Итоговый проект по курсу "Аналитик данных"

Предсказание цены криптовалюты (Time Series)

Выполнила: Шиляева Ольга Геннадьевна

2025 год

Почему Bitcoin?

Bitcoin является крупнейшей и наиболее ликвидной криптовалютой, что делает его идеальным объектом для исследования методов прогнозирования. Его курс часто служит индикатором общего состояния криптовалютного рынка.

- Высокая рыночная капитализация
- Широкое признание и ликвидность
- Индикатор рынка

Актуальность

- Криптовалютный рынок волатилен и быстро развивается
- Прогноз цен важен для инвесторов и аналитиков
- Недостаток фундаментальных факторов усложняет задачу
- Методы ML и временных рядов помогают выявить закономерности

Объект

- Биткойн и его ценовая динамика

Предмет

- Модели прогнозирования временных рядов и факторы (сырьевые, валютные, расчетные индикаторы), влияющие на цену

Цель

- Применить и сравнить модели прогнозирования цены криптовалют для выбора наиболее эффективного подхода

Задачи

1. Сбор и обработка данных
2. Выбор методов анализа временных рядов
3. Построение моделей (ARIMA, Linear, Random Forest, XGBoost, LSTM), графики прогнозов и ошибок
4. Оценка по метрикам (RMSE, MAE, MAPE)
5. Сравнение результатов
6. Выводы и рекомендации

Данные и исследовательский анализ (EDA)

Данные

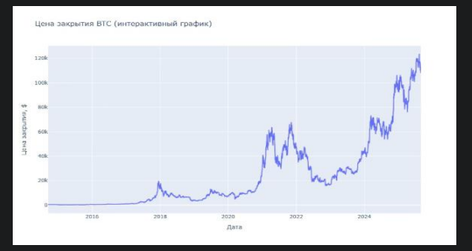
- Источник: Yahoo Finance
- Период: 2014–2025 гг.
- Переменные: Open, High, Low, Close, Volume

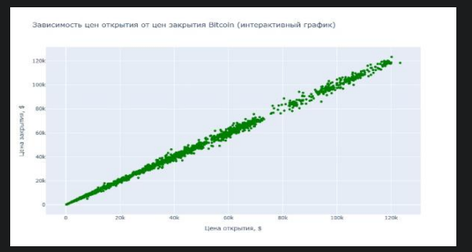
EDA нужно для:

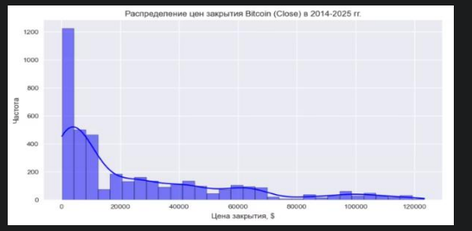
- Проверки качества данных (пропуски, ошибки, выбросы)
- Выявления трендов, сезонности и аномалий
- Поиска связей между переменными
- Подбора признаков и методов моделирования

Инструменты:

- корреляции, линейные графики, гистограммы, тепловая карта

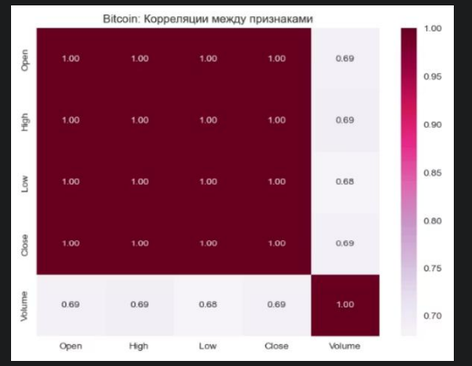






Вывод

- Распределение цен закрытия асимметричное: большинство значений до $ 20 000, редкие пики до $ 120 000 → высокая волатильность и спекулятивный характер рынка.
- Ценовые признаки (Open, High, Low, Close) дублируют друг друга → достаточно использовать **Close**.
- **Volume** слабее коррелирует с ценами и может быть полезен как дополнительный индикатор динамики и аномалий.



Выбор факторов и технических индикаторов

Для анализа взаимосвязей с ценой Bitcoin и повышения точности прогнозов были выбраны ключевые внешние активы и технические индикаторы, отражающие различные аспекты рынка.

Золото: Защитный актив

Традиционно рассматривается как "тихая гавань", помогая выявить корреляцию с рисковыми активами, такими как криптовалюты.

Нефть: Глобальный индикатор

Ключевой сырьевой товар, отражающий мировые экономические циклы и влияющий на поведение инвесторов.

Индийская рупия: Валюта развивающегося рынка

Позволяет оценить влияние валютных факторов на спрос и цену Bitcoin в одном из крупнейших крипторынков.

SMA (50, 200): Выявление трендов

Скользящие средние для определения краткосрочных и долгосрочных тенденций, сглаживания ценовых колебаний.

MACD (12, 26, 9): Сила и направление тренда

Измеряет расхождение экспоненциальных скользящих средних, выявляя силу тренда и возможные развороты.

