AgroNDVI

Спутниковый мониторинг рисовых полей и прогноз урожайности

Пет-проект для портфолио ML/DS

10 дней разработки · бесплатные открытые данные · полный pipeline

Валентюк Евгений · ML/DS-кандидат · переход из системного анализа

Контекст и цель

Куда хочу: ML/DS-команды банковского агро-сегмента

- **РСХБ-Цифровые решения** -- агро-скоринг через спутник, прогноз урожайности залогов;
- **Сбер для агробизнеса** (agroindex.ru) -- отдельная команда внутри Сбера;
- **Совкомбанк Страхование, Ингосстрах, СОГАЗ** -- агростраховой риск-менеджмент;
- **Агрохолдинги:** ЭФКО, Русагро, Содружество, Мираторг;
- **Геотех:** ExactFarming, OneSoil, Геомир.

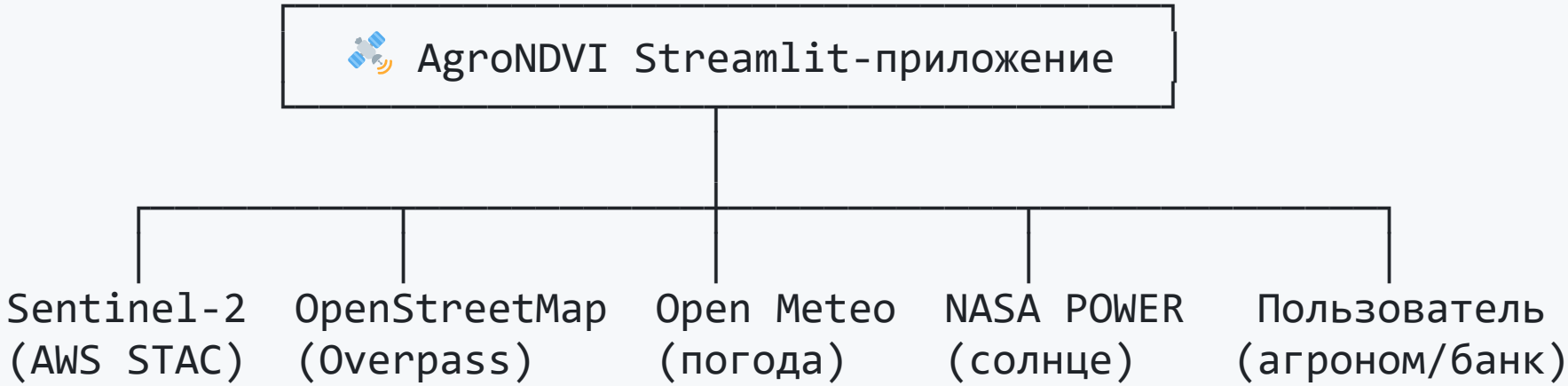
Что нужно показать: работу с Sentinel-2, geopandas/rasterio, time-series ML, агро-доменом, и **зрелость подхода**.

Что внутри (одно предложение)

> **Принимает координаты полей в Краснодарском крае, скачивает Sentinel-2 за сезон, считает NDVI, объединяет с погодой, обучает LightGBM прогноза урожайности, детектирует аномалии тремя независимыми методами, показывает всё в Streamlit-приложении с интерактивной картой.**

End-to-end pipeline за 10 минут. Воспроизводимо одной командой `run_pipeline.ps1`.

Архитектура (Context-уровень)



Все 4 внешних источника **бесплатные**, доступны **из РФ без VPN**.

Главный санкционный обход: Sentinel-2 берём через AWS Open Data, а не Copernicus (последний под блоком для РФ-IP).

Данные

Источник	Что даёт	Объём
Sentinel-2 L2A через AWS Element 84 STAC	45 безоблачных снимков сезона	10 м/px
OpenStreetMap (Overpass API)	Реальные границы 20 рисовых чеков	14 КБ GeoJSON
Open Meteo Historical	Температура, осадки, ET0, влажность	366 дней
NASA POWER	Солнечная радиация ALLSKY_SFC_SW_DWN	366 дней

Регион: Темрюкский район Краснодарского края (юго-запад tile MGRS 37TDK).

