Предсказание инфляции

Артюкевич Арина

31 августа 2025 г.

Оглавление

Глава 1

Введение

1.1. Описание выбранного датасета

Датасет T10YIE, предоставленный Федеральным резервным банком Сент-Луиса (FRED), отражает ожидания рынка относительно средней ставки инфляции. Датасет охватывает период с 2 января 2003 года по 27 августа 2025 года. Данные представлены в процентах, не скорректированы по сезонности и обновляются ежедневно.

Датасет включает более 5700 наблюдений. Показатель отражает динамику ожидаемой инфляции, захватывая различные экономические циклы и кризисные периоды. В 2003-2007 годах значения находились в диапазоне 2.0%-2.5%, демонстрируя стабильные инфляционные ожидания в период экономического роста. Во время финансового кризиса 2008-2009 годов наблюдалось значительное снижение показателя до уровня около 0.5%-1.0%. В последующий период 2010-2019 годов значения постепенно восстановились до диапазона 1.5%-2.5%. В 2020 году вновь произошло снижение из-за пандемии COVID-19, за которым последовал резкий рост в 2021-2022 годах до уровней выше 3.0%. К августу 2025 года значение стабилизировалось на уровне 2.41%, указывая на возврат к умеренным долгосрочным инфляционным ожиданиям.

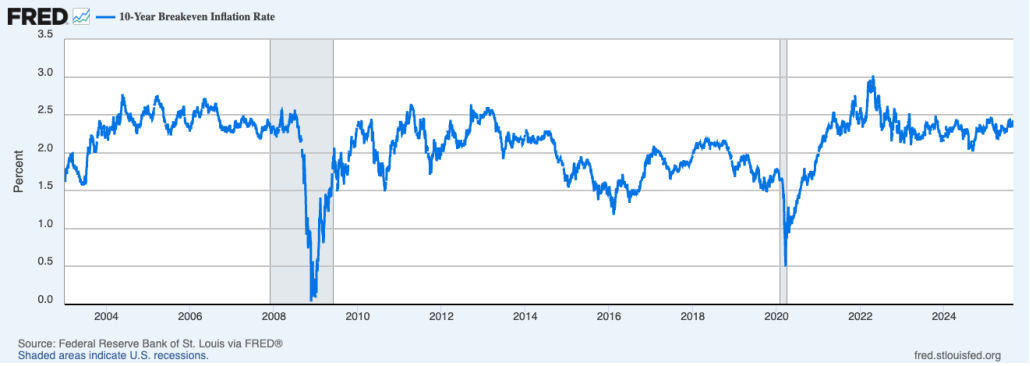

Рис. 1.1: Оригинальный временной ряд T10YIE

1.2. Постановка задачи прогнозирования

Основной целью данной работы является построение моделей машинного обучения для краткосрочного прогнозирования (short-term forecasting) временного ряда T10YIE на один шаг вперед (one-step-ahead forecast).

Формально модель можно описать как функцию f, которая отображает исторические данные и признаки в будущее значение:

$$\hat{y}_{t+1} = f(y_t, y_{t-1}, ..., y_{t-n}, X_t) \tag{1.1}$$

где:

- $\hat{y}_{t+1}$ – прогнозируемое значение T10YIE на день $t+1$,
- $y_t, y_{t-1}, ..., y_{t-n}$ – значения ряда в моменты времени $t, t-1, ..., t-n$ (лаговые переменные),
- X_t – вектор дополнительных признаков (скользящие статистики, календарные признаки и т.д.) на момент времени t,
- n – размер окна исторических данных.

1.2.1. Целевая переменная и горизонт прогнозирования

- **Целевая переменная (Target Variable):** Исходный временной ряд – ежедневные значения процента ожидаемой инфляции (T10YIE).
- **Горизонт прогнозирования (Forecasting Horizon):** $h = 1$ день. Выбор такого горизонта обусловлен высокой волатильностью финансовых данных и практической ценностью для принятия оперативных решений.

1.2.2. Критерии оценки

Для оценки точности прогнозов и сравнения моделей будут использоваться следующие метрики:

1. **MAE (Mean Absolute Error):** Показывает среднюю величину ошибки в абсолютном выражении (в процентах). Основная интерпретируемая метрика.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1.2}$$

2. **RMSE (Root Mean Squared Error):** Уделяет больше внимания крупным ошибкам, что критично в финансовом прогнозировании.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1.3}$$

3. **MAPE (Mean Absolute Percentage Error):** Позволяет оценить ошибку в относительном выражении.

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{1.4}$$

4. **SMAPE (Symmetric Mean Absolute Percentage Error):** Симметричная версия MAPE, менее чувствительная к малым значениям.

$$\text{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{1.5}$$

5. R^2 **(Коэффициент детерминации):** Отражает долю дисперсии целевой переменной, объясненную моделью.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.6}$$

где:

- y_i – истинное значение целевой переменной,
- $\hat{y}_i$ – предсказанное моделью значение,
- $\bar{y}$ – среднее значение целевой переменной,
- n – количество наблюдений в тестовой выборке.

Успешной будет считаться модель, которая демонстрирует стабильно низкие значения ошибок (MAE, RMSE, MAPE, SMAPE) и высокое значение R^2 на тестовой выборке, что будет свидетельствовать о ее способности к обобщению и практической применимости.

1.3. Гипотезы исследования

В рамках данной работы были выдвинуты и проверены следующие гипотезы:

1. **Гипотеза о временной структуре:** Временной ряд T10YIE содержит значимые автокорреляции и сезонные компоненты, что делает возможным его прогнозирование с использованием исторических значений. Предполагается, что значения лагов ряда являются статистически значимыми предикторами будущих значений.
2. **Гипотеза о нелинейных зависимостях:** Взаимосвязь между историческими значениями временного ряда и будущими наблюдениями носит нелинейный характер. Модели, способные учитывать такие сложные зависимости (такие как Gradient Boosting или LSTM), будут показывать более высокую точность прогнозирования по сравнению с линейными методами.
3. **Гипотеза о значимости дополнительных признаков:** Генерация дополнительных признаков (календарных, статистических, скользящих статистик) позволяет улучшить качество прогноза по сравнению с использованием исключительно лаговых значений временного ряда.
4. **Гипотеза о стационарности ряда:** После применения соответствующих разностей временной ряд T10YIE может быть приведен к стационарному виду, что является необходимым условием для применения некоторых классических методов прогнозирования.

Глава 2

Анализ данных

2.1. Описательная статистика

Для проведения первичного анализа временного ряда T10YIE была рассчитана описательная статистика основных характеристик распределения данных. Результаты представлены в Таблице 2.1 .

Таблица 2.1: Описательная статистика временного ряда T10YIE

Параметр	Значение
count	5668.0000
mean	2.0992
std	0.3976
min	0.0400
25%	1.8500
50%	2.2100
75%	2.3600
max	3.0200
skewness	-1.2655
kurtosis	2.7165

- **Центральная тенденция:** Среднее значение ожидаемой инфляции составляет 2.0992% при медиане 2.21%, что указывает на длинный хвост слева.
- **Изменчивость:** Стандартное отклонение (0.3976) свидетельствуют об умеренной волатильности показателя в течение анализируемого периода.
- **Распределение:** Положительный эксцесс (2.7165) характеризует распределение с более острой вершиной по сравнению с нормальным распределением.
- **Диапазон значений:** Значения показателя варьируются от 0.04% до 3.02%, что отражает значительные изменения инфляционных ожиданий в различные экономические периоды.

2.2. Визуализация временного ряда

2.2.1. Распределение данных

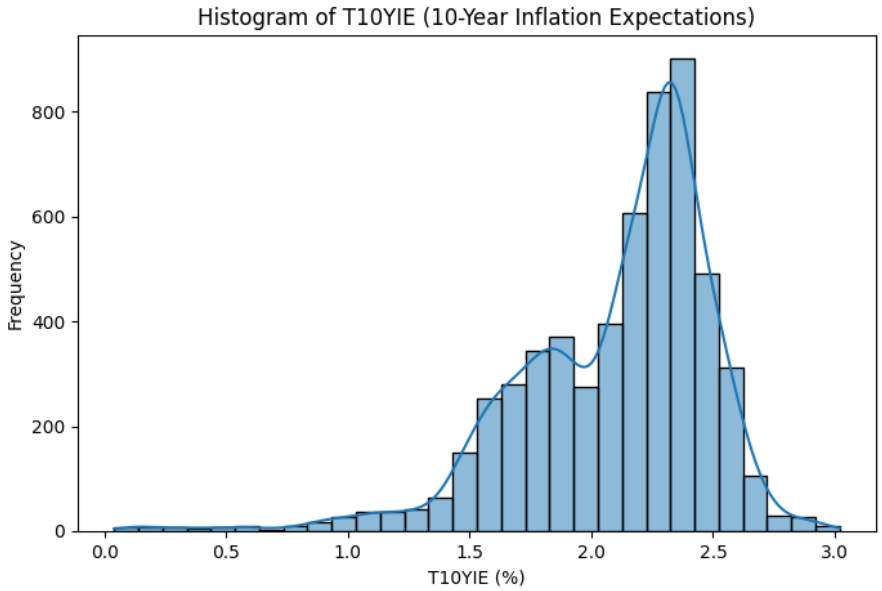

Рис. 2.1: Гистограмма распределения значений T10YIE

Гистограмма (Рис. 2.1) демонстрирует распределение значений T10YIE. Анализ показывает бимодальное распределение с пиками в районе 1.7% и 2.3%, что соответствует различным экономическим состояниям.