ATR (14): Измерение волатильности

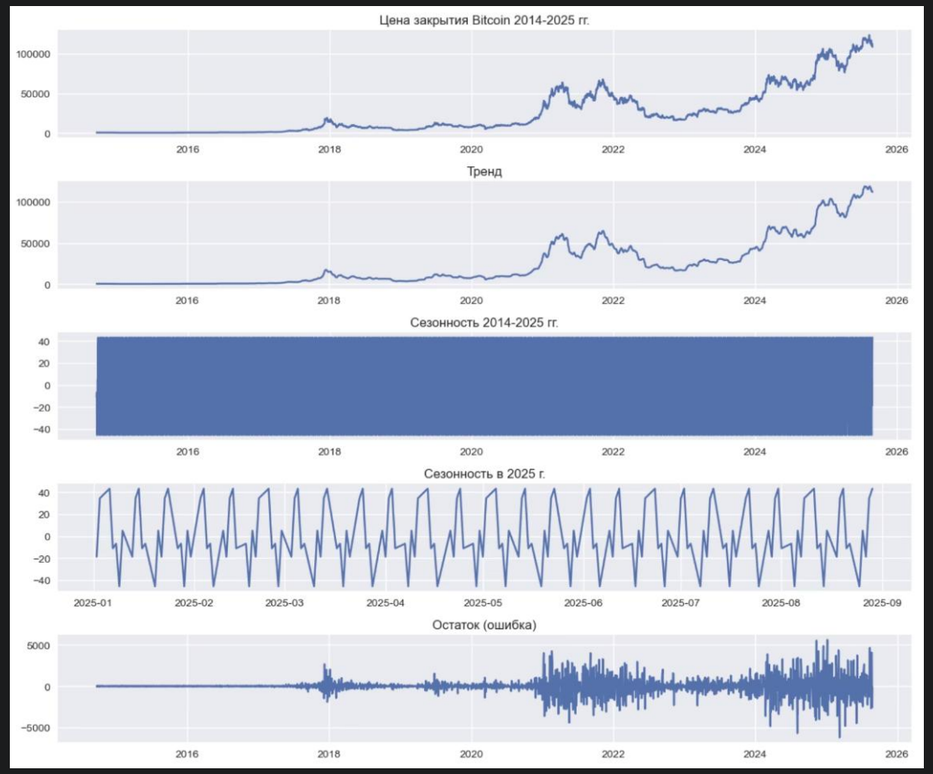
Отражает среднюю амплитуду колебаний цены за период, указывая на интенсивность рыночных движений.

Формирование итогового датасета

По результатам тщательного сбора данных и детального исследовательского анализа (EDA) был сформирован комплексный датасет с признаками, демонстрирующими сильную корреляцию с ценой закрытия Биткойна.

Эти отобранные данные послужат основой для дальнейшего построения и тестирования наших прогностических моделей.

Декомпозиция временного ряда



Декомпозиция временного ряда — это метод статистического анализа, позволяющий разложить наблюдаемый процесс на несколько составляющих: тренд, сезонность и остаточную часть (шум).

Тренд отражает долгосрочную **тенденцию** изменения ряда. **Сезонность** фиксирует регулярные колебания. **Остаток (ошибки)** показывает непредсказуемые флуктуации.

Применение декомпозиции к цене закрытия Биткойна (2014–2025 гг.) выявило:

- Долгосрочный восходящий тренд с резкими скачками и коррекциями.
- Наличие повторяющихся сезонных колебаний с амплитудой до ±50 единиц, однако их вклад относительно мал.
- Нерегулярные и резкие остаточные отклонения, наиболее выраженные в периоды кризисов и бурного роста (2017–2018 гг., 2021 г., 2024–2025 гг.), что указывает на значительные внешние факторы.

Таким образом, простая декомпозиция даёт ценную информацию, но недостаточна для построения точного прогноза. Для учёта автокорреляций, сезонных факторов и волатильности целесообразно переходить к более гибким стохастическим моделям, таким как ARIMA и SARIMA.

Использованные модели прогнозирования

Разведочный анализ данных и декомпозиция временного ряда выявили долгосрочный тренд, слабую сезонность и значительную "необъяснённую" часть, обусловленную внешними факторами. Это определило логику моделирования: от простых и интерпретируемых методов к более сложным, учитывающим нелинейность и расширенный набор признаков.

Цель — прогнозировать цену закрытия на следующий день (t+1) и объективно сравнить модели на едином тестовом интервале, избегая утечки данных. Оценка качества производится по метрикам RMSE, MAE и MAPE (в исходных единицах USD / процентах).

01

ARIMA и SARIMA

Классические модели временных рядов с учётом автокорреляции и сезонности. Параметры подбираются автоматически.

02

LinearRegression

Базовая линия: проверяет инерционную связь цены закрытия "вчера → завтра".

03

Random Forest и XGBoost

Нелинейные ансамбли деревьев, способные улавливать сложные зависимости на коротком горизонте.

04

LSTM (только Close)

Рекуррентная нейросеть, обучающаяся на окне недавних значений цены закрытия для моделирования краткосрочных паттернов.

05

LSTM (Multi-Feature)

Та же архитектура LSTM, но с расширенным набором признаков (Close, SMA_200, Volume, ATR_14, Gold, INR) для учёта внешнего контекста.

Такой подход от простого к сложному позволяет установить базовый уровень, определить вклад более комплексных моделей в точность прогнозов и обосновать выбор итоговой рабочей модели.

Ключевые метрики оценки качества прогнозирования

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE

Корень среднеквадратичной ошибки – среднее квадратичное отклонение предсказанных значений от фактических, чувствителен к большим ошибкам (выбросам) (USD).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MAE

Средняя абсолютная ошибка – среднее значение модуля отклонения предсказаний от фактических данных, значение MAE близко к средней «погрешности» модели (USD).

