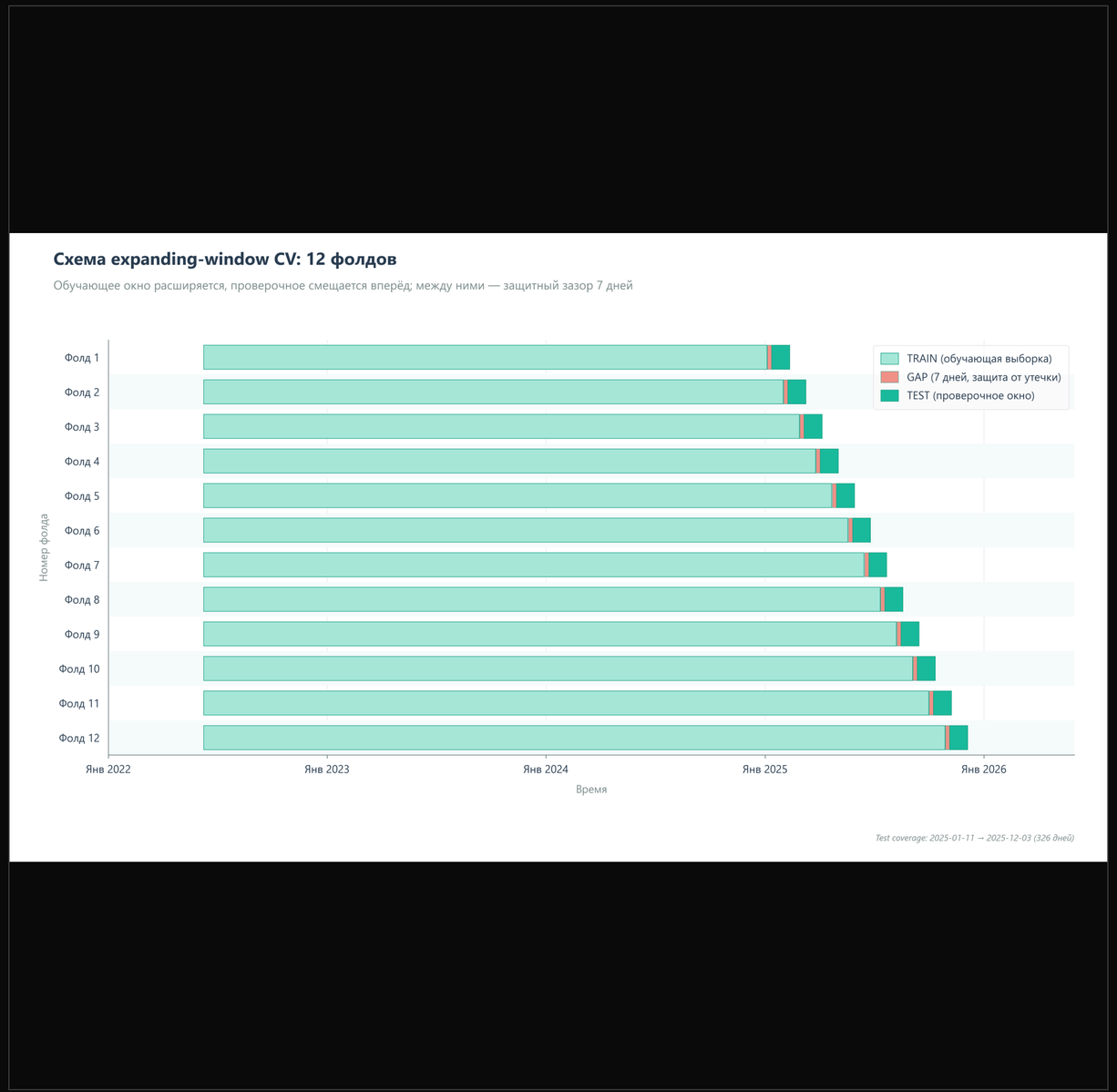
Покрытие тестами: 11 янв 2025 — 3 дек 2025, почти полный годовой цикл.

02 **Gap = 7 дней**

Защита от утечки y_binary_h7 — последняя обучающая цель не должна заглядывать в тест.

03 **Стабильность во времени**

12 раздельных проверок показывают разброс качества между фолдами.



ВЕРОЯТНОСТЬ ПО ГОРИЗОНТУ

Одна и та же частота даёт разные вероятности на разных горизонтах. Из пуассоновской модели редких событий.