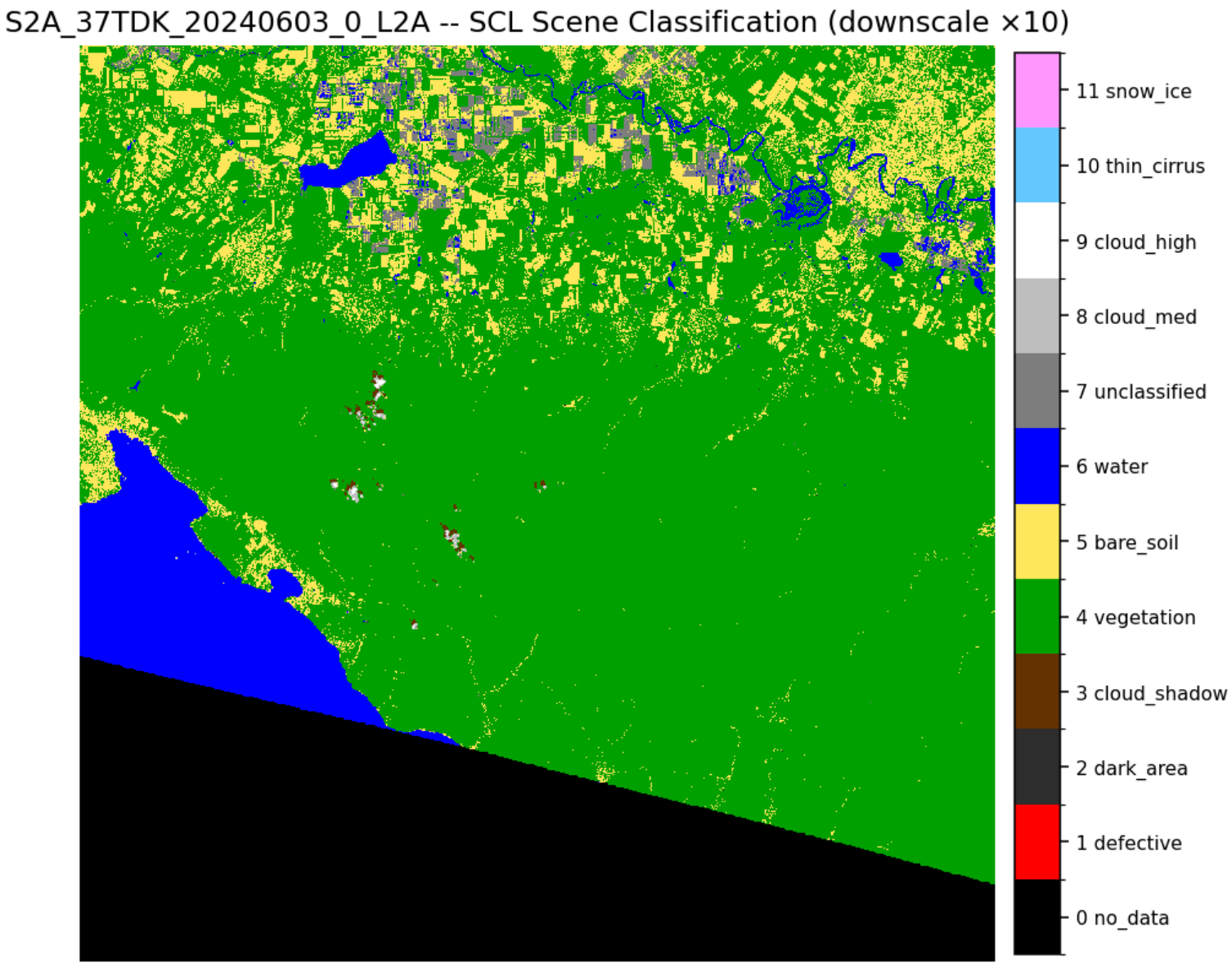
Период: октябрь 2023 -- сентябрь 2024 (полный сезон вегетации риса).

Спутниковый tile (3 июня 2024)

Классификация пикселей

Sentinel-2 L2A:

- 🟢 **62%** вегетация
- 🟡 9.5% голая земля
- 🔵 6% вода (море, лиманы, чеки)
- ⬛ 21% «no data» (UTM-крой)
- ☁️ <0.2% облака