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

MAPE

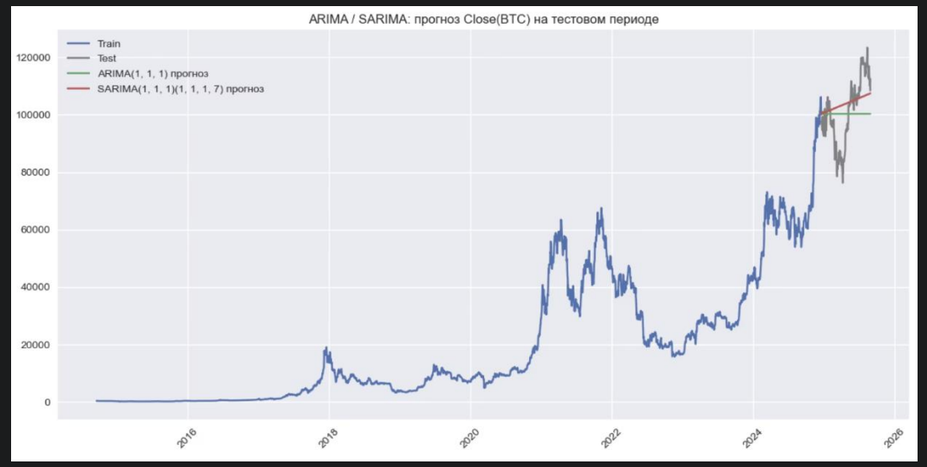
Средняя абсолютная процентная ошибка – отклонение в процентах от фактического значения, полезна для сравнения моделей на разных масштабах (%).

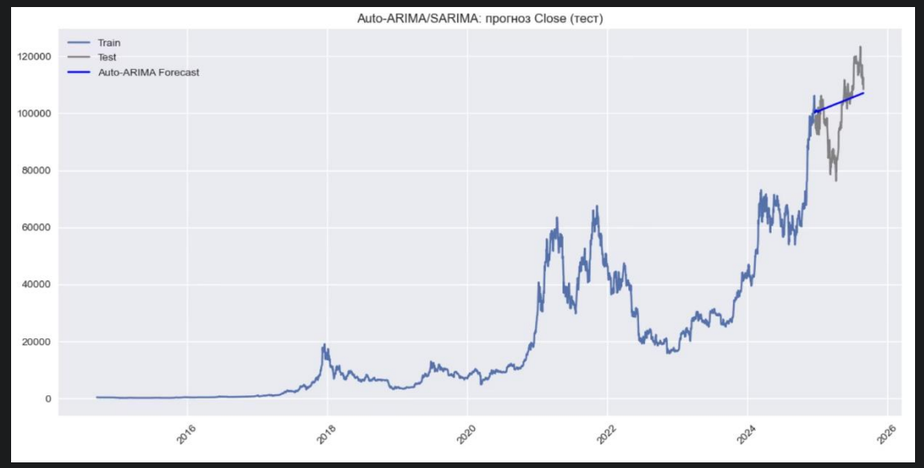
Классические методы прогнозирования: ARIMA, SARIMA, Auto-ARIMA

Модели семейства ARIMA (Autoregressive Integrated Moving Average) — это мощный инструмент для анализа и прогнозирования временных рядов. Они учитывают прошлые значения (AR), устраняют нестационарность (I) и принимают во внимание прошлые ошибки прогноза (MA) для создания точных предсказаний.

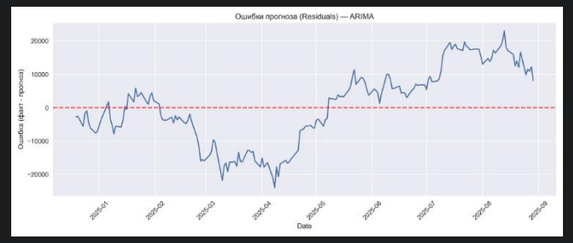
SARIMA (Seasonal ARIMA) расширяет возможности ARIMA, добавляя сезонные компоненты для работы с данными, демонстрирующими регулярные сезонные паттерны. В то же время, Auto-ARIMA автоматизирует подбор оптимальных параметров для ARIMA и SARIMA, упрощая и ускоряя процесс построения модели.

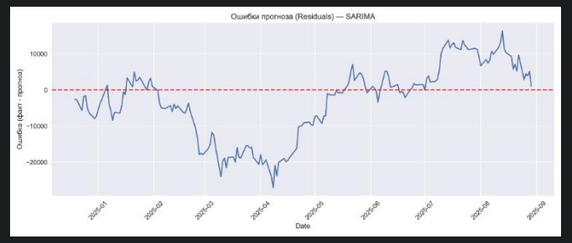
Эти методы эффективно применяются для прогнозирования цен на криптовалюты, включая Bitcoin, так как они способны улавливать сложные тенденции, цикличность и случайные колебания в динамике рынка, обеспечивая ценные инсайты для будущего.