ФОРМУЛА

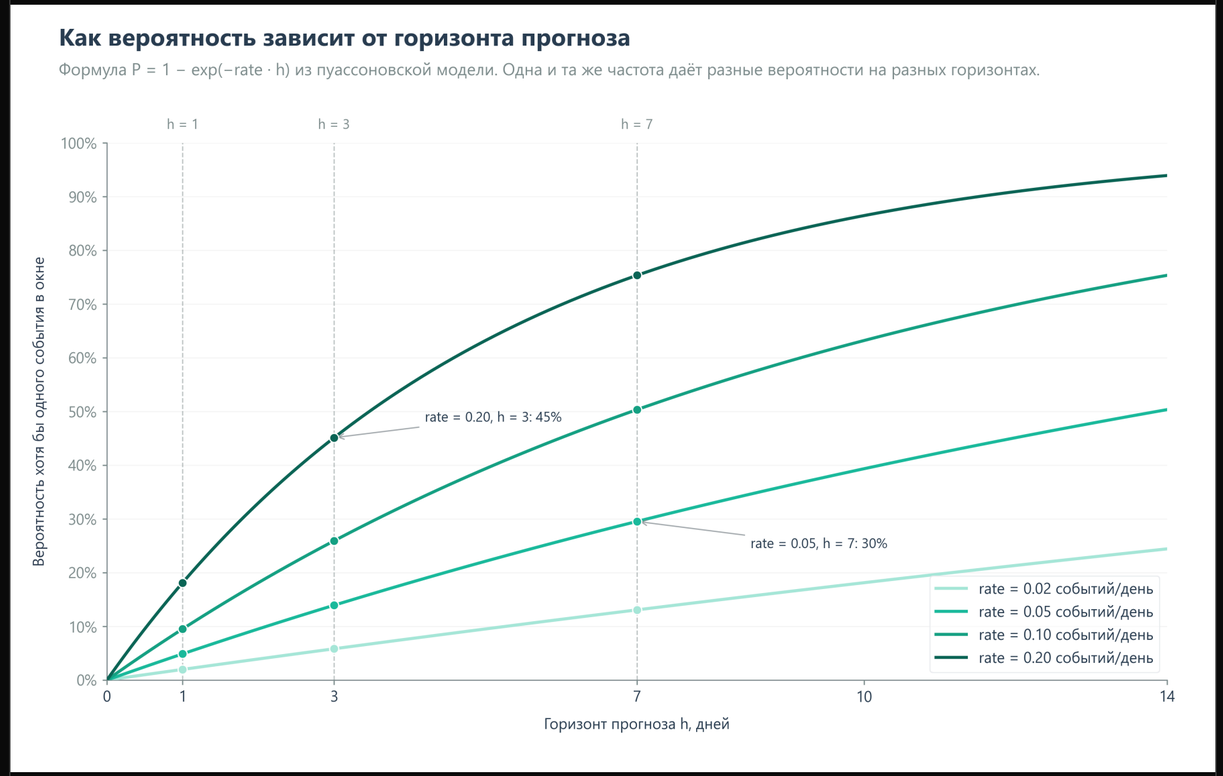
$$P = 1 - \exp(-\text{rate} \cdot h)$$

rate — экспоненциально взвешенная частота событий на день, h — горизонт в днях. Самокалиброванные вероятности без post-hoc регрессии.