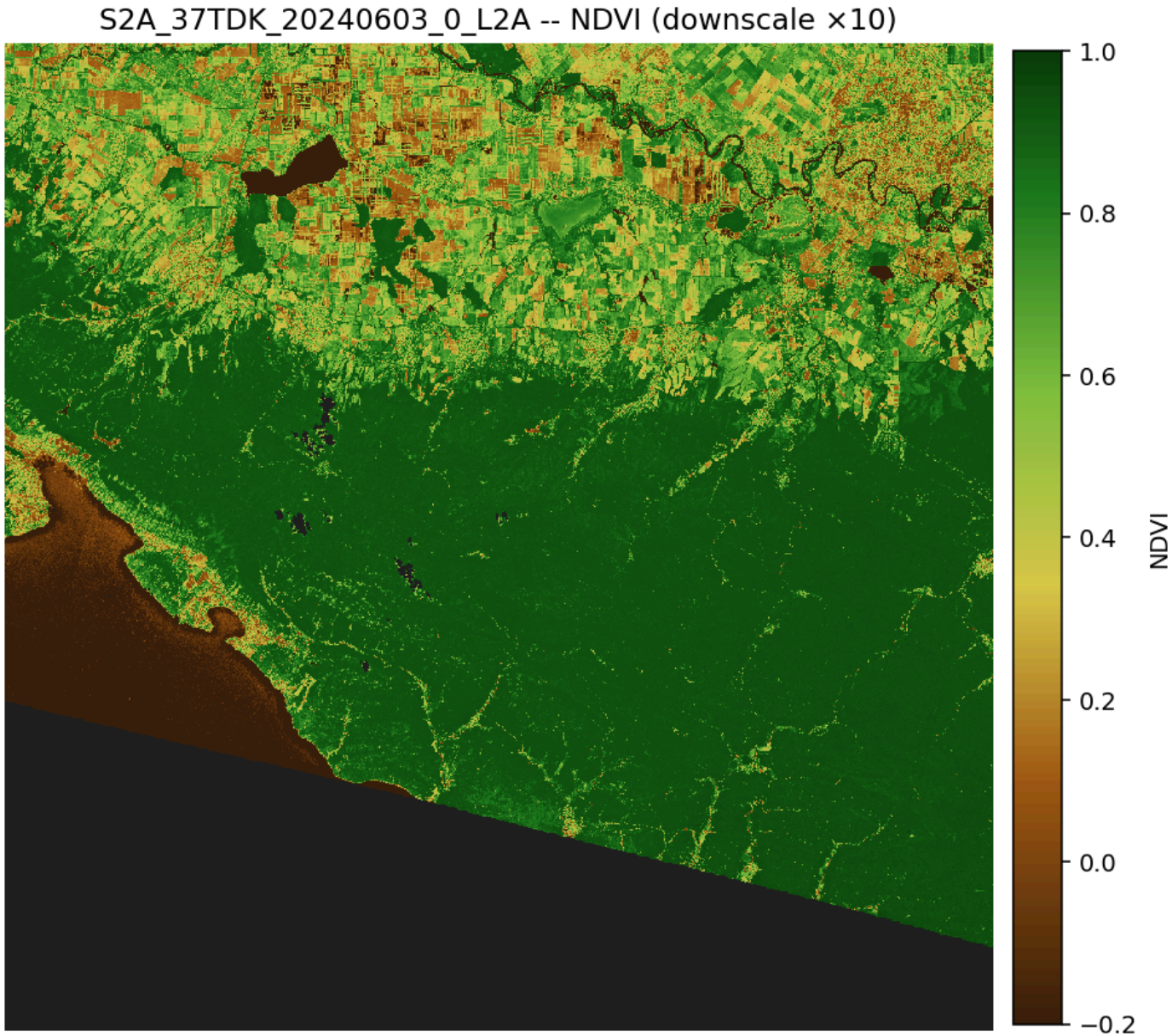
Валентюк Е.Г. · e.valentyuk@yandex.ru · github.com/EValentyuk

NDVI = (NIR - Red) / (NIR + Red)

Физика: хлорофилл поглощает красный (665 нм), клетки листа отражают NIR (842 нм). Чем активнее фотосинтез -- тем сильнее контраст.

Шкала:

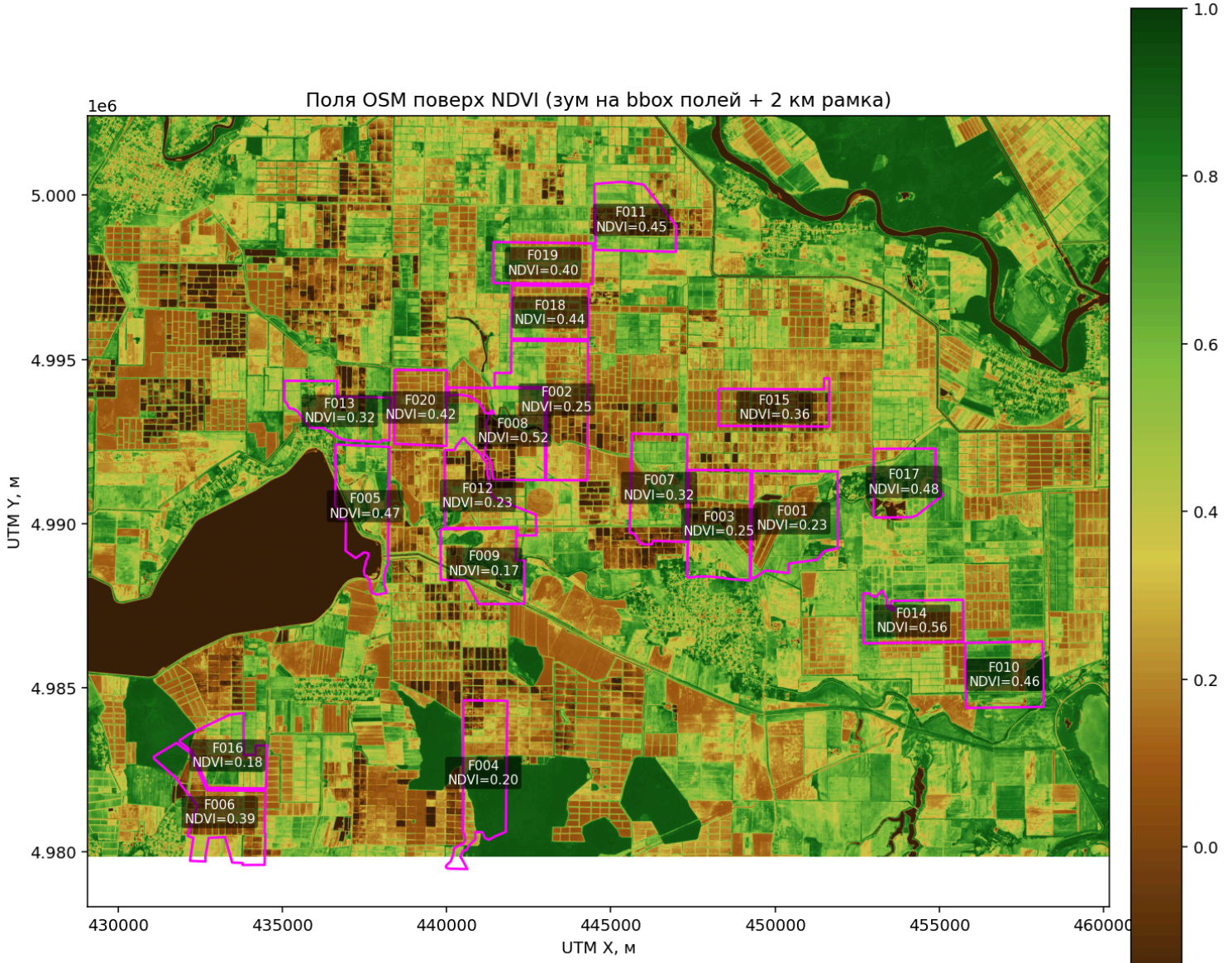
- 0.7-0.9 -- густой лес, пшеница в колошении
- 0.3-0.7 -- здоровая культура



20 рисовых чеков из OSM поверх NDVI

Сюрприз дня 3: OSM показал, что в выбранном районе **рис**, не пшеница (14 из 20 полей с тэгом `crop=rice`).

Темрюк исторически рисовый -- орошение от Кубани и лиманов. Чеки **по 700-1500 га**.