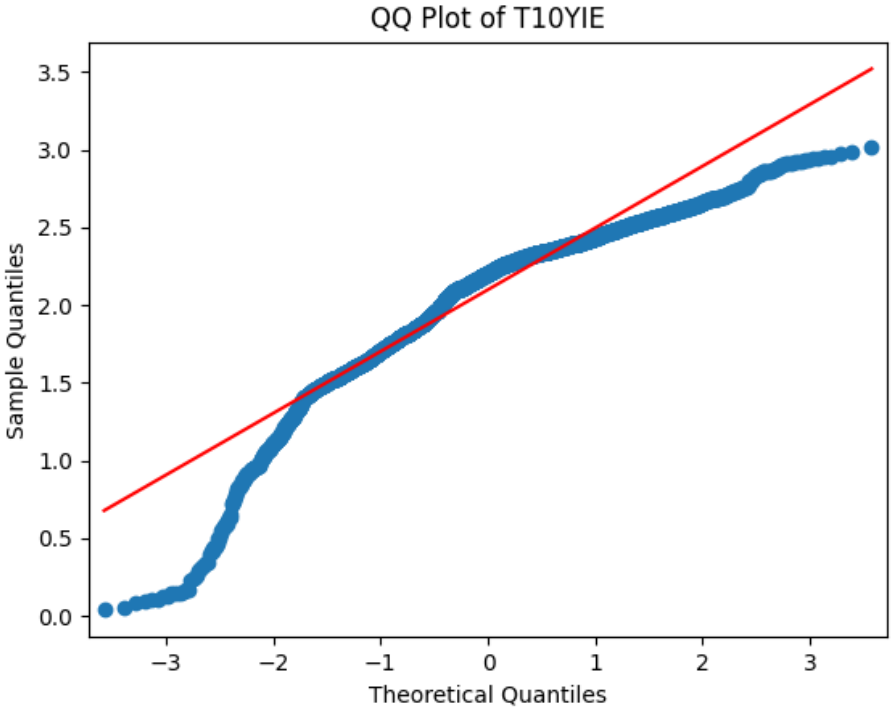

Рис. 2.2: Q-Q plot распределения T10YIE

Q-Q plot (Рис. 2.2) показывает значительное отклонение от нормального распределения, особенно в хвостах.

2.2.2. Анализ выбросов

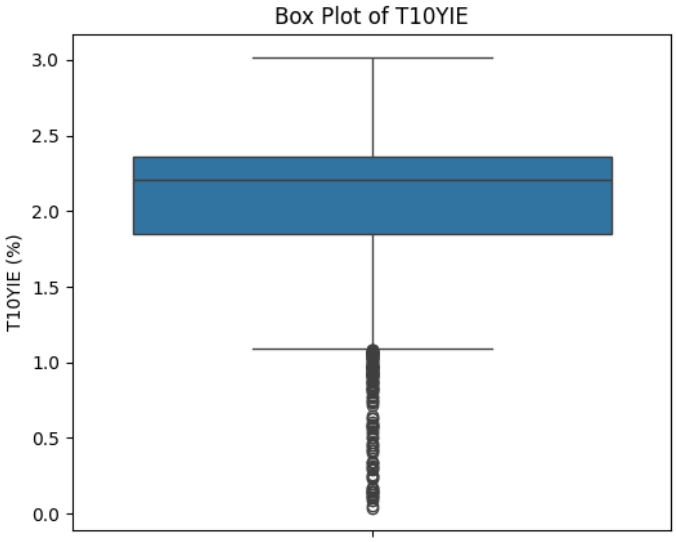

Рис. 2.3: Box plot для выявления выбросов T10YIE

Box plot (Рис. 2.3) показывает наличие значительных выбросов, что соответствует периодам экономической нестабильности.

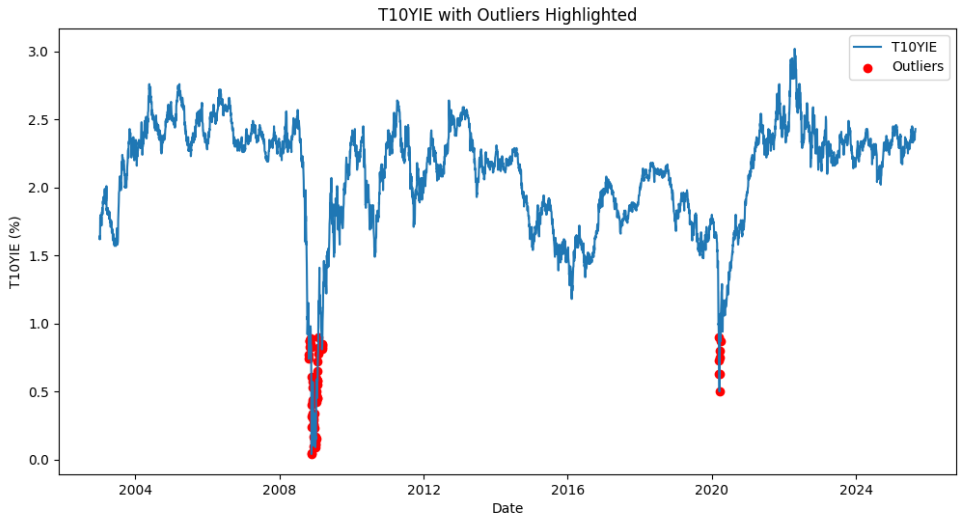

Рис. 2.4: Временной ряд T10YIE с выделенными выбросами

Временной ряд с выбросами (Рис. 2.4) демонстрирует кластеризацию выбросов в периоды финансовых кризисов (2008-2009, 2020-2022 гг.).

2.2.3. Декомпозиция временного ряда

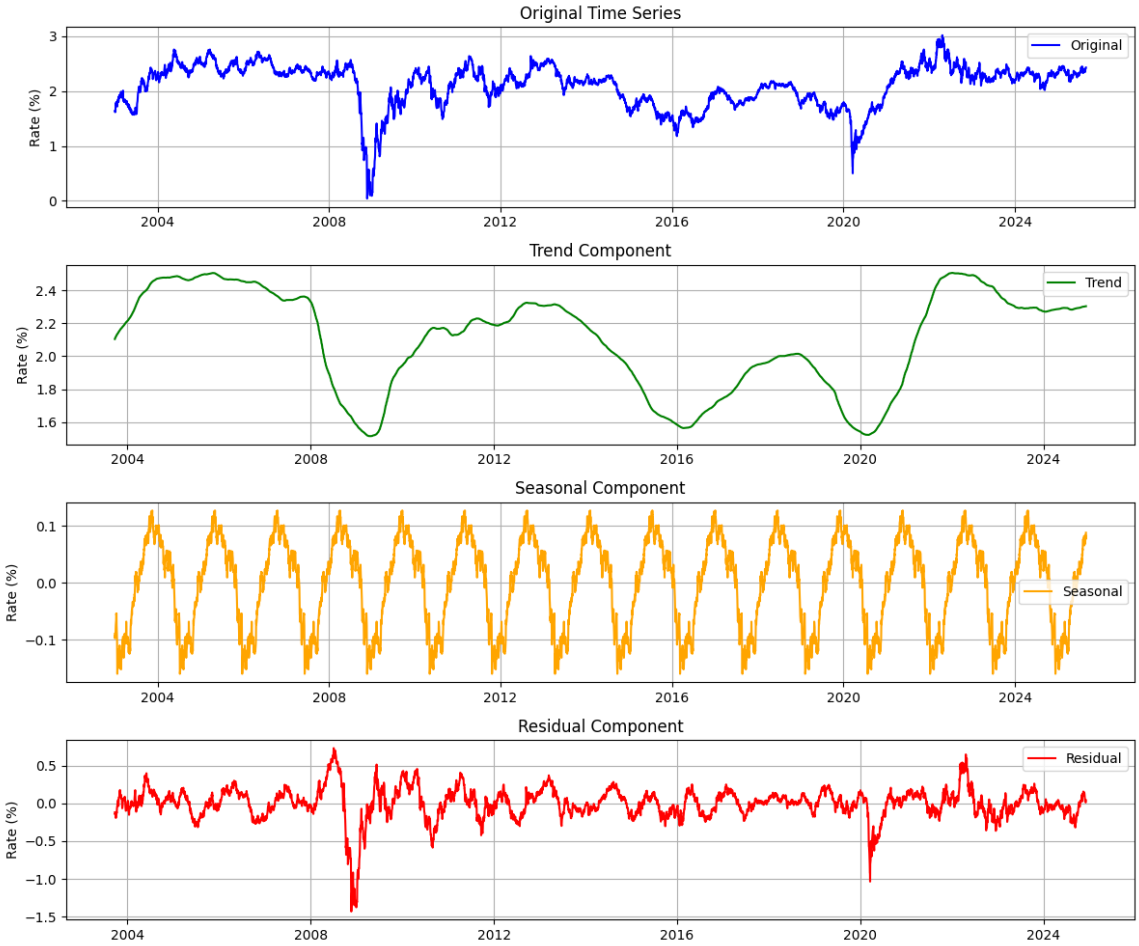

Рис. 2.5: Аддитивная декомпозиция временного ряда T10YIE

Декомпозиция (Рис. 2.5) показывает наличие выраженного тренда и сезонной компоненты, что подтверждает сложную структуру ряда.