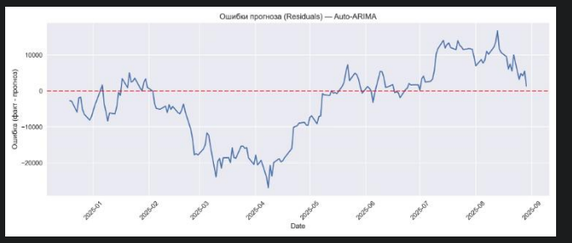


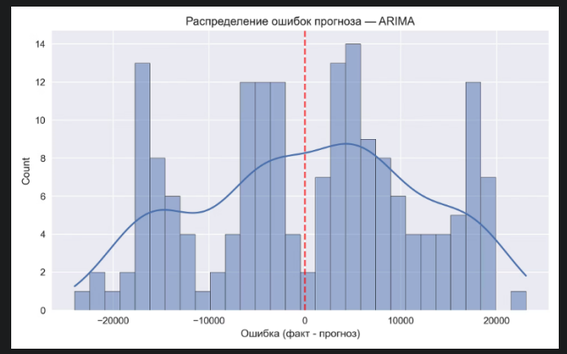


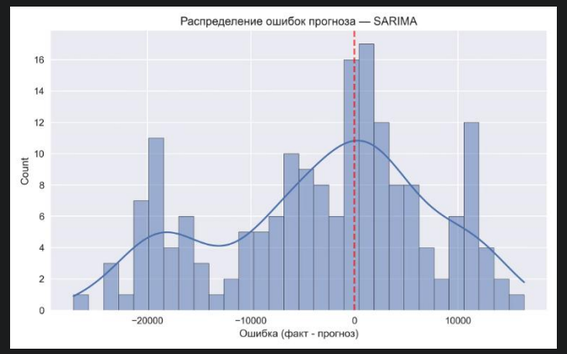
Анализ ошибок прогнозов ARIMA, SARIMA, Auto-ARIMA

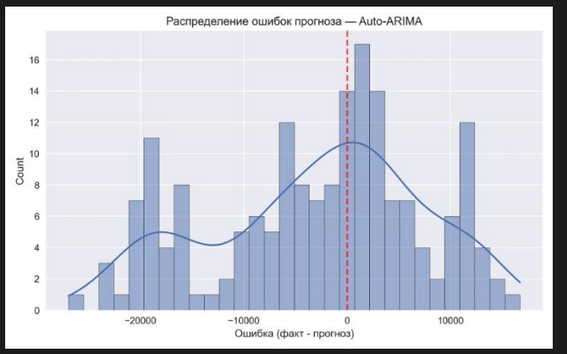












ARIMA

RMSE = $ 11 360,31

MAE = $ 9 592,98

MAPE = 9,73%

SARIMA

RMSE = $ 10 534,83

MAE = $ 8 169,51

MAPE = 8.69%

Auto-ARIMA

RMSE = $ 10 539,07

MAE = $ 8 213,40

MAPE = 8,71%

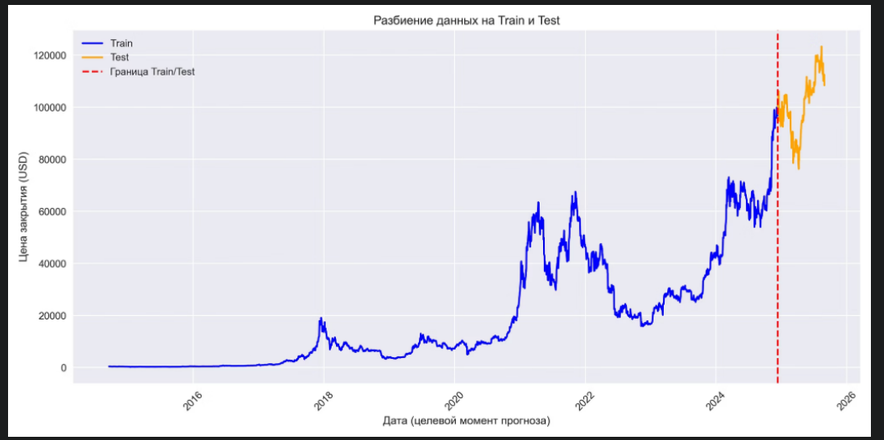
Анализ показал, что ARIMA и SARIMA способны уловить тренд и сезонные колебания в данных, однако их точность остаётся ограниченной: ошибки прогнозов достигают значительных величин, что связано с высокой волатильностью криптовалютного рынка. Даже использование автоматического подбора параметров не устраняет этой проблемы. Такие результаты указывают на необходимость обращения к более сложным методам прогнозирования, способным лучше учитывать нестабильный характер динамики Bitcoin.

Формирование выборки

Для формирования обучающей и тестовой выборок был реализован последовательный процесс подготовки данных:

- Итоговый датасет с выбранными коррелирующими признаками был скопирован и преобразован, индекс с датами перенесён в отдельный столбец.
- Определены основные признаки на момент времени **t** (Close, Gold, INR, SMA_200, Volume, ATR_14) и целевая переменная на **t+1** (Close_{t+1}), с добавлением колонки даты прогноза.
- Данные очищены от пропусков, обеспечивая согласованный набор признаков и целевой переменной.
- Финальная выборка разделена на обучающую и тестовую части по временной оси: обучающая до последних 180 дней, тестовая — этот заключительный период.

Такой подход обеспечивает корректную проверку моделей в условиях временной последовательности данных и позволяет оценить их способность прогнозировать цену криптовалюты на ещё не использованных в обучении отрезках.



Линейная регрессия

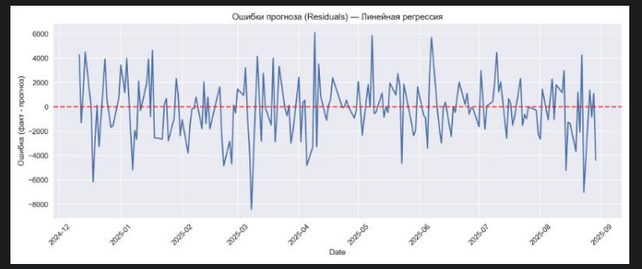
Линейная регрессия — это базовый метод машинного обучения, который моделирует зависимость целевой переменной от набора факторов (признаков) с помощью линейной комбинации. Модель имеет вид:

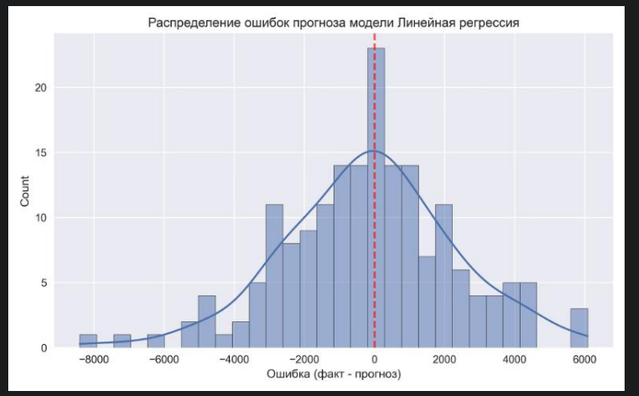
$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Этот метод служит отправной точкой (бенчмарком) для сравнения со сложными моделями и предоставляет простой, интерпретируемый способ оценки влияния признаков на целевую переменную.

В данном проекте линейная регрессия используется для прогнозирования цены закрытия Bitcoin на следующий день. В качестве признаков выступают лаги цены, SMA_200, цена золота, объем торгов и индикатор ATR_14.

Модель поможет оценить линейную зависимость цены Bitcoin от выбранных факторов, а ее результаты станут базой для сравнения с нелинейными моделями (Random Forest, XGBoost) и нейросетями (LSTM).

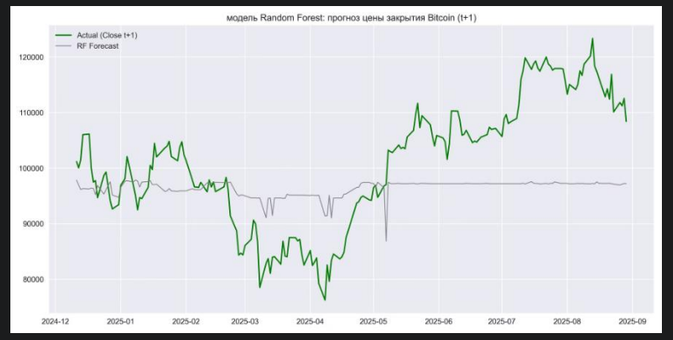


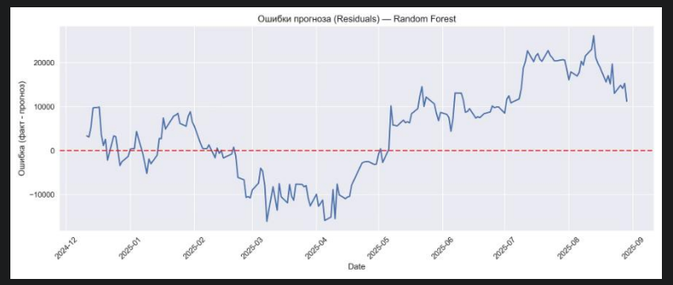


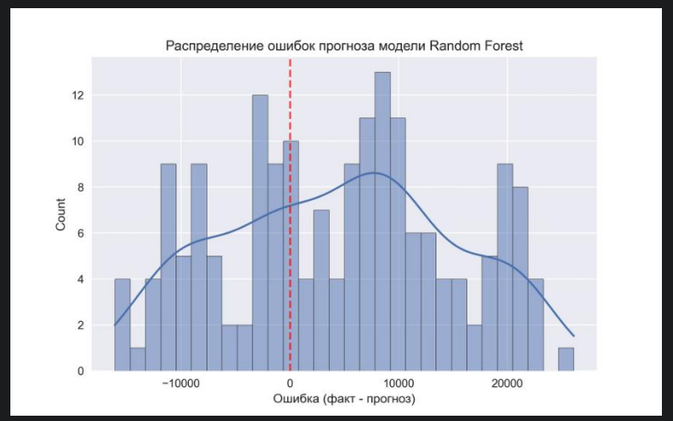
RMSE = $ 2 445,73, MAE = $ 1 867,67, MAPE = 1,90%

- Ошибка прогноза менее **2% (MAPE = 1.9%)** — модель показывает высокую точность.
- Главный фактор прогноза — текущая цена закрытия BTC; влияние остальных признаков минимально.
- Модель интерпретируема и надёжна как **базовый ориентир**, но ограничена при высокой волатильности.

Random Forest







Random Forest — ансамблевый метод, использующий множество решающих деревьев для повышения точности и устойчивости прогнозов.

Преимущества:

- Улавливает нелинейные зависимости.
- Устойчив к выбросам и шуму.

В данном исследовании Random Forest прогнозирует цену закрытия Bitcoin, используя лаги цены, SMA_200, ATR_14, объём торгов и цену золота. Это позволяет оценить, насколько нелинейные методы улучшают качество прогнозов по сравнению с линейной регрессией.

RMSE = $ 11 354,30, MAE = $ 9 458,95, MAPE = 9,18%

- Ошибка прогноза высока (**MAPE ≈ 9.2%**), модель уступила линейной регрессии.
- Основной вклад — текущая цена BTC, остальные признаки малозначимы.
- Наблюдается систематическое смещение: модель недооценивала рост цены.
- Для прогноза на один день вперёд Random Forest оказался менее эффективным.