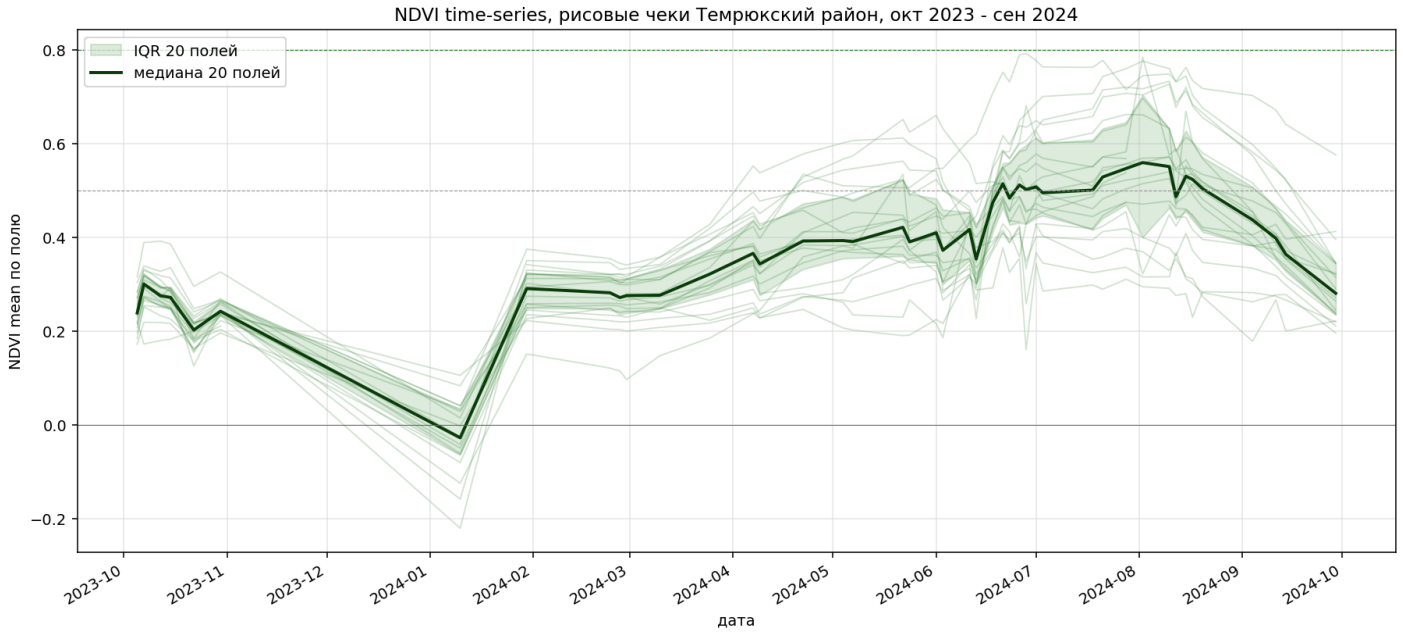


Валентюк Е.Е. · e.valentyuk@yandex.ru · github.com/EValentyuk

Главный артефакт: time-series NDVI

Классическая рисовая кривая за полный сезон 2023-2024:

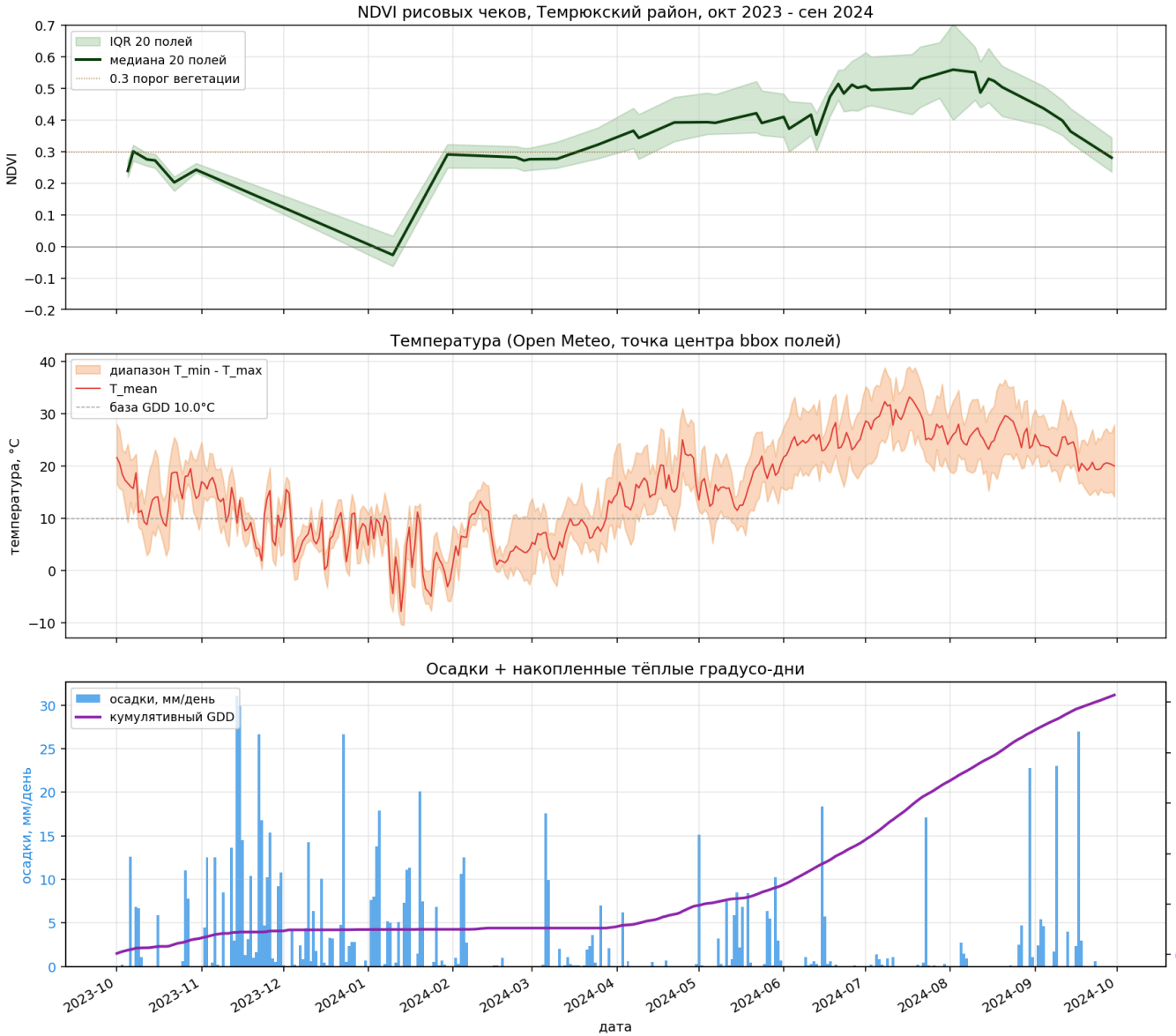
- октябрь: остатки растительности (0.3)
- январь: затопление чеков (~0)
- март-апрель: подготовка (0.15-0.25)
- май-июнь: рост (0.35-0.45)
- август: пик (0.56)