2.2.4. Корреляционный анализ

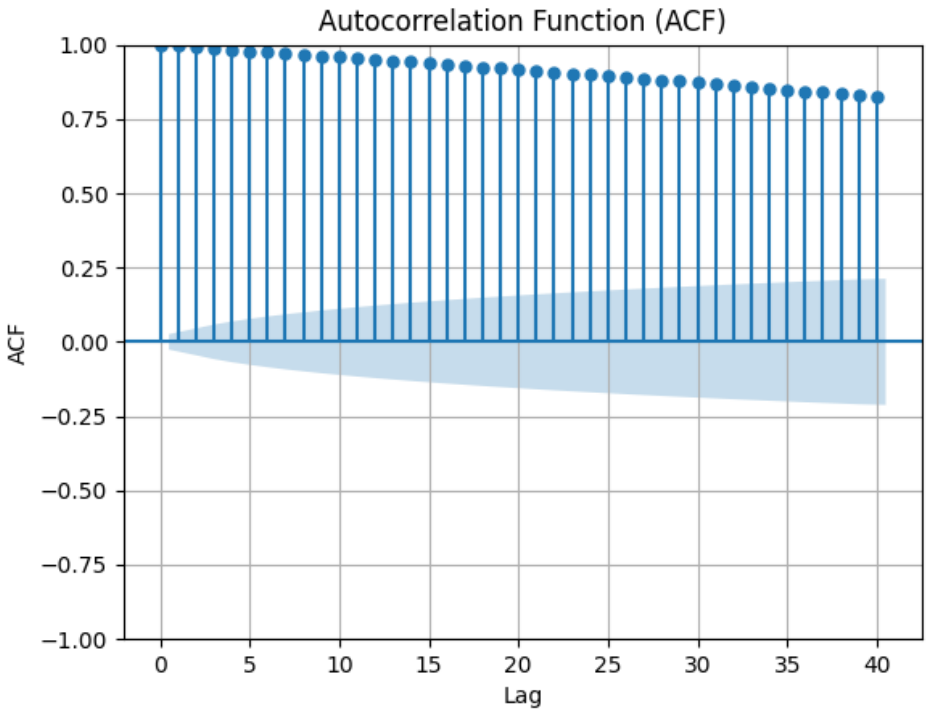

Рис. 2.6: Автокорреляционная функция (ACF) T10YIE (40 лагов)

ACF (Рис. 2.6) показывает:

- Медленное затухание корреляционной функции
- Высокую положительную автокорреляцию на первых лагах
- Значимые автокорреляции на протяжении 40 лагов

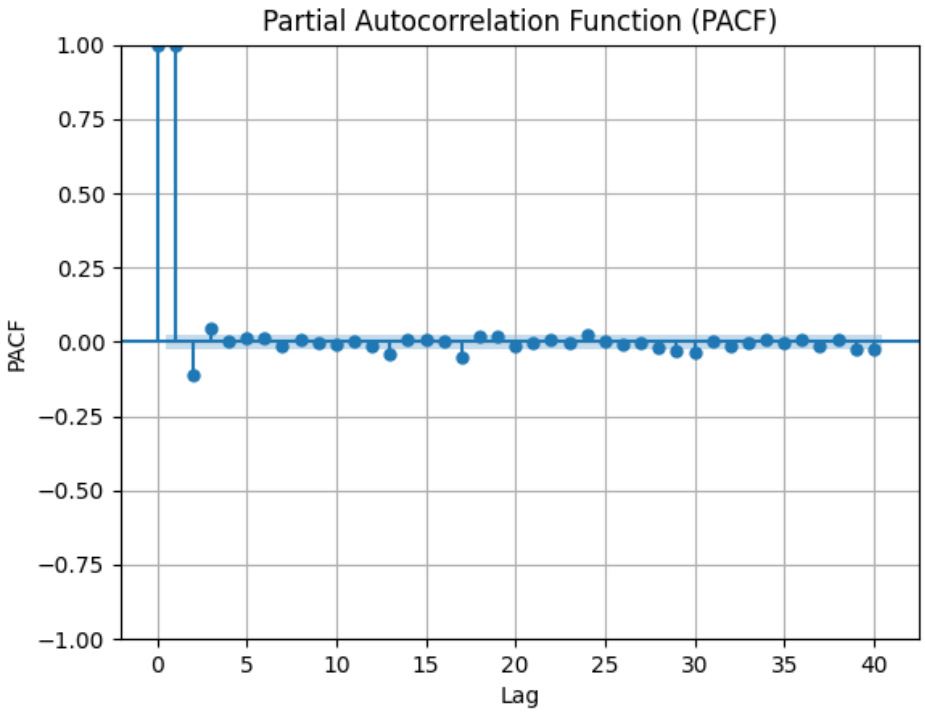

Рис. 2.7: Частная автокорреляционная функция (PACF) T10YIE (40 лагов)

PACF (Рис. 2.7) демонстрирует:

- Резкое падение после первых 2-3 лагов
- Значимые частные автокорреляции на лагах 1 и 2

2.3. Результаты статистических тестов

2.3.1. Тест на стационарность (Augmented Dickey-Fuller)

Результаты теста показывают, что временной ряд является стационарным:

- **ADF статистика:** -3.5010
- **p-value:** 0.0080
- **Критические значения:**
  - 1% уровень: -3.4315
  - 5% уровень: -2.8621
  - 10% уровень: -2.5670

Вывод: Формально ADF статистика превышает критическое значение на уровне 1% значимости, однако визуальный анализ ряда и его компонент указывает на наличие нестационарности. В том числе, наличие выбросов и нестабильной дисперсии (особенно в периоды кризисов). Данное противоречие может быть вызвано структурными особенностями или наличием детерминированного тренда.

2.3.2. Тест на нормальность (Shapiro-Wilk)

Распределение остатков значимо отличается от нормального:

- **Статистика:** 0.8960
- **p-value:** 0.0000

Вывод: Низкое значение статистики (0.8960) и крайне малый p-value (0.0000) свидетельствуют о существенном отклонении от нормального распределения.

2.3.3. Тест на автокорреляцию (Ljung-Box)

В остатках присутствует значимая автокорреляция:

- **Тест на 40 лагах:** $p < 0.05$

Вывод: Тест показывает наличие статистически значимой автокорреляции в остатках на различных лагах.

2.3.4. Интерпретация результатов

На основе проведенных тестов можно сделать следующие выводы:

- **Стационарность:** Формальные критерии указывают на стационарность, но визуальный анализ предполагает обратное. Нестационарность ряда также объясняет медленное затухание автокорреляционной функции и наличие значимых автокорреляций на больших лагах.
- **Нормальность распределения:** Остатки имеют ненормальное распределение, что указывает на необходимость использования методов, устойчивых к отклонениям от нормальности, или применения соответствующих преобразований.
- **Автокорреляция:** Наличие автокорреляции в остатках указывает на то, что модель не полностью улавливает все зависимости в данных.

Полученные результаты указывают, что модели, учитывающие автокорреляционную структуру и не предполагающие нормальности распределения, будут наиболее подходящими для данного временного ряда.