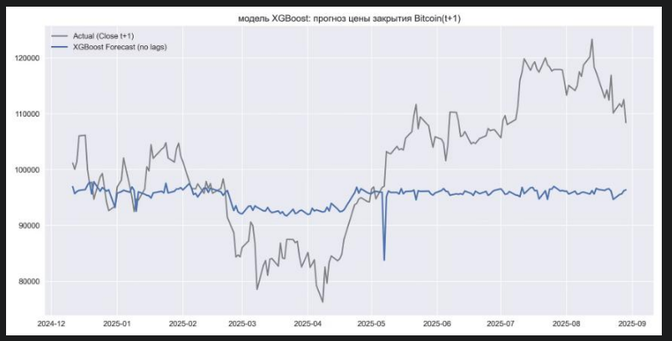
XGBoost (Extreme Gradient Boosting)

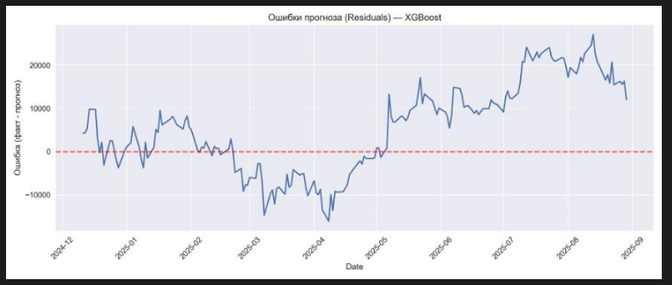
XGBoost — это мощная реализация градиентного бустинга, использующая ансамбль деревьев решений. Модель последовательно строит деревья, каждое из которых исправляет ошибки предыдущих, создавая сильный предсказатель сложных зависимостей в данных.

Преимущества:

- Высокая точность за счёт учёта нелинейных связей и взаимодействий признаков.
- Встроенная регуляризация предотвращает переобучение.
- Возможность оценки важности признаков (Feature Importance, SHAP-анализ).

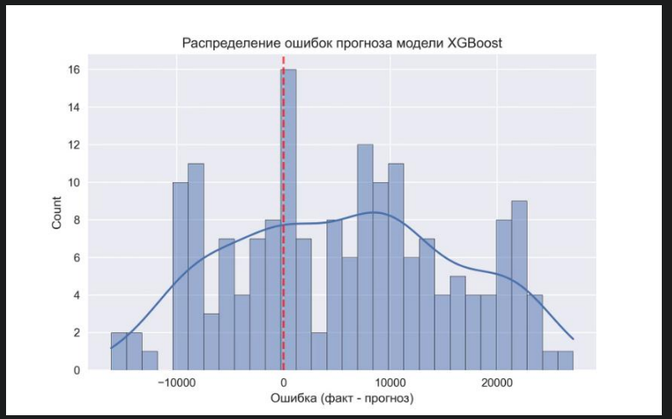
В проекте XGBoost прогнозирует цену закрытия Bitcoin, используя лаги цены, SMA_200, объем торгов, волатильность (ATR_14), цену золота и курс INR. Это позволяет оценить, насколько бустинговый алгоритм улучшит прогноз по сравнению с линейной регрессией и Random Forest.



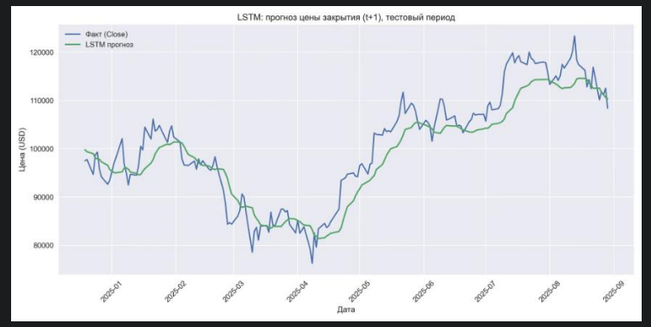


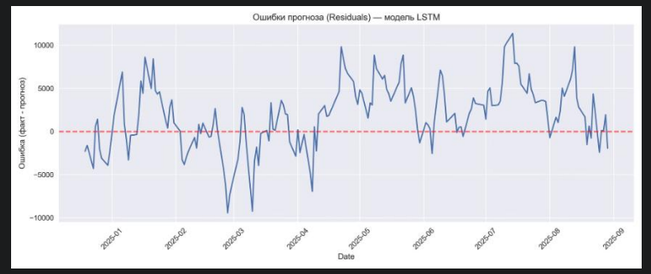
RMSE = $ 11 780,89, MAE = $ 9 637,86, MAPE = 9,23%

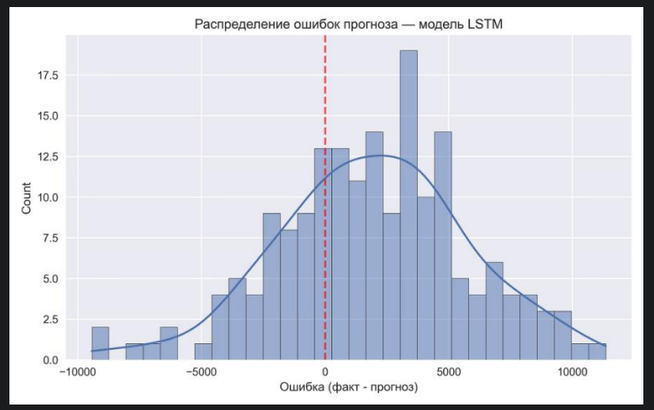
- Ошибка прогноза высокая (**MAPE ≈ 9.2%**), результат сопоставим с Random Forest и хуже линейной регрессии.
- Ключевой фактор — текущая цена BTC (76%); SMA_200 и ATR_14 имеют умеренное влияние, остальные признаки незначимы.
- Модель систематически недооценивала цену (ошибки до $25 000).
- Несмотря на сложность алгоритма, XGBoost оказался неэффективным без дополнительных временных признаков.



