

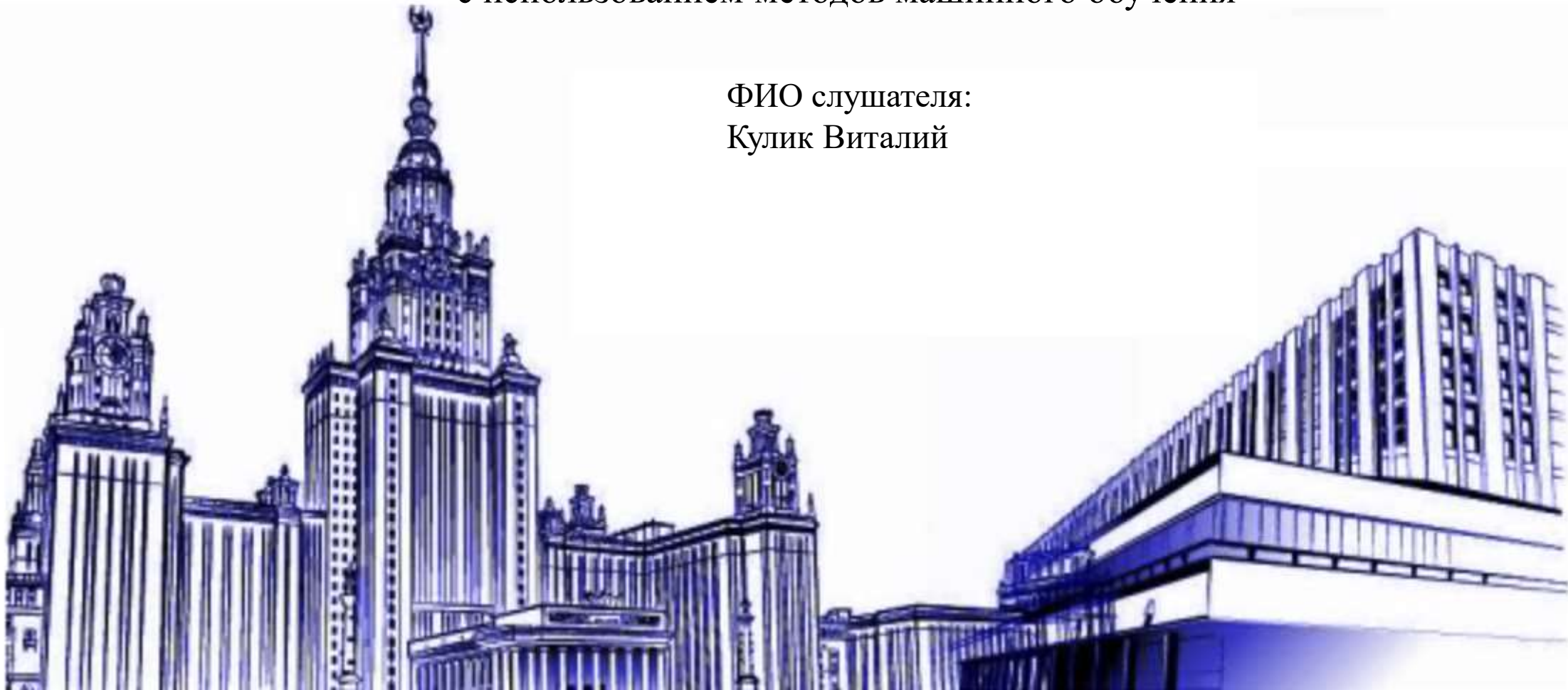
Факультет вычислительной математики и кибернетики МГУ имени М.В. Ломоносова



Программа профессиональной переподготовки
«Машинное обучение и управление большими данными»

Тема: Разработка системы мониторинга и прогнозирования цен на товары
с использованием методов машинного обучения

ФИО слушателя:
Кулик Виталий





Актуальность темы



Актуальность темы:

- Высокая волатильность цен на товары в России
- Прямое влияние на уровень жизни населения и инфляцию
- Необходимость точного прогнозирования для граждан, бизнеса и государства

Исследования ранее:

- Россия: Росстат, ИНП РАН (ARIMA, Prophet)
- За рубежом: Walmart, Target, FAO, World Bank (XGBoost, LSTM)

Проблемы существующих решений:

- Слабый учёт сезонности
- Отсутствие автоматизированных систем
- Низкая доступность для конечного пользователя



Цели и задачи



Цель работы:

- Разработать систему мониторинга и прогнозирования цен на товары с использованием методов машинного обучения и технологий Big Data.