h=1
Операционный
Что произойдёт завтра

h=3
Тактический
Устойчивая тенденция

h=7
Стратегический
Активность на неделю



— СЕКЦИЯ 06

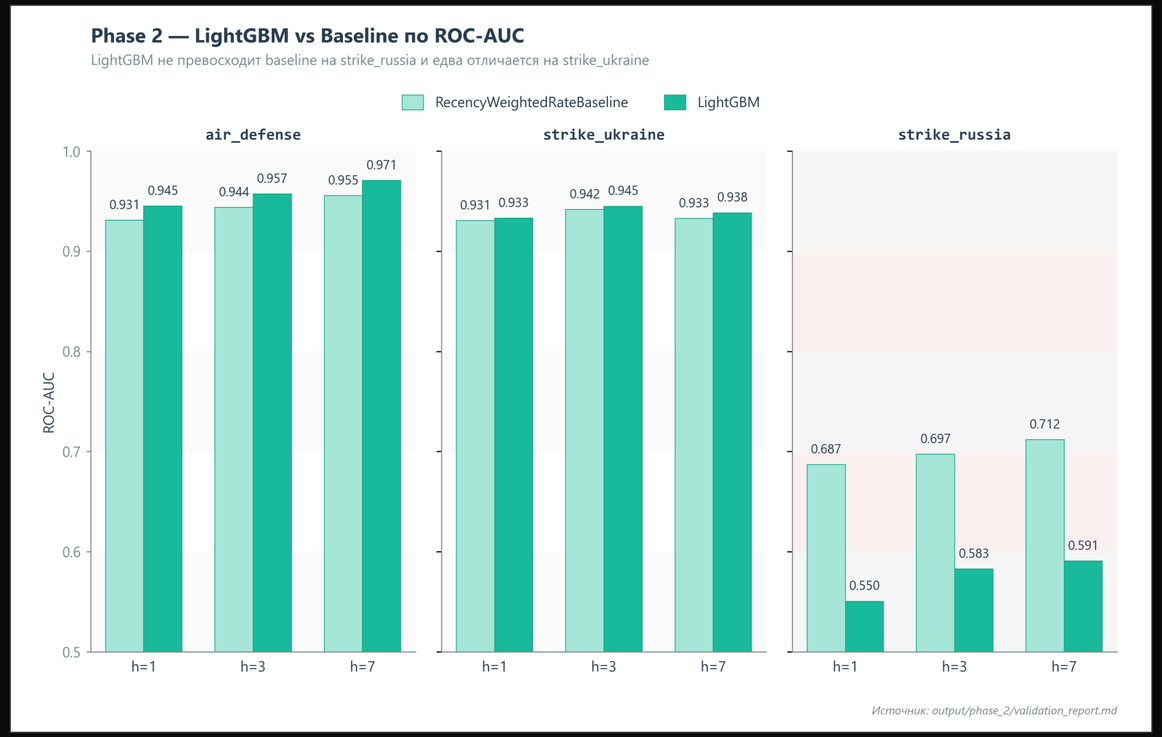
06

РЕЗУЛЬТАТЫ.

LIGHTGBM VS BASELINE

Сравнение по ROC-AUC на трёх категориях × трёх горизонтах. Заранее зафиксированный критерий Phase 3 — ΔAUC ≥ 10%.

КАТЕГОРИЯ	H	BASELINE	LIGHTGBM	Δ AUC
air_defense	1	0.931	0.945	+0.014
air_defense	3	0.944	0.957	+0.014
air_defense	7	0.955	0.971	+0.015
strike_ukraine	1	0.931	0.933	+0.002
strike_ukraine	3	0.942	0.945	+0.003
strike_ukraine	7	0.933	0.938	+0.006
strike_russia	1	0.687	0.550	-0.137
strike_russia	3	0.697	0.583	-0.114
strike_russia	7	0.712	0.591	-0.121