NDVI + погода = агрономический контекст

Три синхронизированных шкалы:

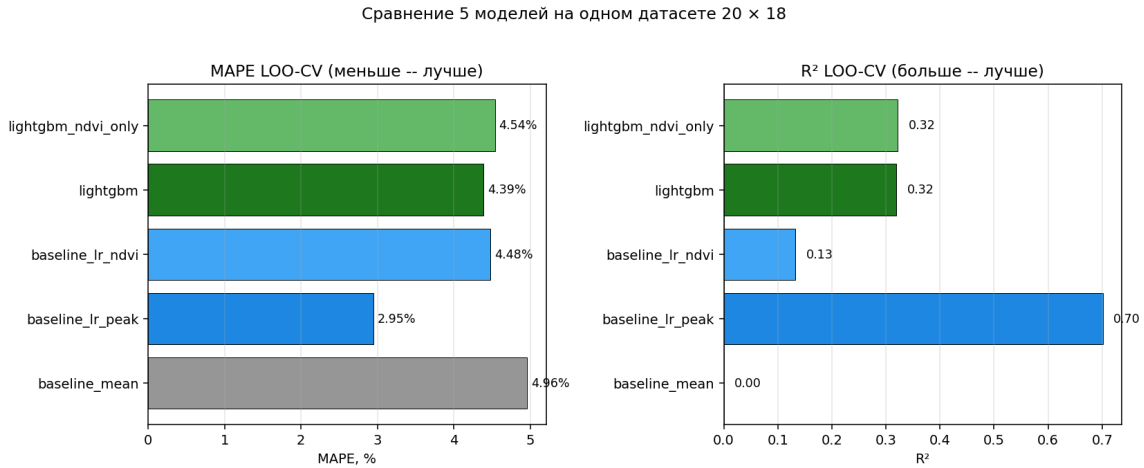
1. NDVI медиана 20 полей + IQR;
2. Температура min-max-mean;
3. Осадки + накопленный GDD.



Видно физически

ML pipeline: 5 моделей × Leave-One-Out CV

Модель	Фичи	MAPE %	R^2
baseline_mean	0	4.96	0.00
baseline_lr_peak	**1**	**2.95**	**0.70**
baseline_lr_ndvi	9	4.48	0.13
LightGBM	18	4.39	0.32
LightGBM (NDVI only)	9	4.54	0.32



💡 Главный методологический сигнал

> **На N=20 простая LinearRegression на одной фиче `ndvi_peak` побеждает LightGBM на 18 фичах.**
>
> MAPE: **2.95% vs 4.39%** · R^2: **0.70 vs 0.32**
>
> LightGBM train MAPE = 0.03%, LOO MAPE = 4.39% -- **классическое переобучение**.

Включил это в отчёт как есть. Это и есть зрелый ML: начинаешь с baseline, доказываешь оправданность сложной модели. А не «лишь бы LightGBM запустить».

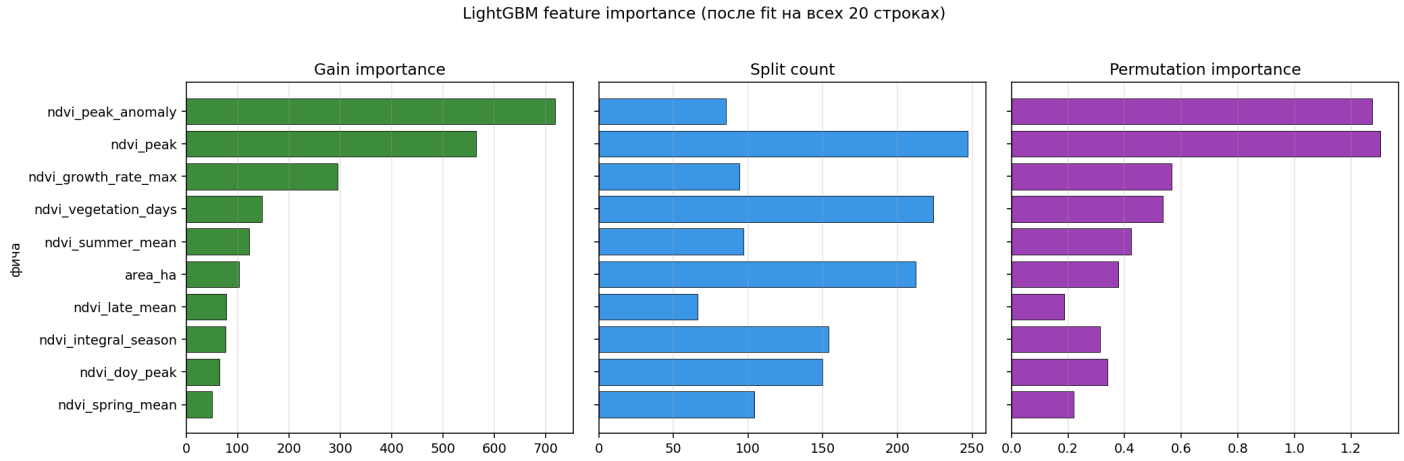
Это **методологический сигнал для рекрутёра**, что я понимаю bias-variance trade-off, а не делаю модели наугад.