Задачи:

- Организовать автоматический сбор данных из открытых источников
- Выполнить предобработку данных и Feature Engineering
- Обучить и оценить модель на основе XGBoost
- Разработать удобный веб-дашборд
- Обеспечить автоматизацию обновления данных



Методы и инструменты



- Сбор данных: Python, requests, Росстат
- Хранение: SQLite
- Анализ и обработка: pandas, seaborn, matplotlib
- Feature Engineering: 22 признака (лаги, rolling, сезонные)
- Модель: XGBoost (градиентный бустинг)
- Интерфейс: Streamlit + Plotly
- Автоматизация: скрипт `update_data.py`

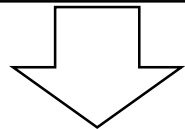


Архитектура системы



1. Сбор данных:

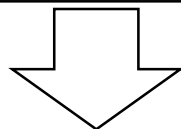
- Автоматический сбор еженедельных цен с сайта Росстата
- Генерация тестовых данных при отсутствии свежих
- Источники: rosstat.gov.ru + синтетические данные



Выход: Сырые данные в формате Excel / DataFrame

2. Предварительная обработка данных

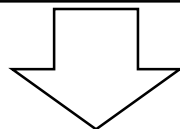
- Загрузка данных в SQLite базу (food_prices.db)
- Очистка и проверка данных
- Добавление временных колонок (year, month, week и т.д.)



Выход: Структурированные данные в базе

3. Создание признаков:

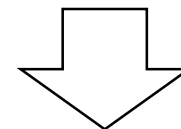
- Создание 22 признаков
- Лаговые признаки (price_lag_1, _2, _4, _8)
- Скользящие статистики (mean и std за 4, 8, 12 недель)
- Сезонные признаки (is_winter, is_summer)
- Цена год назад (price_year_ago)



Выход: features.csv с расширенным датасетом

4. Обучение и оценка модели:

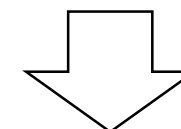
- Обучение отдельных моделей XGBoost для каждой пары «товар-регион»
- Разделение на train/test (80/20) с учётом времени
- Оценка по метрикам: MAE, RMSE, MAPE, R^2



Выход: 10 обученных моделей (.joblib)

5. Прогнозирование:

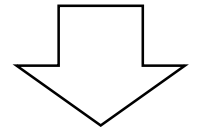
- Загрузка последней модели
- Генерация прогноза на 30/60/90 дней
- Добавление контролируемого шума для реалистичности
- Сохранение прогноза



Выход: Таблица с прогнозными значениями

6. Реализация интерфейса:

- Веб-дашборд на Streamlit
- Вкладки: «Аналитика по товару», «Общая аналитика», «Сравнение товаров»
- Интерактивные графики (Plotly)
- Кнопка «Построить прогноз»



Выход: Готовый веб-сервис



Сбор данных, обработка и создание признаков



1. Сбор данных

- Автоматический парсинг еженедельных цен с сайта Росстата
- Создание тестового датасета (1740 записей)
- Источник: rosstat.gov.ru

2. Предварительная обработка

- Загрузка в SQLite (food_prices.db)
- Проверка и очистка данных
- Добавление временных колонок (год, месяц, неделя)

3. Feature Engineering

- Создано **22 признака**
- Лаговые: price_lag_1, price_lag_2, price_lag_4, price_lag_8
- Скользящие: price_roll_mean_4/8/12, price_roll_std_4/8/12
- Сезонные: is_winter, is_summer, price_year_ago