2.3.5. Анализ разностного ряда

Для преобразования исходного нестационарного ряда к стационарному рассматривается разность первого порядка. Анализ разностного ряда (2.8) показывает:

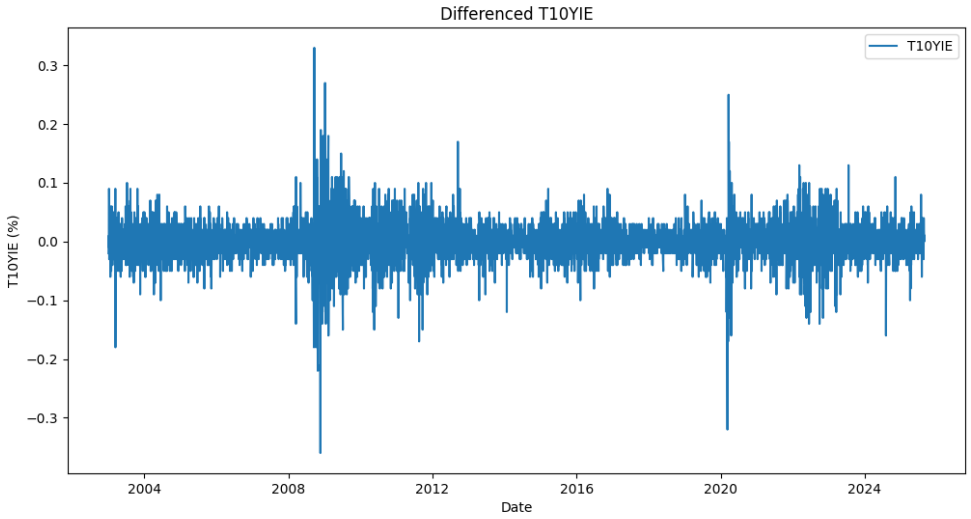

Рис. 2.8: Разностный временной ряд T10YIE

- **Стабилизация среднего уровня:** Разностный ряд колеблется вокруг нулевого уровня.
- **Постоянная дисперсия:** Амплитуда колебаний стала более однородной по сравнению с исходным рядом, хотя некоторые периоды повышенной волатильности сохраняются.
- **Выявление краткосрочных изменений:** Разностный ряд более четко показывает ежедневные изменения инфляционных ожиданий, устраняя долгосрочные тенденции.
- **Сохранение кластеров волатильности:** Периоды повышенной изменчивости соответствуют временам финансовой нестабильности (2008-2009, 2020-2022 гг.).

Глава 3

Генерация признаков

3.1. Подробное описание всех категорий признаков

Для построения эффективных прогнозных моделей был сгенерирован набор признаков, включающий следующие категории:

3.1.1. Лаговые признаки (Lag Features)

- **Простые лаги (lag_1...lag_20):** Значения ряда за предыдущие дни. Полезны, так как текущее значение часто зависит от недавнего прошлого (инерция рынка).
- **Сезонные лаги (seasonal_lag_20...60):** Значения с фиксированным сдвигом (20 дней). Могут выявлять повторяющиеся месячные/цикличные паттерны.
- **Разности (diff_1, diff_2):** Первая и вторая разности ряда. Преобразуют нестационарный ряд в стационарный.
- **Лаги разностей (diff_1_lag_1...2):** Изменения за предыдущие дни. Полезны для прогнозирования динамики (ускорения/замедления).
- **Логарифмические лаги (log_lag_1...10):** Натуральный логарифм значений. Сжимает шкалу, делает данные более нормальными, полезен для моделирования процентных изменений.

3.1.2. Скользящие статистики (Rolling Window Features)

Признаки, рассчитанные на скользящем окне (6, 8, 12, 14, 18 дней), которые описывают локальные свойства ряда:

- **Среднее (ma_w) и Медиана (median_w):** Определяют локальный уровень (тренд) ряда. Медиана устойчива к выбросам.
- **Стандартное отклонение (std_w) и Дисперсия (var_w):** Измеряют локальную волатильность.
- **Минимум (min_w) и Максимум (max_w):** Определяют локальные экстремумы.
- **Размах (range_w):** Общая амплитуда колебаний в окне.
- **Асимметрия (skew_w) и Эксцесс (kurt_w):** Описывают форму распределения. Асимметрия указывает на смещение, эксцесс — на "тяжелые хвосты".

- **Квантили (q25_w, q75_w) и IQR (iqr_w):** Описывают центральную часть распределения. IQR — устойчивая мера разброса.

3.1.3. Признаки экспоненциального сглаживания (EMA Features)

Динамические средние, придающие больший вес недавним наблюдениям (с параметрами α=0.1-0.9):

- **EMA (ema_α):** Базовое экспоненциальное среднее. Сглаживает шум, определяет краткосрочный тренд.
- **DEMA (dema_α) и TEMA (tema_α):** Двойное и тройное сглаживание. Уменьшают lag (запаздывание) обычного EMA, быстрее реагируют на изменения.
- **Адаптивное EMA (adaptive_ema):** Параметр сглаживания α динамически приспосабливается к волатильности.

3.1.4. Технические индикаторы (Technical Indicators)

Стандартные индикаторы технического анализа, адаптированные для временного ряда:

- **RSI (rsi_14/21/30):** Index Relative Strength Index. Индекс относительной силы для периодов 14, 21, 30.
- **Stochastic Oscillator (stoch_k, stoch_d):** Сравнивает цену закрытия с диапазоном за период. Сигнальная линия (D) сглаживает основную (K).
- **Williams %R (williams_r):** Моментум-индикатор, аналогичный Stochastic, но с инвертированной шкалой.
- **Rate of Change (roc_10/20/30) и Momentum (momentum_10/20/30):** Измеряют абсолютное и относительное изменение цены за период.
- **MACD (macd, macd_signal, macd_hist):** Moving Average Convergence Divergence. Разница между краткосрочной и долгосрочной EMA.
- **Parabolic SAR (psar):** Stop and Reverse. Точечный индикатор, следующий за трендом и указывающий точки изменения тренда.
- **ADX (adx):** Average Directional Index. Измеряет силу тренда.
- **Bollinger Bands (bb_upper, bb_middle, bb_lower, bb_width, bb_position):** Границы скользящего среднего. Ширина границ (bb_width) — мера волатильности. Позиция внутри границ (bb_position) — относительный уровень.
- **ATR (atr):** Average True Range. Измеряет волатильность, основанную на полном диапазоне.
- **Volatility Ratio (vol_ratio):** Отношение краткосрочной (14д) к долгосрочной (50д) волатильности. Рост указывает на учащение резких изменений.

3.1.5. Временные признаки (Date Features)

Признаки, кодирующие дату:

- **Линейные (dayofmonth, quarter, dayofyear, weekofyear):** Номерные атрибуты даты. Могут выявлять циклы (например, конец квартала).
- **Категориальные (dayofweek_0-4, month_1-12):** One-Hot Encoding дня недели и месяца. Учитывают weekly/monthly эффекты (например, волатильность по пятницам).
- **Циклические (sin/cos dayofweek, month, dayofyear):** Преобразуют циклическое время в координаты на окружности. Сохраняют непрерывность (например, воскресенье (0) близок к понедельнику (1)).
- **Праздники (is_holiday):** Бинарный признак. Рынки часто ведут себя иначе в праздники (низкие объемы, аномальные движения).
- **Начало/Конец месяца (is_start/end_month):** Периоды ребалансировок, приток / отток средств.

3.1.6. Статистические признаки (Statistical Features)

Признаки, описывающие статистические свойства ряда в окне:

- **Скользящая ACF (acf_lag_1/2/3/7/14/30):** Автокорреляция на разных лагах в окне. Изменения могут предвещать смену тренда.
- **Скользящие Skewness (skew_w) и Kurtosis (kurt_w):** Динамика асимметрии и "тяжести хвостов" распределения.
- **Энтропия Шеннона (shannon_entropy_w):** Мера неопределенности/случайности в окне. Низкая энтропия может указывать на тренд.
- **Mean Absolute Deviation (mad_w):** Устойчивая к выбросам мера разброса.
- **Coefficient of Variation (coeff_var_w):** Отношение std к mean (нормализованная волатильность). Полезен для сравнения разных периодов.
- **Z-Score Difference (z_score_diff_w):** Изменение стандартизованной цены за окно. Мера momentum.
- **Variance Ratio (var_ratio_5_20, var_ratio_10_50):** Отношение дисперсий. Проверка на кластеризацию волатильности (Volatility Clustering).
- **Rolling Correlation (corr_lag_1/7/14):** Корреляция текущих значений с лагами в окне. Показывает устойчивость краткосрочной автокорреляции.

3.1.7. Комбинированные признаки (Mixed Features)

Синтетические признаки, объединяющие информацию из разных источников:

- **Отношения EMA (ema_ratio_0.5_0.1, ema_ratio_0.7_0.3):** Сравнение трендов разных периодов. Может сигнализировать о смене тренда.
- **Разности EMA (ema_diff_0.5_0.1):** Разрыв между трендами.

- **Расстояние до EMA (dist_to_ema_0.3/0.5/0.7):** Отклонение цены от тренда.
- **Композитный моментум (momentum_composite):** Усреднение нескольких нормированных моментум-индикаторов. Дает более устойчивый сигнал.
- **Z-Score цены и RSI (z_score_rate, z_score_rsi):** Нормированные отклонения. Показывают, насколько текущее значение экстремально в недавнем контексте.
- **Экстремумы (is_peak, is_trough):** Бинарные признаки локальных максимумов/минимумов. Могут маркировать точки смены тренда.
- **Направление тренда (trend_direction):** Знак изменения скользящего среднего (+1 вверх, -1 вниз). Определяет основной тренд.