Feature importance: правильное поведение модели

Top-5 фичей по gain:

1. ndvi_peak_anomaly (718)
2. ndvi_peak (565)
3. ndvi_growth_rate_max (295)
4. ndvi_vegetation_days (148)
5. ndvi_summer_mean (123)

Все 8 weather-фичей: gain = 0.

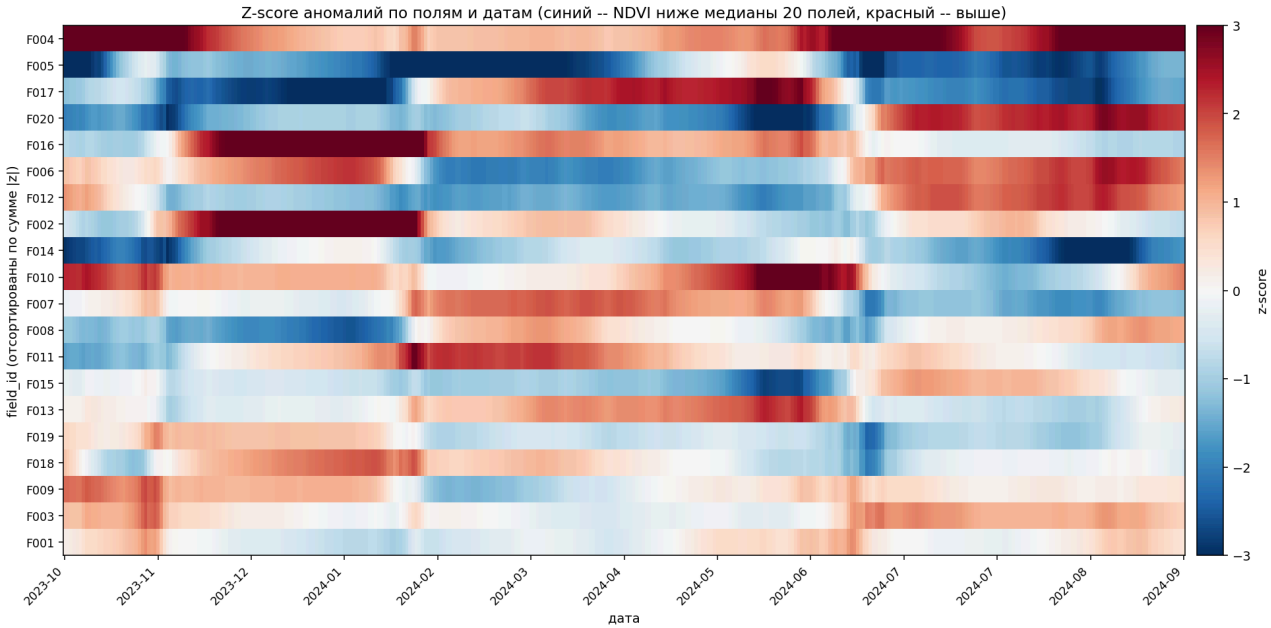


Anomaly detection: 3 метода независимо

Top-3 аномалии (одинаковы во всех методах):

Поле	mean_rank	Природа
F004	1.00	Чемпион (max NDVI 0.77)
F005	2.33	Отстающий (min NDVI 0.45)
F017	3.33	Сдвиг (пик на 2 недели раньше)

Spearman ρ между методами: 0.70 -- 0.96. **Устойчивый сигнал**, не



Чем полезен -- для агро-страхования

Сценарий 1: предупреждение об отставании

«Поле F005 имеет NDVI ниже коридора 20 соседей. Возможно: пересев, проблема с орошением, другая культура. Агроному выехать.»

Сценарий 2: верификация залога

«Банк выдал кредит под рисовый чек. Через 3 месяца спутник показывает NDVI 0.15 в фазу, когда должен быть 0.50. Залог не оправдывает оценку.»

Сценарий 3: страховое урегулирование

«Страхователь заявляет убыток. Спутник за 3 года показывает: 2024 NDVI в норме, отклонения нет. Страховой случай не подтверждён.»

UI: Streamlit + folium + plotly

5 KPI: полей, снимков, NDVI, прогноз yield, top-K аномалий.

Карта (слева): 20 полей раскрашены по предсказанному yield. Top-3 аномалии в красной рамке.

Панель деталей (справа): селектор поля → метрики + NDVI кривая в коридоре 20 полей + погода сезона + OSM-метаданные.