Наименование	на 6 апреля	на 13 апреля	на 20 апреля
Говядина (кроме бескостного мяса), кг	729,59	731,62	732,77
Свинина (кроме бескостного мяса), кг	394,11	395,17	395,92
Баранина (кроме бескостного мяса), кг	810,41	812,71	813,20
Куры охлажденные и мороженые, кг	229,25	228,74	228,26
Сосиски, сардельки, кг	550,56	549,65	550,32
Колбаса полукопченая и варено-	730,04	729,19	727,72
Колбаса вареная, кг	580,95	581,53	579,34
Консервы мясные для детского питания,	1305,44	1303,25	1312,69
Рыба мороженая неразделанная, кг	348,78	349,60	350,91
Масло сливочное, кг	1142,35	1139,98	1138,54
Масло подсолнечное, л	152,56	152,84	152,92
Маргарин, кг	295,65	296,22	297,33
Молоко питьевое цельное пастеризованное 2,5-3,2% жирности, л	98,13	98,07	97,84

date	product	region	price	price_lag_1	price_lag_2	price_roll_mean_4	price_roll_std_4	is_winter	month
2023-01-08	Говядина (кроме бескостного мяса)	Российская Федерация	451.48	450.84	NaN	453.7288	3.289975	1	1
2023-01-15	Говядина (кроме бескостного мяса)	Российская Федерация	450.87	451.48	450.84	452.7788	2.844275	1	1
2023-01-22	Говядина (кроме бескостного мяса)	Российская Федерация	453.47	450.87	451.48	452.9458	2.348567	1	1
2023-01-29	Говядина (кроме бескостного мяса)	Российская Федерация	449.78	453.47	450.87	451.3888	1.547066	1	1
2023-02-05	Говядина (кроме бескостного мяса)	Российская Федерация	451.75	449.78	453.47	451.4675	1.559324	1	2
2023-02-12	Говядина (кроме бескостного мяса)	Российская Федерация	459.86	451.75	449.78	453.5158	3.902263	1	2
2023-02-19	Говядина (кроме бескостного мяса)	Российская Федерация	454.19	459.86	451.75	453.6958	4.005767	1	2
2023-02-26	Говядина (кроме бескостного мяса)	Российская Федерация	453.64	454.19	459.86	454.6668	3.113015	1	2



Обучение и оценка модели



Подготовка данных

1. Загрузка и первичная обработка

- Загрузка данных из features.csv (1730 записей после очистки)
- Фильтрация по конкретному товару и региону
- Разделение на features (X) и target (y = price)
- Сохранение временной последовательности (time-series split)

2. Разделение на обучающую и тестовую выборки

- Используется хронологическое разделение (без случайного shuffle)
- Train: 80% данных (ранние периоды)
- Test: 20% данных (последние периоды)

3. Создание новых признаков (Feature Engineering)

- 22 признака, включая:
 - Лаговые признаки (price_lag_1 ... price_lag_8)
 - Скользящие статистики (price_roll_mean_4/8/12, price_roll_std_4/8/12)
 - Сезонные и временные признаки
 - price_year_ago

4. Нормализация / Масштабирование — не требовалась (XGBoost устойчив к масштабу)

Параметр	Значение	Описание
n_estimators	200	Количество деревьев в ансамбле
learning_rate	0.05	Скорость обучения (малое значение = более точное, но медленное обучение)
max_depth	6	Максимальная глубина каждого дерева
subsample	0.8	Доля выборки для обучения каждого дерева
colsample_bytree	0.8	Доля признаков для построения каждого дерева
random_state	42	Для воспроизводимости результатов

Метрика	Значение	Интерпретация
MAE (Mean Absolute Error)	2.35 руб	Средняя абсолютная ошибка
RMSE (Root Mean Squared Error)	3.08 руб	Чувствительна к большим ошибкам
MAPE (Mean Absolute Percentage Error)	3.46 %	Относительная ошибка
R ² (Coefficient of Determination)	0.95	Доля объяснённой дисперсии