КАЛИБРОВКА ВЕРОЯТНОСТЕЙ

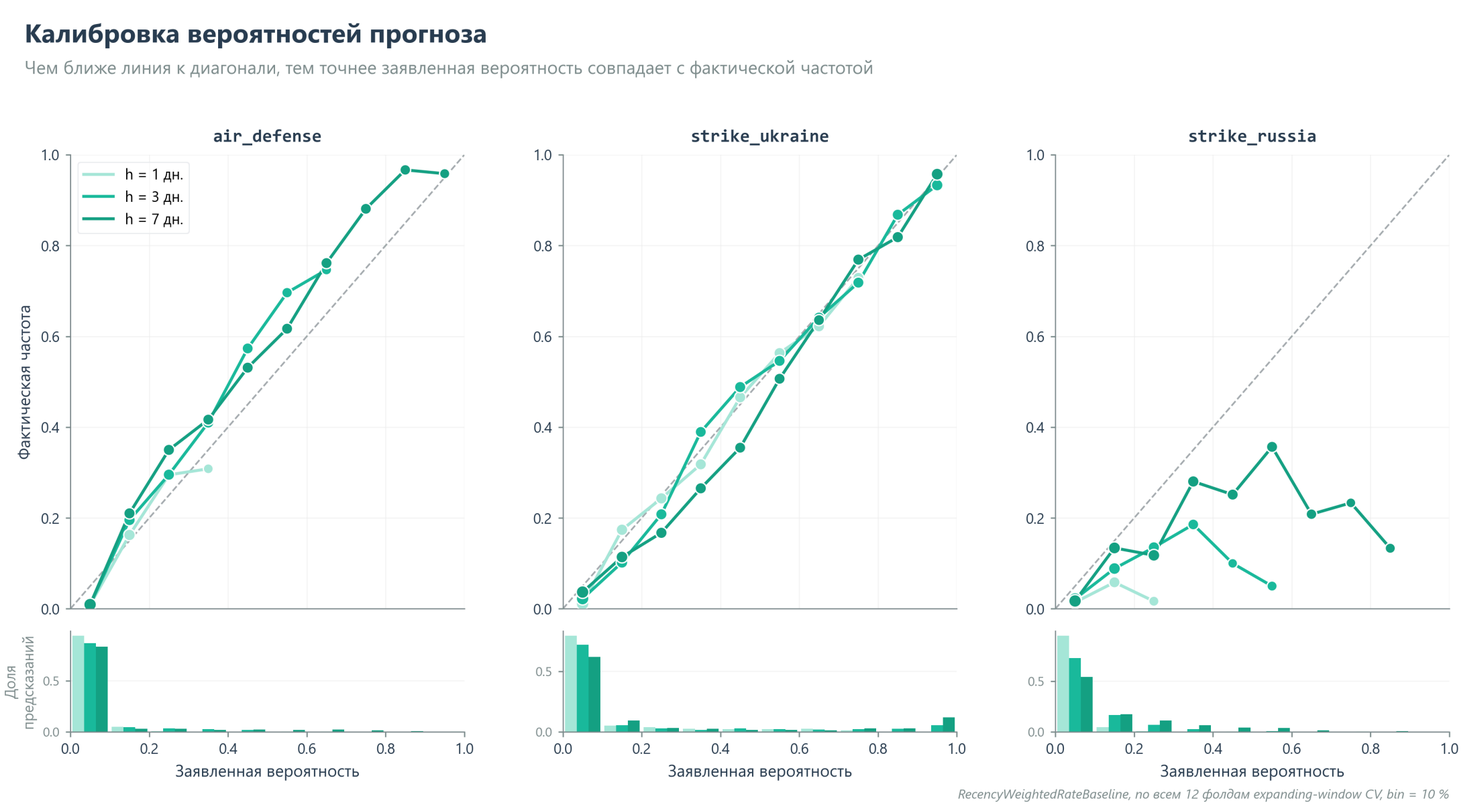
Чем ближе линия к диагонали, тем точнее заявленная вероятность совпадает с фактической частотой.

01 **air_defense, strike_ukraine**

Близко к диагонали на всех трёх горизонтах. Числа на карте означают то, что кажется.

02 **strike_russia — отклонение**

Разреженная категория, 2% заполнение панели. Один всплеск сдвигает rate сильно.