LSTM (Long Short-Term Memory)







LSTM – это тип рекуррентных нейронных сетей, разработанный для эффективной работы с последовательными данными и временными рядами. Она способна "запоминать" информацию на больших интервалах, решая проблемы обычных RNN, и идеально подходит для улавливания нелинейных зависимостей.

В этом проекте мы используем LSTM для прогнозирования цены Bitcoin, чтобы оценить её способность превосходить классические модели (линейная регрессия, Random Forest, XGBoost) в условиях высокой волатильности, используя те же метрики качества: RMSE, MAE, MAPE.

RMSE = $ 4 267,35, MAE = $ 3 405,81, MAPE = 3,36%

- Высокая точность прогноза (**MAPE < 4%**) и отсутствие переобучения.
- Модель хорошо улавливает как долгосрочные, так и краткосрочные зависимости.
- Ошибки сбалансированы, распределены вокруг нуля; смещение небольшое.
- LSTM показала одно из лучших качеств среди всех протестированных моделей.

Многопризнаковая (multi-feature) модель LSTM

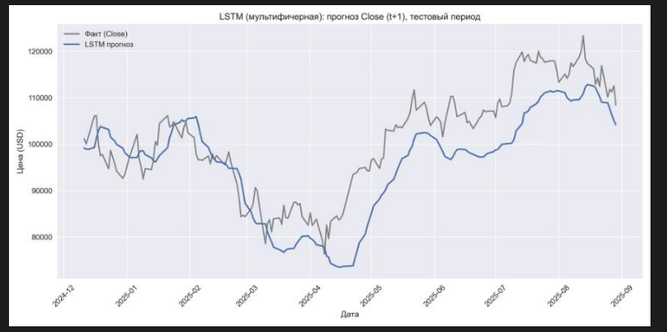
Мы расширяем вход LSTM, включая дополнительные признаки, описывающие состояние рынка и внешнюю среду: **SMA_200**, **Volume**, **ATR_14**, **Gold** и **INR**.

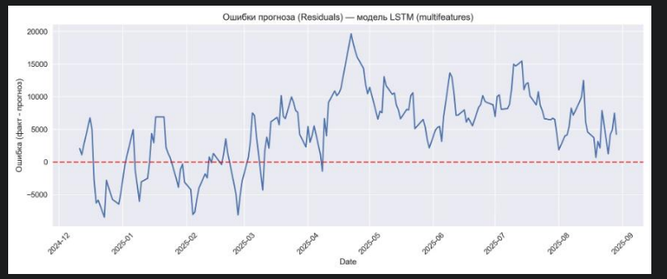
Это преобразует одномерный вход в многомерный тензор признаков. Целевая переменная (цена закрытия на следующий день) остаётся неизменной.

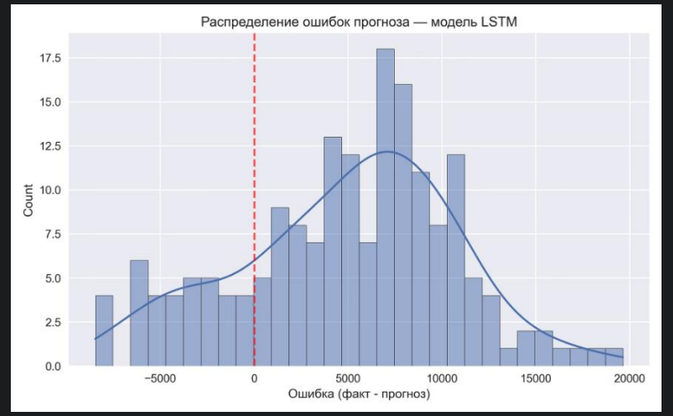
Ожидается снижение метрик ошибок **RMSE**, **MAE** и **MAPE**, что повысит предсказательную способность модели.

RMSE = $ 7 539,22, MAE = $ 6 463,17, MAPE = 6,40%

- Качество прогноза хуже, чем у одномерной модели (**MAPE ≈ 6.4% против 3.4%**).
- Модель сглаживает динамику и недооценивает пики (ошибки до +10–15 тыс. USD).
- Добавленные признаки внесли шум и не улучшили прогноз, частично дублируя информацию из Close.







Интерпретация результатов

Сравнительный анализ различных моделей прогнозирования цены Bitcoin позволил выявить их сильные и слабые стороны.

1

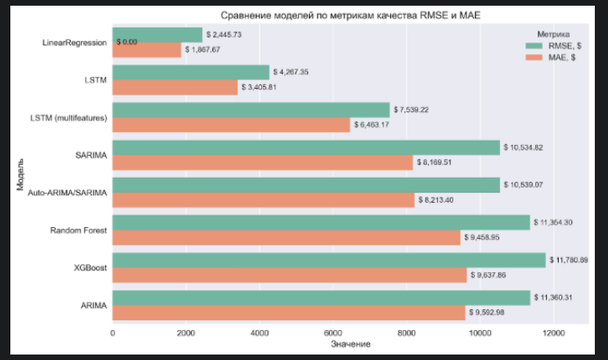
Линейная регрессия

Несмотря на простоту, показала удивительно низкую ошибку (MAPE = 1.90%), став надёжным бенчмарком. Основной фактор — текущая цена BTC, что указывает на высокую автокорреляцию временного ряда.

2

LSTM (одномерная)

Однопризнаковая LSTM (MAPE = 3.36%) продемонстрировала одно из лучших качеств прогноза. Сеть успешно «запомнила» долгосрочные и краткосрочные зависимости, обеспечив сбалансированное распределение ошибок.