Обучение и оценка модели



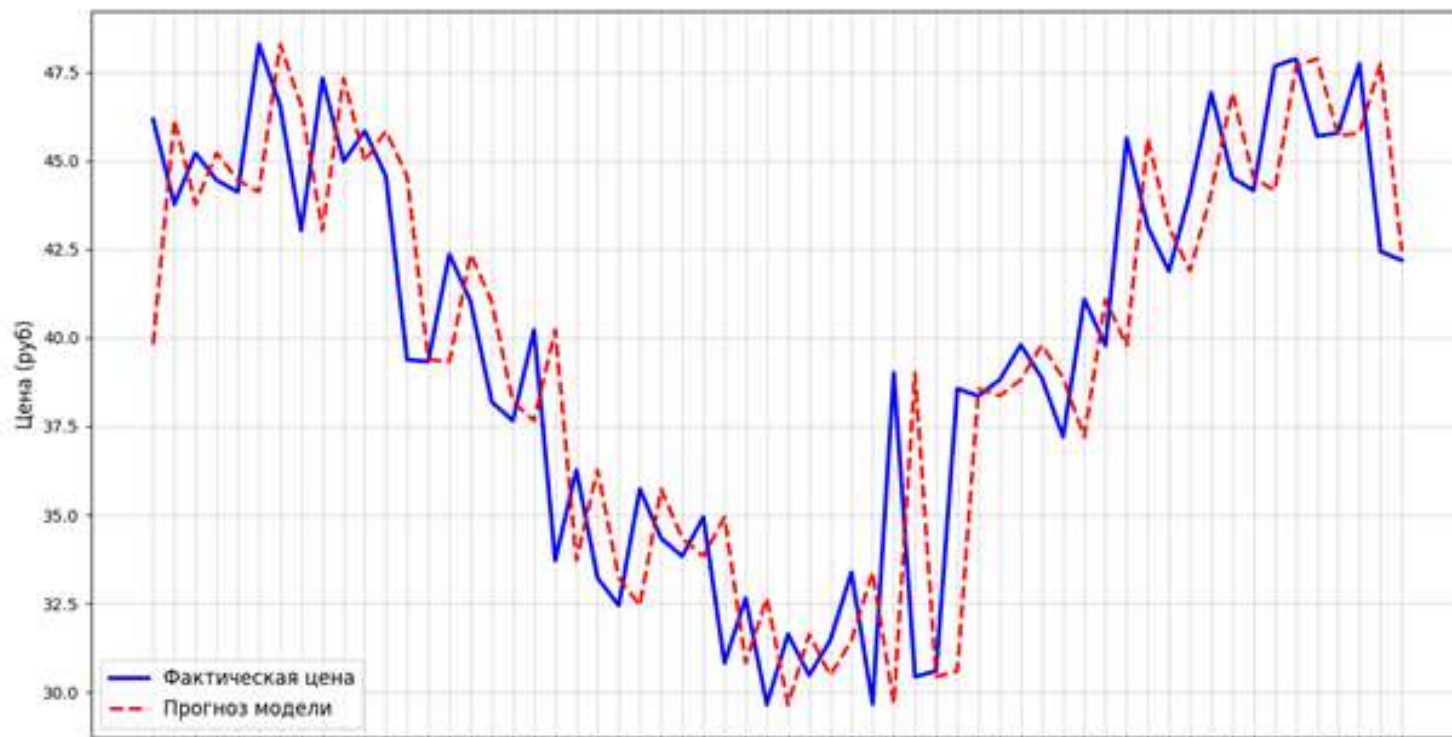
- Обучено 10 моделей (по одной на каждую пару «товар-регион»)
- Алгоритм: XGBoost (градиентный бустинг)
- Гиперпараметры: $n_estimators=200$ (количество деревьев), $learning_rate=0.05$ (скорость обучения), $max_depth=6$ (максимальная глубина каждого дерева)

Лучшие результаты:

Картофель (г. Москва) — MAE = **1.73 руб**

Говядина (г. Москва) — MAE = **2.41 руб**

Метрика	Значение (среднее)
MAE	2.1 – 2.7 <u>руб</u>
RMSE	2.1 – 3.5 <u>руб</u>
MAPE	2.8% – 5.4%
RMSLE	0.04 – 0.07
R^2	0.92 – 0.97