Глава 4

Отбор признаков

4.1. Описание алгоритмов Forward и Backward Selection

Для отбора наиболее информативных признаков из сгенерированного набора были применены два классических алгоритма последовательного отбора: Forward Selection и Backward Selection.

4.1.1. Forward Selection (Последовательное включение)

Алгоритм начинается с пустого набора признаков и на каждом шаге добавляет один признак, который наиболее улучшает качество модели. Критерием остановки служит отсутствие значимого улучшения (порог: 0.1%) при добавлении нового признака либо достижение максимального количества признаков (20).

4.1.2. Backward Selection (Последовательное исключение)

Алгоритм начинается с полного набора всех сгенерированных признаков и на каждом шаге удаляет один признак, удаление которого приводит к наименьшему ухудшению качества модели. Критерием остановки служит превышение допустимой деградации качества (порог: 0.1%) либо достижение минимального количества признаков (20).

4.1.3. Методика оценки

Для оценки качества модели на каждом шаге использовалась 5-кратная кросс-валидация с метрикой R^2 (коэффициент детерминации). В качестве базовой модели применялась линейная регрессия предсказания инфляции на следующий день.

4.2. Результаты применения каждого метода

4.2.1. Forward Selection

Метод forward selection отобрал 18 наиболее значимых признаков. На рисунке 4.1 показана динамика изменения R^2 по мере добавления признаков.

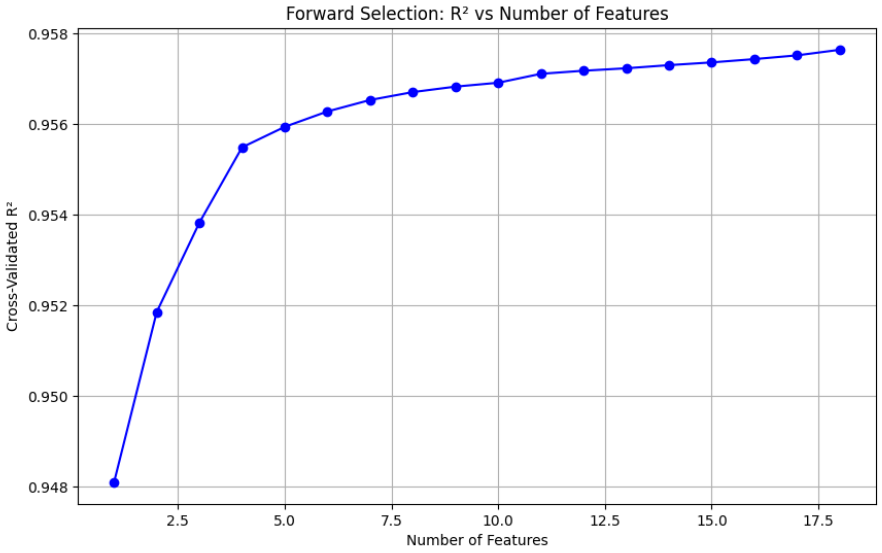

Рис. 4.1: Динамика R^2 при последовательном добавлении признаков (Forward Selection)

Наибольший вклад в качество модели внесли следующие признаки (порядок добавления):

1. `T10YIE` - текущее значение ряда
2. `is_peak` - индикатор локального максимума
3. `is_trough` - индикатор локального минимума
4. `lag_1` - значение ряда в предыдущий день
5. `month_2` - индикатор февраля

Финальное значение R^2 на кросс-валидации составило 0.9576, что демонстрирует высокую объясняющую способность отобранных признаков.

4.2.2. Backward Selection

Метод backward selection сократил исходный набор со 213 до 34 признаков. На рисунке 4.2 показана динамика R^2 при последовательном удалении признаков.

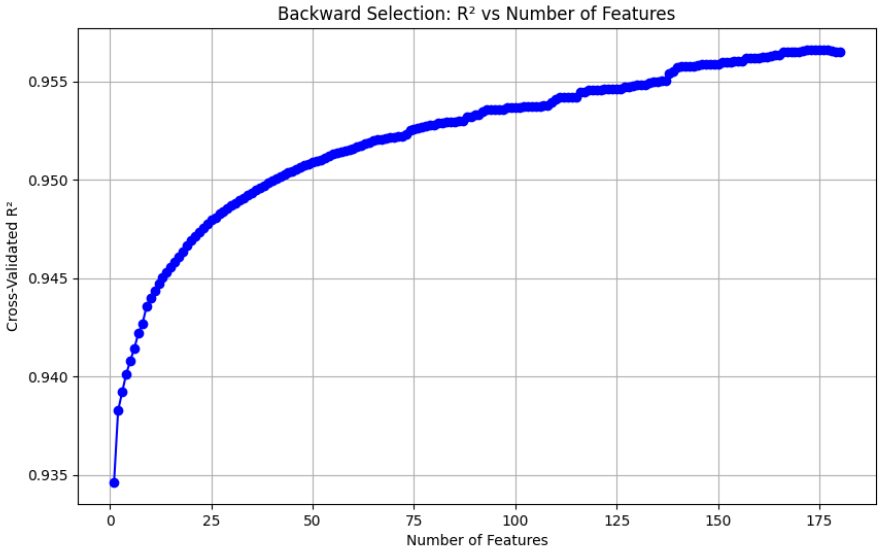

Рис. 4.2: Динамика R^2 при последовательном удалении признаков (Backward Selection)

Процесс отбора выявил следующие закономерности:
Наименее значимые признаки (удалены в первую очередь):

1. Сезонные лаги с большими периодами: `seasonal_lag_60`, `seasonal_lag_40`
2. Производные признаки: `ema_ratio_0.5_0.1`, `ema_ratio_0.7_0.3`
3. Слабо информативные статистики: `acf_lag_7`, `coeff_var_10`, `shannon_entropy_10`
4. Избыточные лаговые переменные: `lag_7`, `lag_16`, `lag_10`

Наиболее устойчивые признаки (сохранены в финальном наборе):

1. Ключевые лаги (`lag_4`, `lag_14`, `lag_15`)
2. Экспоненциальные скользящие средние (`ema_0.1`, `ema_0.2`, `ema_0.3`)
3. Бинарные индикаторы экстремумов (`is_peak`, `is_trough`)
4. Технические индикаторы (`macd_signal`, `macd_hist`, `adx`)

Финальное значение R^2 составило 0.9565, что на 0.11 ниже результата forward selection.

4.3. Сравнительный анализ отобранных признаков

4.3.1. Пересечение наборов признаков

Оба метода выделили 6 общих наиболее значимых признаков:

- `lag_4` - значение ряда с лагом 4 дня
- `is_peak` - индикатор локального максимума
- `is_trough` - индикатор локального минимума
- `lag_14` - значение ряда с лагом 14 дней
- `month_2` - индикатор февраля
- `adx` - Average Directional Index (сила тренда)