Запуск: `streamlit run src/streamlit_app.py` → http://localhost:8501

Стек технологий

Слой	Технологии
Язык	Python 3.13
Геоданные	rasterio 1.5 + GDAL 3.12 · geopandas 1.1 · shapely · pyproj · folium
ML	LightGBM 4.6 · scikit-learn 1.8 · pandas · numpy
Time-series	rolling smoothing · LOO-CV · z-score · IsolationForest
UI	Streamlit 1.57 · streamlit-folium · plotly 6.7
API клиенты	pystac-client · osm2geojson · requests + retry
Документация	Markdown + Mermaid (C4 диаграммы)
Инфраструктура	venv · PowerShell · run_pipeline.ps1

Всё **бесплатное**, всё **открытое**, доступно из РФ.

Что осознанно НЕ сделал

- **Не использовал нейросети** (Conv-LSTM, U-Net) -- на 20 точках они переобучатся ещё хуже LightGBM. NN -- следующий этап при multi-year датасете;
- **Не делал tuning гиперпараметров** -- любой grid-search на 20 точках переобучится на CV;
- **Не подключил Google Earth Engine** -- заблокирован для РФ. AWS STAC выбран как замена;
- **Не использовал платные сервисы** (Sentinel Hub, Planet, Climate FieldView) -- всё бесплатно;
- **Таргет урожайности синтетический.** Реальной пол-уровневой урожайности в открытом доступе нет. Зафиксировано в коде и в отчёте.

Зрелость -- это умение сказать «не сделал, потому что...», а не «не сделал и не

Что дальше -- planned extensions

1. **Multi-year датасет:** 2020-2024 = 5 лет × 20 полей = 100 строк. Появится реальная дисперсия погоды между годами → LightGBM получит шанс показать преимущество над линейной моделью;
2. **Multi-region:** добавить Калининский / Славянский районы (пшеница). Микс культур → бустинг сможет различать;
3. **Реальный таргет:** запросить пол-уровневые урожайности у хозяйств или через open data partnership;
4. **CNN/Conv-LSTM** на сырой NDVI-кривой длиной 45 точек -- вместо плоских фичей;
5. **U-Net** для автодетекции границ полей по снимкам (вместо OSM);
6. **Backtest на отложенном годе:** train 2020-2023, test 2024 -- индустриальная оценка.

Что в портфолио из этих 10 дней

✅ Полный pipeline «спутник → прогноз → UI», воспроизводимо одной командой
✅ Работа с реальными OSM-полями (не выдуманными)
✅ Time-series ML (LightGBM + 4 baseline + LOO-CV)
✅ Anomaly detection в 3 метода с проверкой согласованности
✅ Streamlit-приложение, готовое демо
✅ 7 документов: brief, architecture (C4 Mermaid), metrics, portfolio-report, 2 эксперимент-отчёта, sentinel-download
✅ Параллельный «человеческий» лог для нетехнических читателей
✅ **Bias-variance trade-off** как методологический сигнал

Связь с первым проектом MiniProctor

MiniProctor (CV-прокторинг на MediaPipe + YOLOv8):
github.com/EValentyuk/MiniProctor

Переиспользовано:

- Структура документации (brief + experiment reports + portfolio report);
- Streamlit + folium как UI-комбо;
- Логика честного отчёта (включая «что не работает»);
- Шаблон C4-диаграмм через Mermaid.

Что AgroNDVI добавляет:

- **Новый домен:** агро/спутник вместо CV-прокторинга;
- **Новый стек:** rasterio + geopandas + LightGBM (геоданные + табличный ML);

Контакты и ссылки

- 📦 **Репозиторий:** github.com/EValentyuk/AgroNDVI
- 📄 **Отчёт для работодателя:** docs/portfolio-report.md
- 📊 **Метрики:** docs/metrics.md
- 🏗️ **Архитектура (C4):** docs/architecture.md
- 🧪 **Эксперимент LightGBM:** docs/experiments/2026-05-26-lgb-baseline.md
- 🔍 **Эксперимент аномалий:** docs/experiments/2026-05-26-anomaly.md
- ✉️ **Email:** e.valentyuk@yandex.ru
- 🐱 **GitHub:** github.com/EValentyuk

Открыт к собеседованиям в ML/DS-команды банковского агро-сегмента, агрохолдингов, страховых, геотех-стартапов.

Спасибо

🛰️ **AgroNDVI** -- доказательство того, что я умею:

- собирать данные из 4 разных источников;
- работать с реальными ограничениями (санкционный блок, маленький N);
- делать честный ML с baseline-сравнением;
- документировать осмысленно (C4, эксперимент-отчёты, portfolio brief);
- доводить пет-проект до законченного UI за 10 дней.

Вопросы?

`e.valentyuk@yandex.ru` · `github.com/EValentyuk`