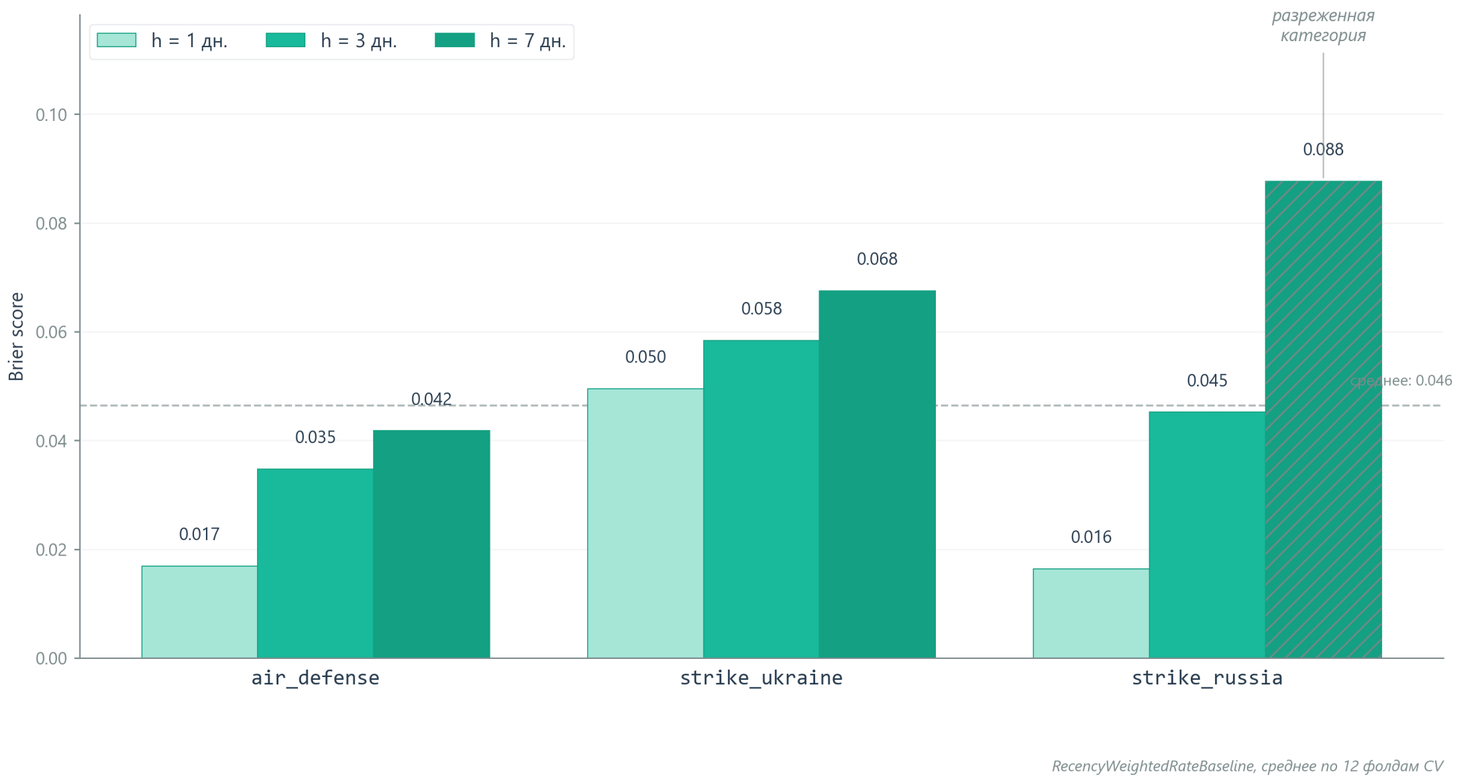


BRIER ПО 9 ПАРАМ

Brier score — средний квадрат разницы между предсказанной вероятностью и фактическим исходом. Ниже — точнее.

air_defense		strike_ukraine		strike_russia	
h=1	0.017	h=1	0.050	h=1	0.016
h=3	0.035	h=3	0.058	h=3	0.045
h=7	0.042	h=7	0.068	h=7	0.088

СРЕДНЕЕ ПО ВСЕМ 9 ПАРАМ — 0.046



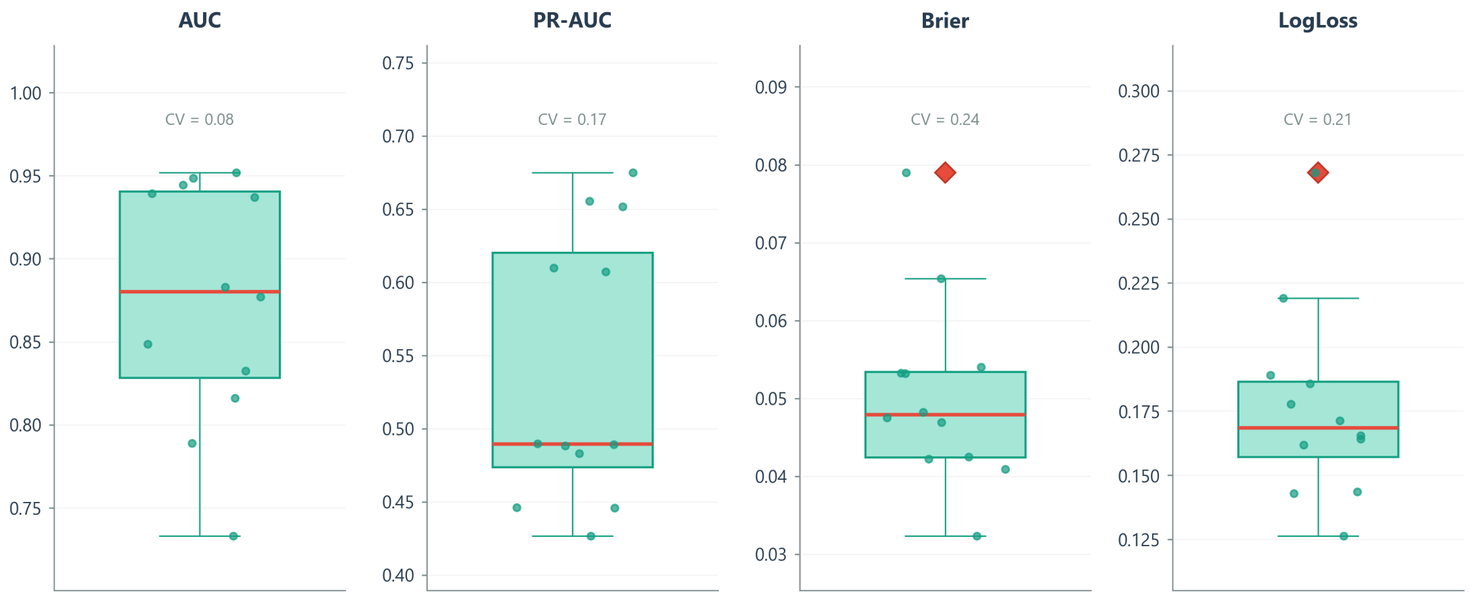
СТАБИЛЬНОСТЬ ПО 12 ФОЛДАМ

Boxplot с точечным отображением каждого фолда. CV — коэффициент вариации (std / mean).