4.3.2. Особенности каждого метода

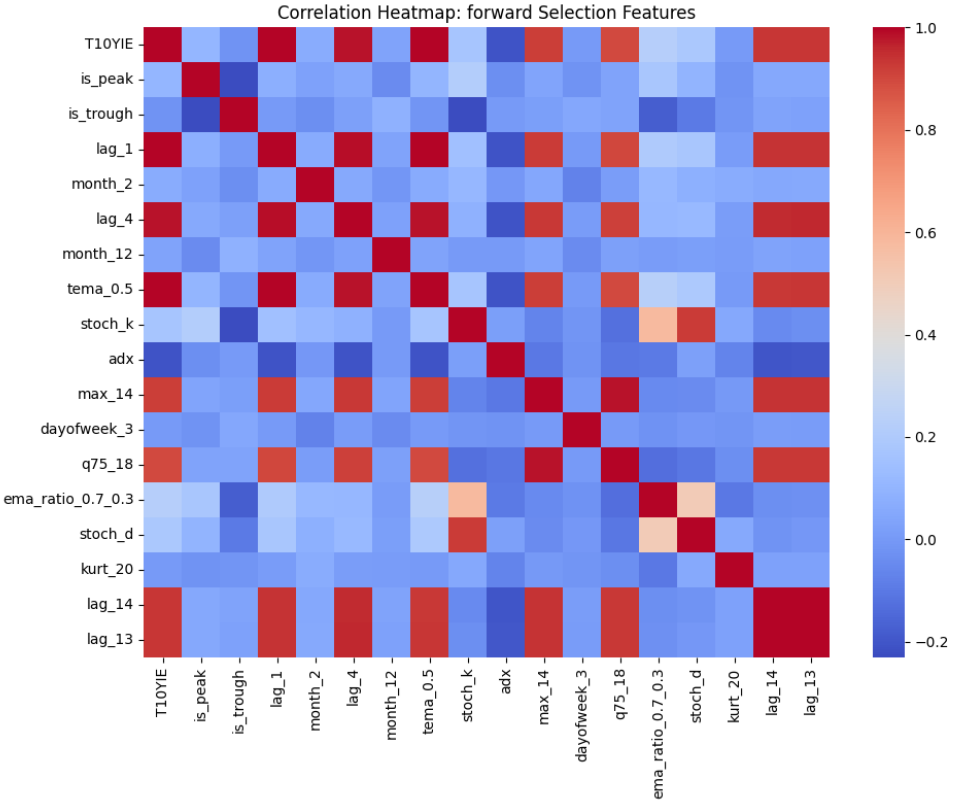

Рис. 4.3: Корреляционная матрица признаков Forward Selection

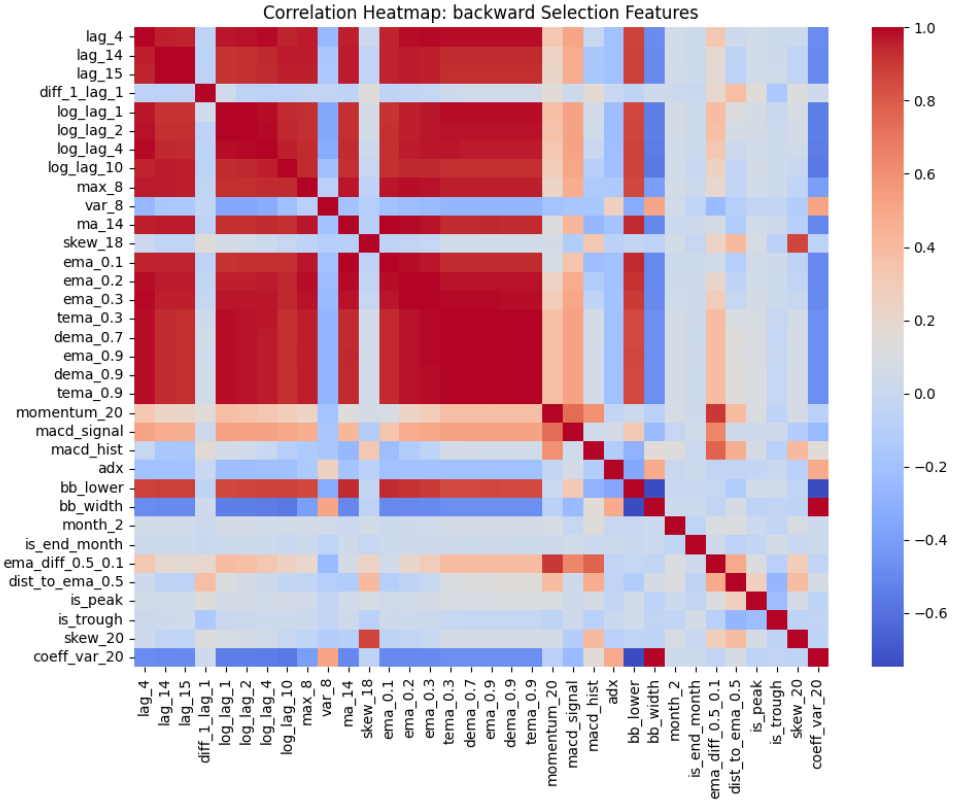

Рис. 4.4: Корреляционная матрица признаков Backward Selection

Forward Selection (18 признаков, $R^2 = 0.9576$):

- Сфокусирован на краткосрочных зависимостях (`lag_1`, `T10YIE`)
- Включил больше технических индикаторов (`stoch_k`, `stoch_d`, `tema_0.5`)
- Сохранил сезонные компоненты (`month_12`)
- Имеет высокую корреляцию между признаками (некоторые пары > 0.99)

Backward Selection (34 признака, $R^2 = 0.9565$):

- Включил более разнообразный набор экспоненциальных средних (`ema_0.1`, `ema_0.2`, `ema_0.3`, `tema_0.3`, `dema_0.7`)
- Сохранил логарифмические преобразования (`log_lag_1`, `log_lag_2`, `log_lag_4`)
- Добавил больше статистических метрик (`var_8`, `skew_18`, `skew_20`, `coeff_var_20`)
- Демонстрирует более сбалансированную корреляционную структуру

4.3.3. Сравнение эффективности

Таблица 4.1: Сравнительная характеристика методов отбора признаков

Параметр	Forward Selection	Backward Selection
Количество признаков	18	34
Final R^2	0.9576	0.9565
Улучшение от исходного	+2.30%	+2.19%
Время вычисления	Быстрее	Медленнее
Интерпретируемость	Высокая	Средняя
Корреляция между признаками	Высокая	Умеренная
Разнообразие признаков	Среднее	Высокое

4.3.4. Выводы

- **Forward selection** показал лучшее качество при меньшем количестве признаков
- **Backward selection** обеспечил более разнообразный и сбалансированный набор признаков
- **Общее ядро** из 6 признаков подтверждает их критическую важность для прогнозирования
- Для финального моделирования будет использован **forward selection** набор как более компактный и интерпретируемый

Глава 5

Моделирование

5.1. Описание архитектур и гиперпараметров моделей

Для решения задачи прогнозирования временного ряда T10YIE были реализованы три модели: линейная регрессия, модель градиентного бустинга (XGBoost) и рекуррентная нейронная сеть (LSTM). Данные были разделены на обучающую, валидационную и тестовую выборки с сохранением временного порядка:

- Train: 2005-05-24 – 2017-05-19 (≈70% данных)
- Validation: 2017-05-22 – 2021-05-20 (для настройки гиперпараметров)
- Test: 2021-05-21 – 2025-08-26 (для финальной оценки моделей)

Для настройки гиперпараметров и предотвращения переобучения использовался метод временных скользящих окон (Time Series Cross-Validation) с 5 фолдами.

5.2. Реализация и анализ линейной регрессии

5.2.1. Архитектура и принцип работы

В качестве алгоритма была выбрана линейная регрессия с L2-регуляризацией. Формально модель описывается уравнением:

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \tag{5.1}$$

где оптимизация весов происходит с учетом регуляризационного члена:

$$\min_w \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{m} w_j^2 \right) \tag{5.2}$$

5.2.2. Подбор гиперпараметров и обучение

Для настройки силы регуляризации α был рассмотрен следующие значения:

```
alpha_grid = (0.01, 0.1, 1, 10, 100, 1000)
```

- Лучший параметр регуляризации: $\alpha = 0.01$ в результате кросс-валидации
- Качество модели на кросс-валидации: $R^2 = 0.9367$, MSE = 0.0033

5.2.3. Анализ коэффициентов модели

Таблица 5.1: Коэффициенты линейной регрессионной модели

Признак	Коэффициент	Абсолютная важность
T10YIE	1.2131	1.0000
lag_1	-0.2559	0.2109
max_14	0.0781	0.0644
tema_0.5	0.0779	0.0642
is_trough	0.0462	0.0381
lag_13	0.0406	0.0335
month_12	0.0354	0.0292
is_peak	-0.0382	0.0315
lag_4	-0.0635	0.0523
month_2	0.0188	0.0155
q75_18	-0.0528	0.0435
ema_ratio_0.7_0.3	-0.0533	0.0439
lag_14	-0.0513	0.0423
dayofweek_3	-0.0079	0.0065
stoch_k	0.0007	0.0006
adx	-0.0005	0.0004
stoch_d	-0.0005	0.0004
kurt_20	-0.0013	0.0011

Анализ коэффициентов показывает:

- **Текущее значение ряда** (`T10YIE`) имеет наибольший положительный вклад в прогноз
- **Лаговые переменные** демонстрируют сложную структуру влияния с разными знаками
- **Технические индикаторы** (`fema`, `stoch`, `adx`) имеют относительно небольшое влияние
- **Сезонные факторы** (`month`) показывают ожидаемую периодичность влияния

5.2.4. Визуализация результатов

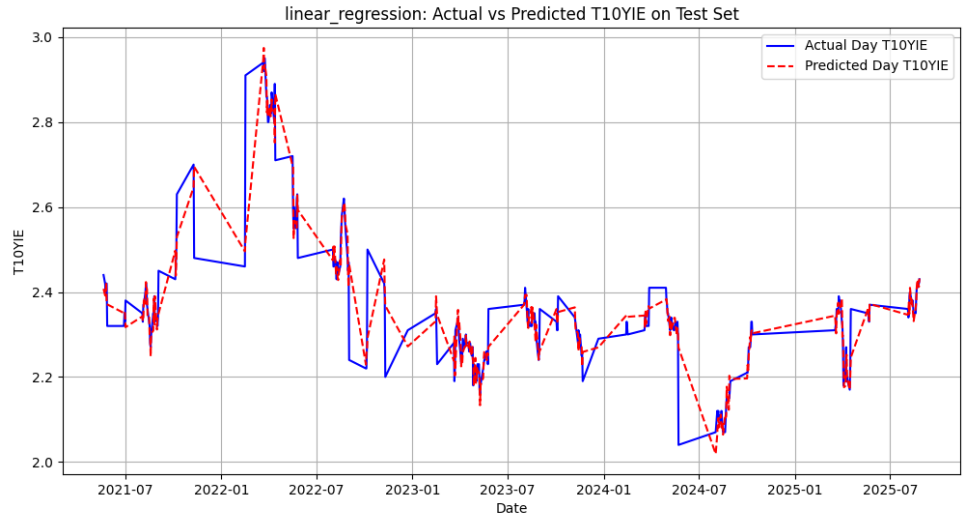

Рис. 5.1: Сравнение фактических и предсказанных значений T10YIE

График демонстрирует хорошее соответствие между фактическими и предсказанными значениями. Модель успешно отслеживает как долгосрочные тренды, так и краткосрочные колебания инфляционных ожиданий.