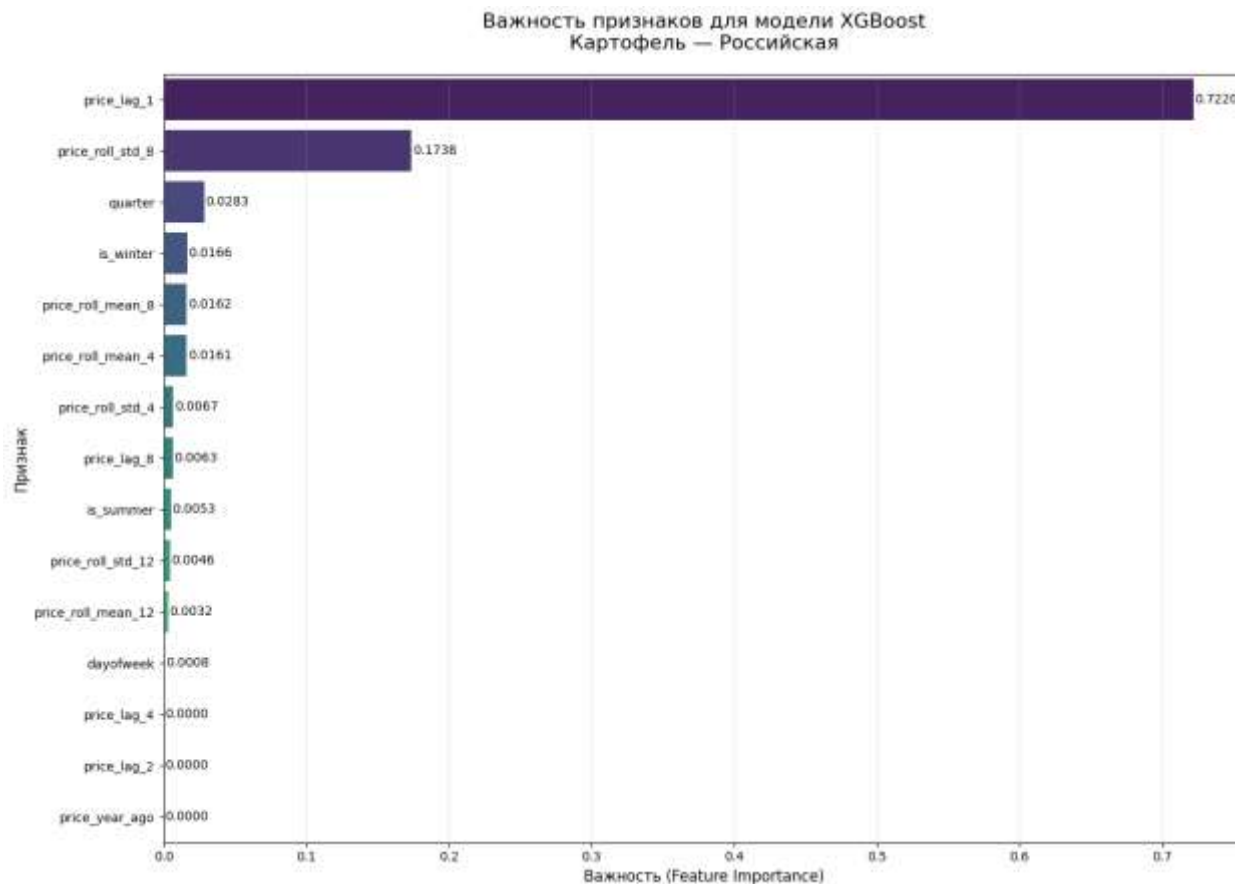




Анализ результатов



- Модель хорошо улавливает тренд и сезонность
- Наиболее важные признаки: price_lag_1, скользящие средние, месяц
- Остатки модели распределены близко к нулю (нет систематической ошибки)

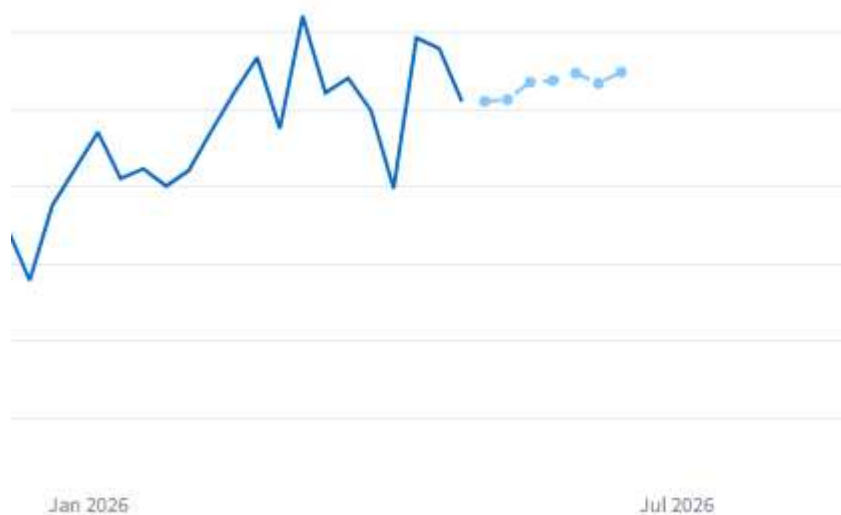




Прогнозирование



- Процесс прогнозирования:
- Загрузка последней обученной модели
- Генерация дат на 30/60/90 дней вперёд
- Применение модели + контролируемый шум (для реалистичности)
- Вывод прогноза в удобном виде