AUC	CV = 0.08	наиболее стабильна
PR-AUC	CV = 0.17	
LogLoss	CV = 0.21	
Brier	CV = 0.24	выброс на strike_russia h=7

Стабильность метрик по 12 фолдам CV

Boxplot с точечным отображением каждого фолда; CV — коэффициент вариации (std/mean)



Максимальный разброс у Brier (CV = 0.24), минимальный — у AUC (CV = 0.08). Ранг стабильности: AUC > PR-AUC > LogLoss > Brier.

Источник: output/phase_2/lgbm_metrics.parquet

ВЕРДИКТ PHASE 2 / 2.5

Критерии запуска следующих фаз были зафиксированы заранее. Ни один не выполнен.

PHASE 2 · LIGHTGBM

Критерий ΔAUC ≥ 10% не достигнут

На air_defense LightGBM даёт +1.4–1.5% AUC. На strike_ukraine — в пределах шума. На strike_russia проигрывает baseline на 11–14 пп.

PHASE 2.5 · ABLATION

Критерий ΔPR-AUC ≥ 3 пп не достигнут

spatial_v2 + cross_cat_v2 + is_unbalance + isotonic — все 9 PR-AUC ухудшились или в шуме. Отрицательный результат задокументирован.

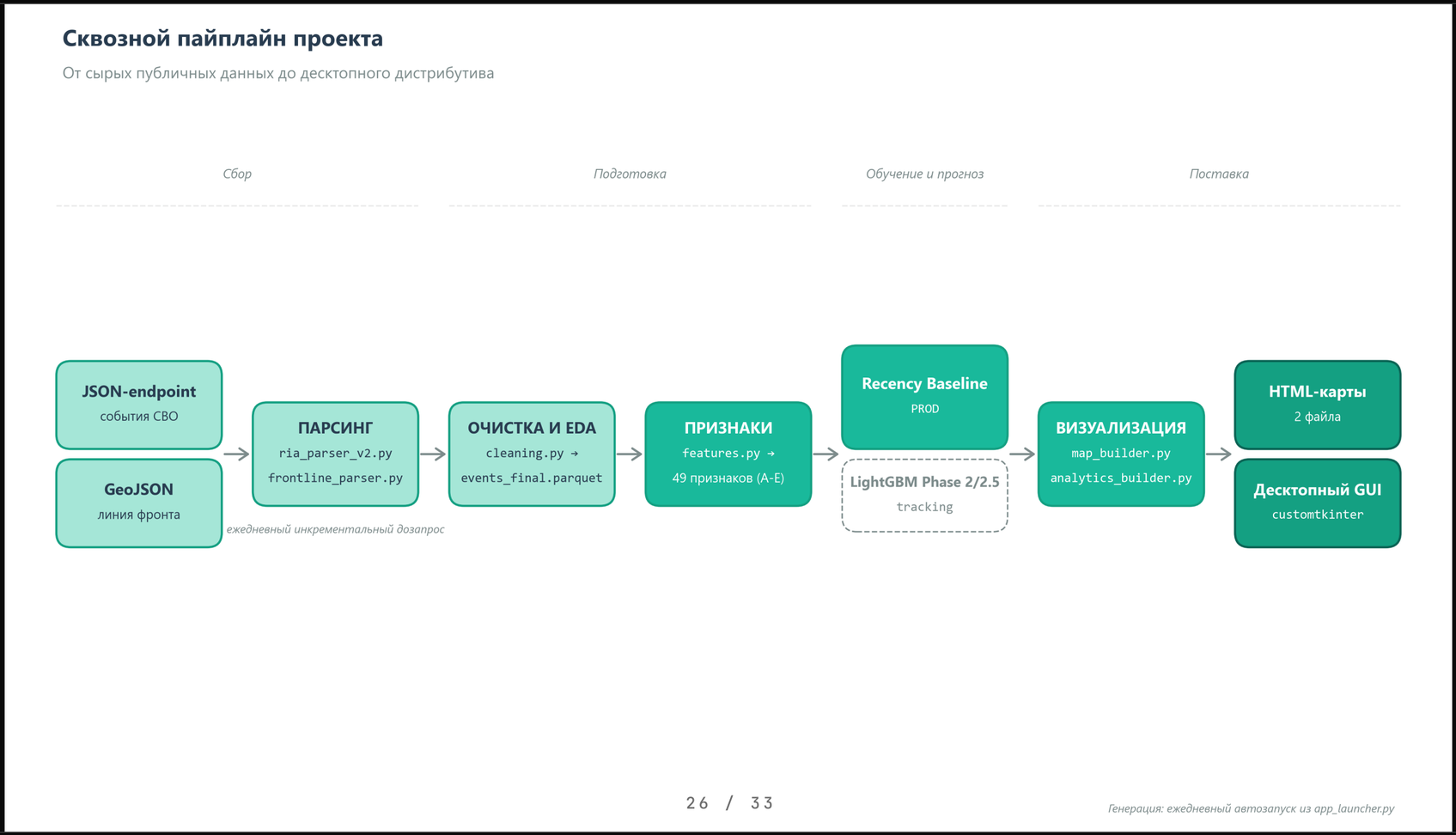
В ПРОДАКШЕН **RecencyWeightedRateBaseline · half-life 30**