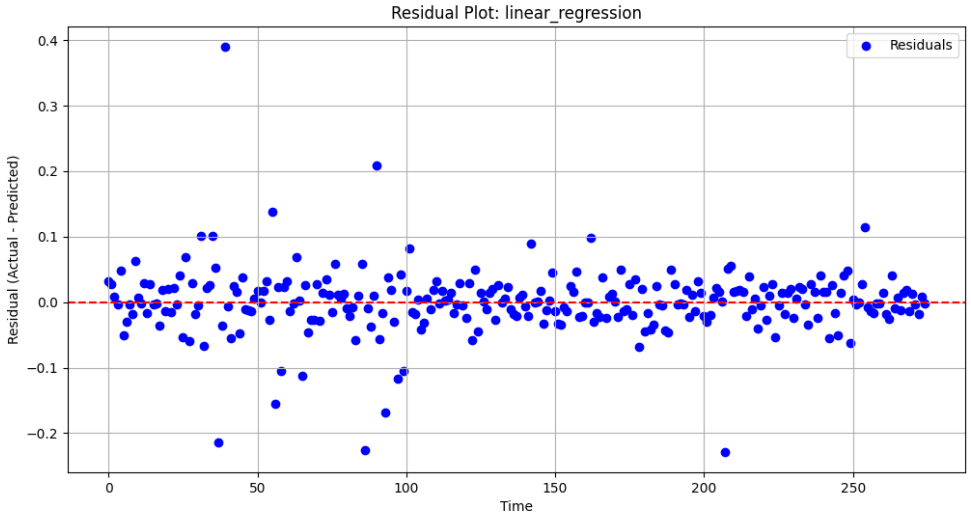

Рис. 5.2: Распределение остатков модели по времени

Анализ остатков показывает их случайный характер распределения вокруг нуля, что свидетельствует об отсутствии систематической ошибки в модели.

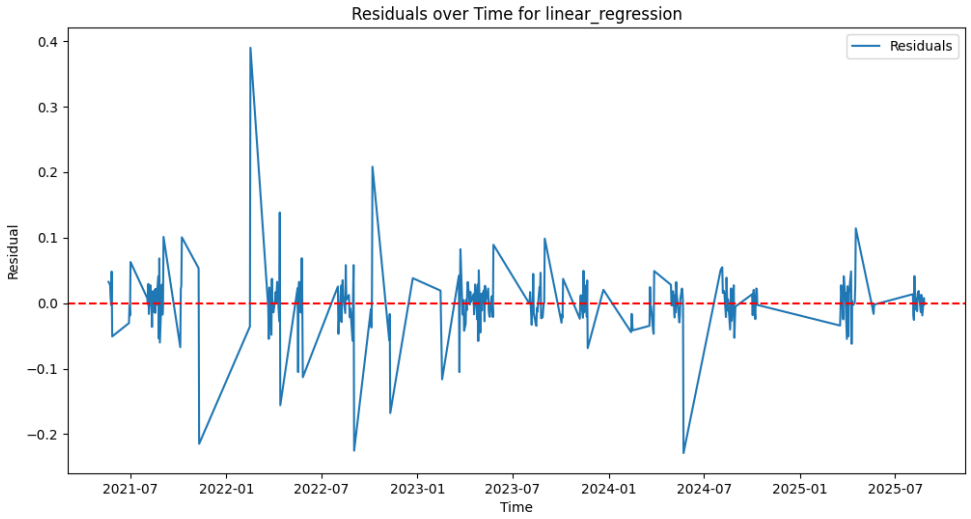

Рис. 5.3: Временная динамика остатков модели

Остатки равномерно распределены во времени без видимых паттернов, что подтверждает адекватность модели. Отдельные выбросы соответствуют периодам высокой волатильности рынка.

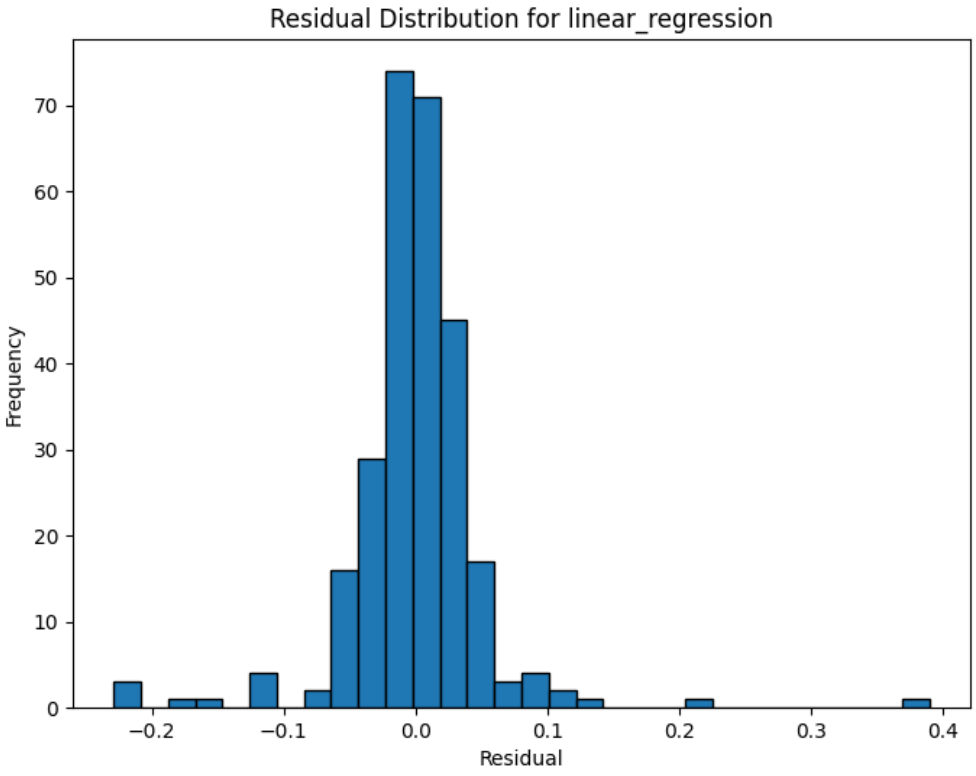

Рис. 5.4: Гистограмма распределения остатков

Распределение остатков близко к нормальному с средним равным нулю.

5.2.5. Анализ ошибок прогнозирования

Анализ периодов с наибольшими ошибками выявил характерные закономерности:

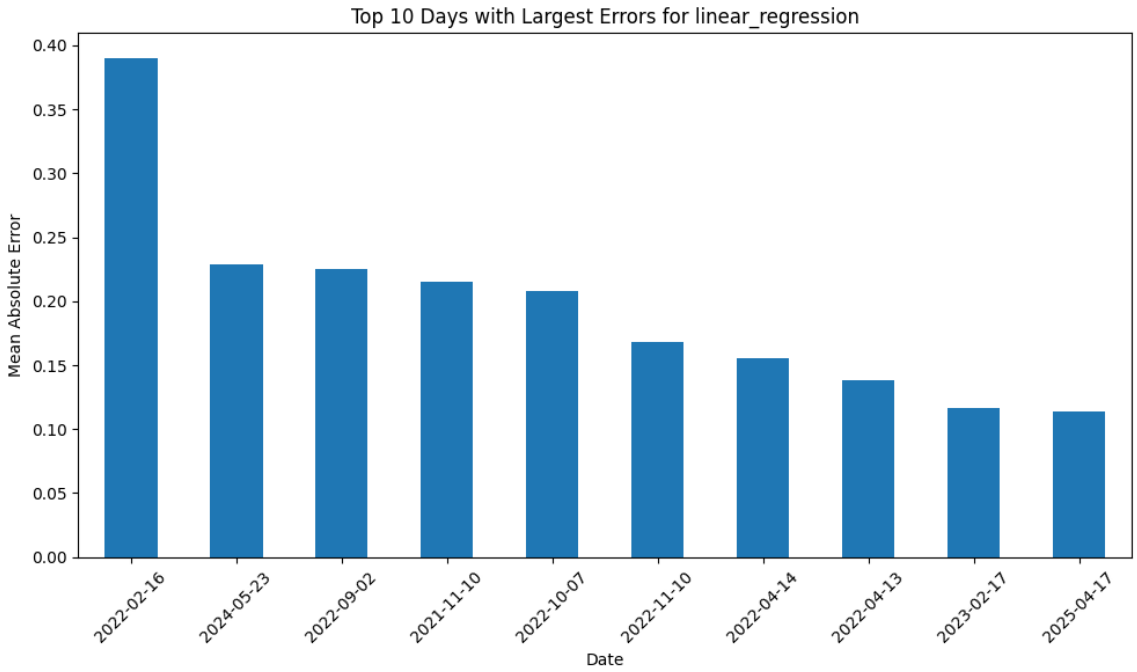

Рис. 5.5: Периоды с наибольшими ошибками прогнозирования

Наибольшие ошибки приходятся на 2022 год.