— Фактические данные
— Прогноз XGBoost



— Фактические данные
— Прогноз XGBoost



Итоговый интерфейс



Разработанный дашборд на Streamlit

- Вкладка 1: Аналитика по товару (графики, метрики, прогноз)
- Вкладка 2: Общая аналитика
- Вкладка 3: Сравнение товаров

Возможности:

- Выбор товара и региона
- Построение прогноза на 30/60/90 дней
- Интерактивные графики (Plotly)
- Автоматическое обновление данных

Аналитика по товару Общая аналитика **Сравнение товаров**

Сравнение товаров

product	mean	min	max
Картофель	37.75	20	52.63
Молоко питьевое цельное	67.68	52.89	80.75
Сахар-песок	77.86	63.28	92.8
Яйца куриные	98.07	83.05	112.44
Говядина (кроме бескостного мяса)	453.09	437.97	466.54

Аналитика по товару Общая аналитика Сравнение товаров

Аналитика по выбранному товару

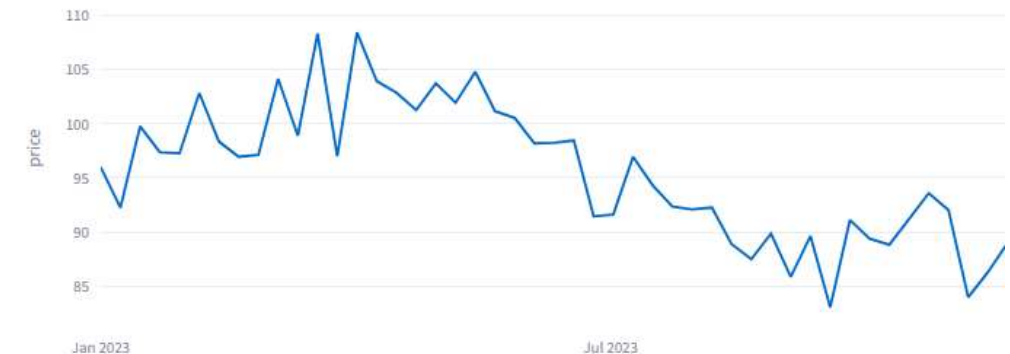
Товар

Яйца куриные

Текущая цена

112.44 ₽

Яйца куриные — Российская Федерация





Выводы и перспективы развития



Возможные перспективы:

- Добавление данных с маркетплейсов (Wildberries, Яндекс.Маркет, Ozon)
- Интеграция внешних факторов (погода, курс валют, инфляция)
- Реализация Telegram-бота с уведомлениями
- Создание гибридной модели (Prophet + XGBoost)
- Развёртывание в облаке (Docker + VPS)

Вывод:

- Цель работы достигнута в полном объёме
- Разработана система мониторинга и прогнозирования цен
- Получена модель XGBoost с высокой точностью ($MAE = 2,35$ руб)
- Создан удобный веб-дашборд и механизм автоматизации
- Работа имеет практическую значимость



Список литературы



1. Росстат. Официальная статистика потребительских цен [Электронный ресурс]. – URL: <https://rosstat.gov.ru/statistics/price> (дата обращения: 22.06.2026).
2. Streamlit Documentation [Электронный ресурс]. – URL: <https://docs.streamlit.io/>
3. XGBoost Documentation [Электронный ресурс]. – URL: <https://xgboost.readthedocs.io/>
4. Friedman J. H. Greedy Function Approximation: A Gradient Boosting Machine // Annals of Statistics. – 2001. – Vol. 29. – No. 5. – P. 1189–1232.
5. Breiman L. Random Forests // Machine Learning. – 2001. – Vol. 45. – P. 5–32.