— СЕКЦИЯ 07

07 АРХИТЕКТУРА.

СКВОЗНОЙ ПАЙПЛАЙН

От публичных данных до десктопного дистрибутива. LightGBM сохранён как tracking-референс — в проде работает baseline.



ДВУХУРОВНЕВАЯ МОДЕЛЬ

Одна и та же baseline на двух разных пространственных сетках — выбор продиктован разреженностью данных.

УРОВЕНЬ 1 · ГЕКСЫ H3

strike_ukraine + air_defense

~1000 ячеек, ресолюшн 5. Достаточно положительных примеров для устойчивых оценок rate в каждой ячейке.

УРОВЕНЬ 2 · РЕГИОНЫ

strike_russia

Категория слишком разрежена для гексов: 2% заполнение, медиана интервала 51 день. На уровне регионов сигнал плотнее.

СЛОЙ ОБЩЕЙ ОПАСНОСТИ

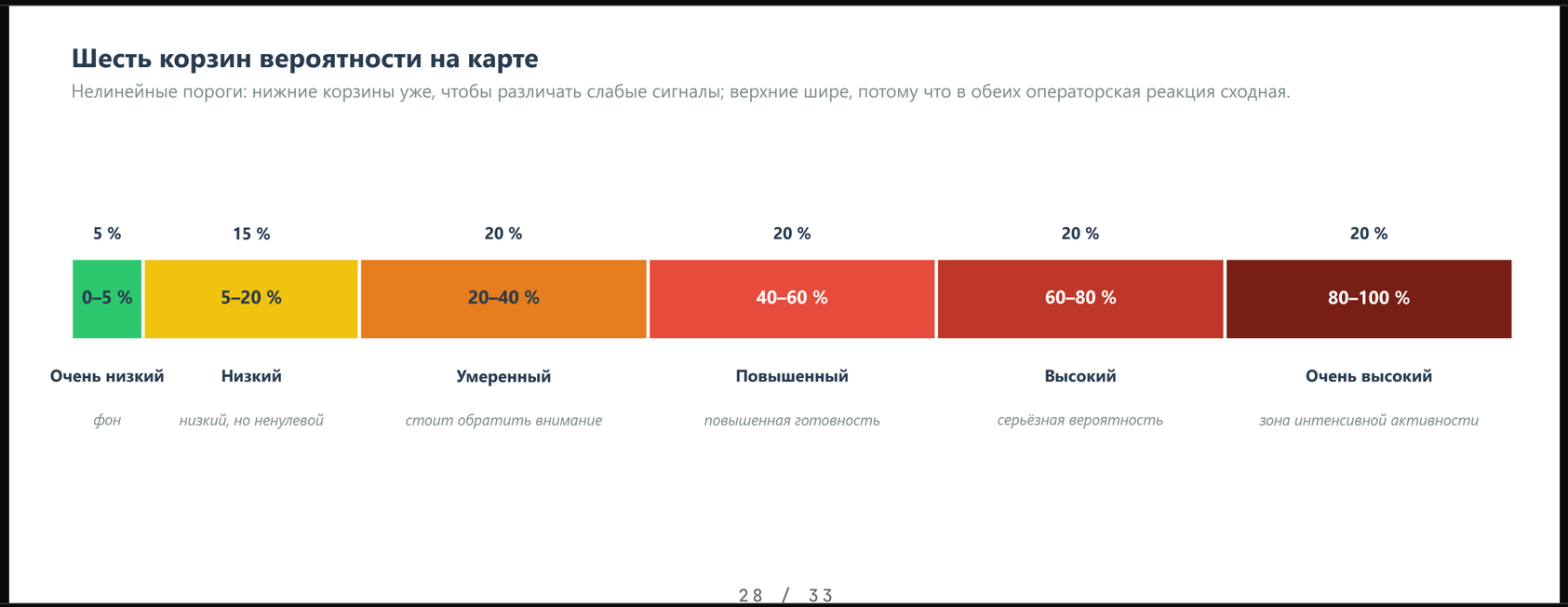
$$P(\text{общая}) = 1 - \prod (1 - p_i)$$