RMSE и MAE

4

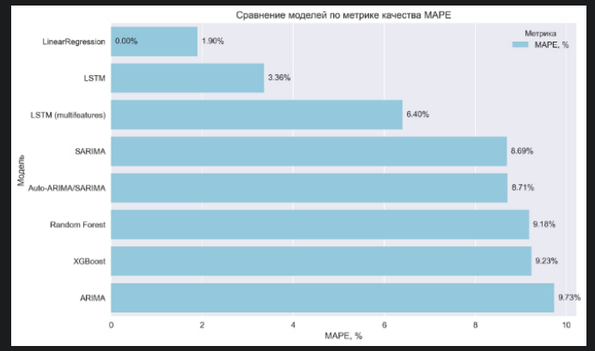
Ансамблевые модели (RF, XGBoost)

Random Forest (MAPE = 9.18%) и XGBoost (MAPE = 9.23%) показали худшие результаты. Они не смогли эффективно уловить динамику временного ряда без явных временных признаков, переобучаясь на шуме и недооценивая пики.

3

LSTM (многопризнаковая)

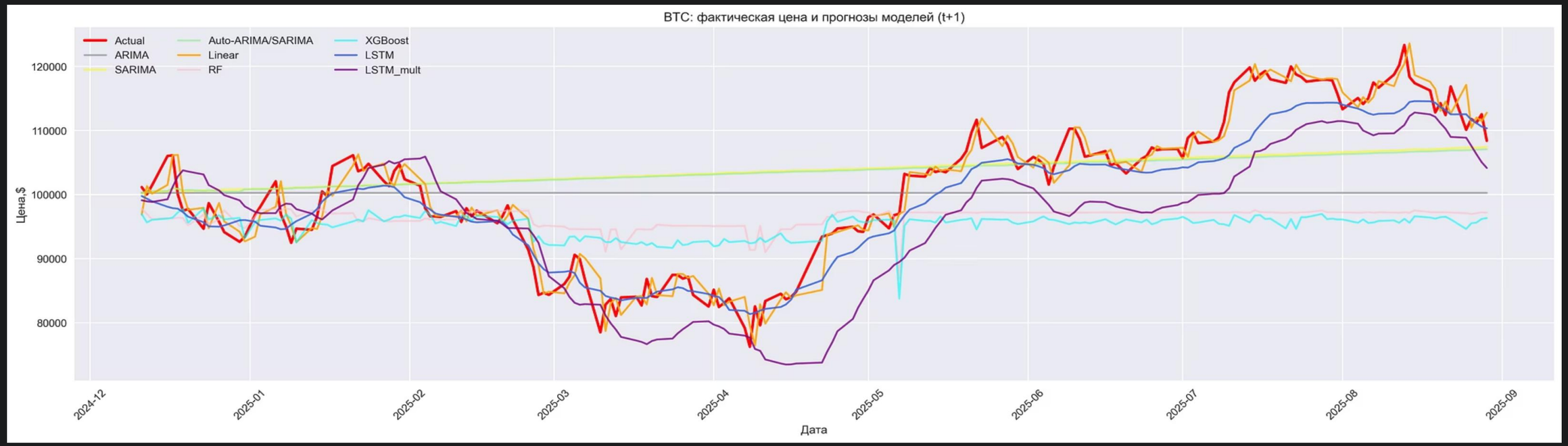
Добавление внешних признаков к LSTM привело к ухудшению качества (MAPE = 6.40%). Это указывает на то, что дополнительные данные в текущем виде внесли больше шума, чем полезной информации для прогноза цены BTC.



MAPE

Наиболее эффективным оказался простой подход с использованием одномерной LSTM и линейной регрессии, что подчеркивает сложность прогнозирования криптовалют.

Сводный график прогнозов моделей



На данном графике представлены прогнозы цен Bitcoin от всех рассмотренных моделей в сравнении с фактическими данными. Визуальное сопоставление позволяет оценить, насколько точно каждая модель улавливает динамику рынка и её способность реагировать на изменения цены.

Отчётливо видно, что одномерная модель LSTM и линейная регрессия показали наиболее близкие к реальности прогнозы, в то время как ансамблевые методы, такие как Random Forest и XGBoost, демонстрируют большее отклонение и сглаживание пиков и падений.

Заключение

Эффективность моделей

Линейная регрессия и одномерная LSTM показали наилучшие результаты для краткосрочного прогнозирования цены Bitcoin (MAPE 1.9% и 3.36% соответственно). Эти модели продемонстрировали высокую точность, несмотря на простоту или одномерный вход.

Ограничения и вызовы

Ансамблевые модели (Random Forest, XGBoost) и многофакторная LSTM оказались менее эффективными, вероятно, из-за шума от дополнительных признаков и сложности улавливания динамики. Рынок криптовалют остаётся высоко волатильным и чувствительным к внешним факторам.

Практическая ценность

Применение методов анализа данных и машинного обучения позволяет формировать более обоснованные прогнозы и снижать неопределенность, что повышает надежность инвестиционных решений на криптовалютном рынке.

Перспективы исследований

Дальнейшие исследования могут включать расширение набора признаков (новости, соцсети, данные о майнинге) и использование гибридных моделей для повышения точности и устойчивости прогнозов.

В целом, проект подтверждает применимость продвинутых методов для анализа и прогнозирования цен криптовалют, подчеркивая важность качественной подготовки данных и адаптивного выбора модели.

Спасибо за внимание