Допущение независимости категорий. Производная величина, не отдельный прогноз.

ШЕСТЬ КОРЗИН ВЕРОЯТНОСТИ

Нелинейные пороги. Нижние корзины уже — важно различать слабые сигналы между собой; верхние шире — поведение пользователя там одинаково.

0–5%	5–20%	20–40%	40–60%	60–80%	80–100%
ОЧЕНЬ НИЗКИЙ	**НИЗКИЙ**	**УМЕРЕННЫЙ**	**ПОВЫШЕННЫЙ**	**ВЫСОКИЙ**	**ОЧЕНЬ ВЫСОКИЙ**
Фон	Низкий, но ненулевой	Стоит обратить внимание	Повышенная готовность	Серьёзная вероятность	Зона интенсивной активности



— СЕКЦИЯ 08

08 ОГРАНИЧЕНИЯ.

ЧТО МОДЕЛЬ НЕ ДЕЛАЕТ

Явные границы — основа ответственного использования результата. Это не оперативный прогноз и не предсказание намерений.

01 **Не предсказывает конкретные инциденты**

Не «в такой-то точке в такой-то час с такими жертвами», а вероятностное распределение активности по территории.

02 **Не учитывает стратегических решений**

Опирается на зафиксированные паттерны прошлого. Лаг реакции на смену режима — от нескольких дней до пары недель.

03 **Только суточная гранулярность**

Часы пиковой активности, ночные и дневные окна модели не видны. Ограничение источника.

04 **Не делает прогнозов вне исторических зон**

Если в ячейке за всю историю не было событий — вероятность близка к нулю по построению.

05 **Не обновляется в реальном времени**

Пайплайн запускается при старте приложения. Между запусками прогноз не меняется.

КУДА РАСТИ

От готовых к реализации в следующем релизе — к пилотным экспериментам.

01 **Расстояние до линии фронта**

Самое дешёвое улучшение. Откладывалось из-за мультиколлинеарности — пересмотреть при изменении feature set.

02 **Веб-версия**

Серверное развёртывание той же аналитики. Большая часть кода переиспользуется без изменений.

03 **Другие источники**

Альтернативные OSINT-ленты для кросс-валидации событий. Слой согласования разметок.

01 **Декомпозиция strike_russia**

При накоплении данных — переход с регионов на гексы. Унификация пространственной модели.

02 **Ансамбль LightGBM + baseline**

Взвешенная комбинация. Потенциал — 2–3% метрик без архитектурных изменений.

03 **Новые целевые категории**

sabotage, terrorist_attack из контекстных в прогнозируемые — при достаточной плотности.

ДОКУМЕНТАЦИЯ ПОД КЛЮЧ

Четыре слоя docx-документации под четыре аудитории. README на двух языках. Метрики, диаграммы и черновики — в репозитории.

01 `docs/01_technical_description.docx`

Техническое описание

Полный пайплайн: данные, признаки, валидация, архитектура

02 `docs/02_probability_interpretation.docx`

Интерпретация вероятностей

Как читать карту, корзины, пороги принятия решений

03 `docs/03_executive_summary.docx`

Executive summary

Краткое представление для руководства, ~3 страницы

04 `docs/04_user_instruction.docx`

Инструкция пользователя

Десктоп-сборка: установка, запуск, GUI, обновление данных

`README.md`
для разработчиков · EN

`README.ru.md`
то же · RU

`docs/diagrams/`
все рисунки в репозитории

`docs/metrics_report.json`
машиночитаемые метрики

END-TO-END.
ОТ JSON ДО EXE

ПАРСИНГ

ria_parser_v2.py
frontline_parser.py

МОДЕЛЬ

RecencyWeightedRate
LightGBM tracking

ПОСТАВКА

PyInstaller onedir
customtkinter GUI