5.2.6. Выводы

Линейная регрессия с L2-регуляризацией показала хорошие результаты прогнозирования временного ряда T10YIE:

- **Высокая объясняющая способность**: $R^2 = 0.9140$ на тестовой выборке
- **Точность прогнозов**: MAE = 0.0307, MAPE = 1.29%
- **Интерпретируемость**: четкая структура влияния признаков

Основные ограничения модели:

- Неспособность полностью справиться с экстремальными движения во время кризисных периодов
- Линейная природа модели ограничивает учет сложных нелинейных зависимостей

5.3. Реализация и анализ XGBoost

XGBoost — ансамблевая модель, основанная на градиентном бустинге деревьев решений, где каждое последующее дерево корректирует ошибки предыдущих.

5.3.1. Подбор гиперпараметров и обучение

Для настройки модели были рассмотрены следующие гиперпараметры:

- **n_estimators**: Количество деревьев (100, 300, 500, 1000)
- **max_depth**: Максимальная глубина дерева (3, 5, 7, 10)
- **learning_rate**: Скорость обучения (0.01, 0.05, 0.1, 0.2)
- **subsample**: Доля наблюдений для обучения (0.8, 0.9, 1.0)

В результате оптимизации была определена следующая оптимальная конфигурация:

- **learning_rate**: 0.01 - обеспечивает постепенное, устойчивое обучение модели
- **max_depth**: 3 - создает достаточно простые деревья, что предотвращает переобучение и улучшает интерпретируемость
- **n_estimators**: 500 - количество деревьев
- **subsample**: 0.8 - использование 80% данных для каждого дерева

Обучение модели проводилось с использованием ранней остановки (early_stopping_rounds=50) на валидационной выборке для предотвращения переобучения. Модель демонстрирует умеренную прогнозирущую способность, объясняя 76.37% дисперсии целевой переменной на тестовой выборке.

5.3.2. Анализ важности признаков

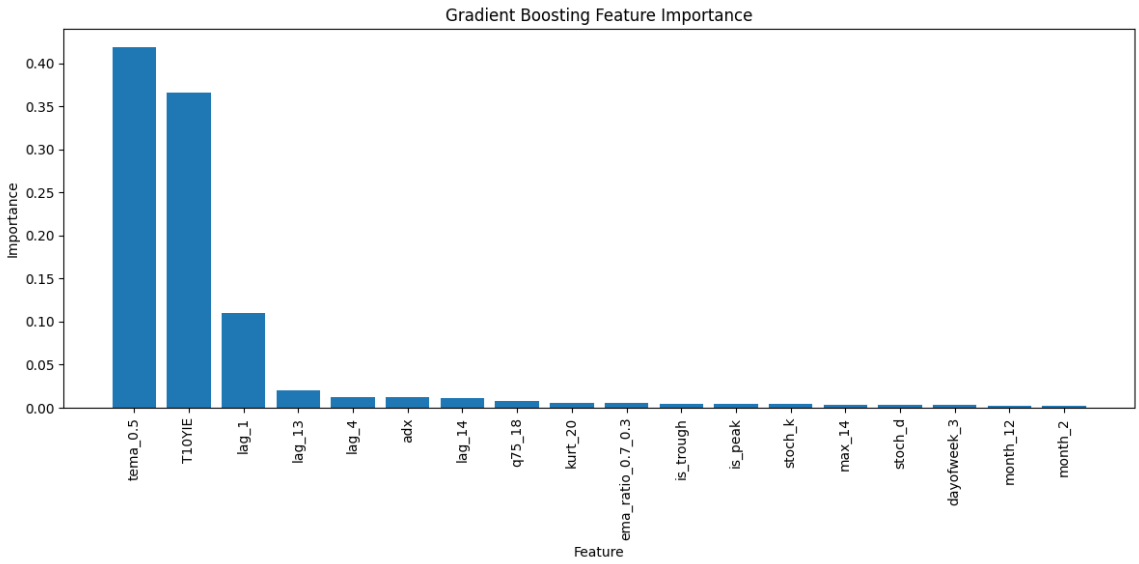

Рис. 5.6: Важность признаков в Gradient Boosting модели

Анализ важности признаков показывает, что модель в основном полагается на:

- **Статистические характеристики**: Наиболее важными признаками стали трендовые индикаторы (tema_0.5), квантили (q75_18) и эксцесс (kurt_20)
- **Текущее и лаговые значения**: Текущее значение ряда и лаговые переменные сохранили умеренную значимость

5.3.3. Визуализация результатов

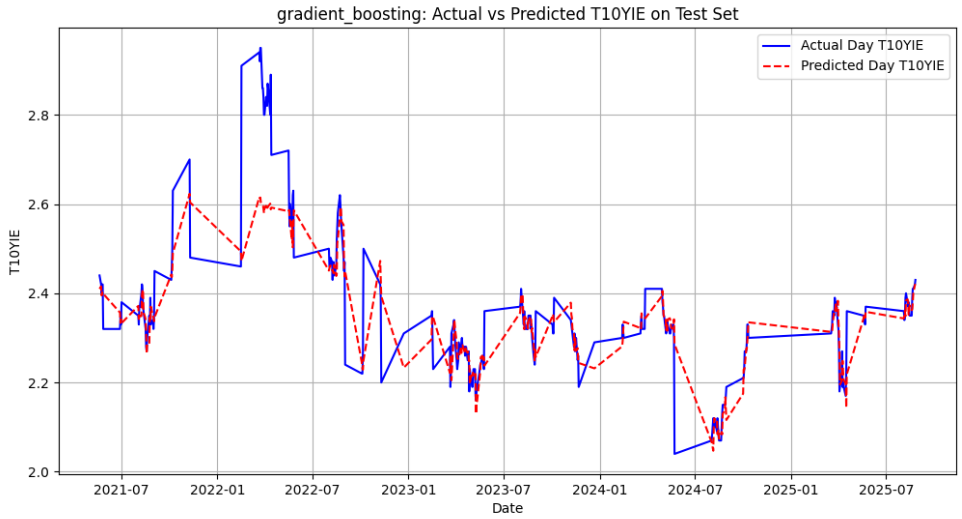

Рис. 5.7: Сравнение фактических и предсказанных значений (Gradient Boosting)

График показывает, что модель успешно отслеживает общие тенденции временного ряда, но может испытывать трудности с точным прогнозированием экстремальных значений и резких изменений.

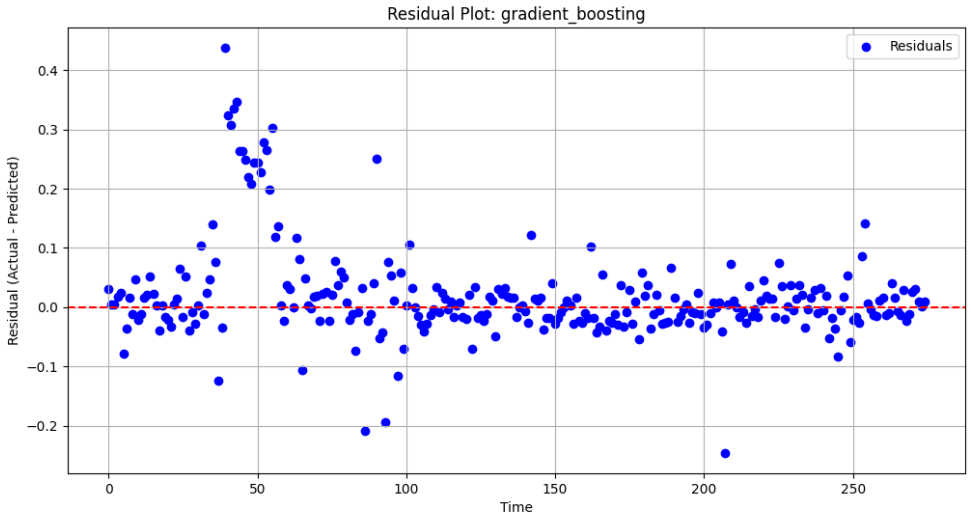

Рис. 5.8: Распределение остатков Gradient Boosting модели

Остатки демонстрируют случайное распределение вокруг нуля, хотя наблюдается несколько выбросов, соответствующих периодам высокой волатильности.

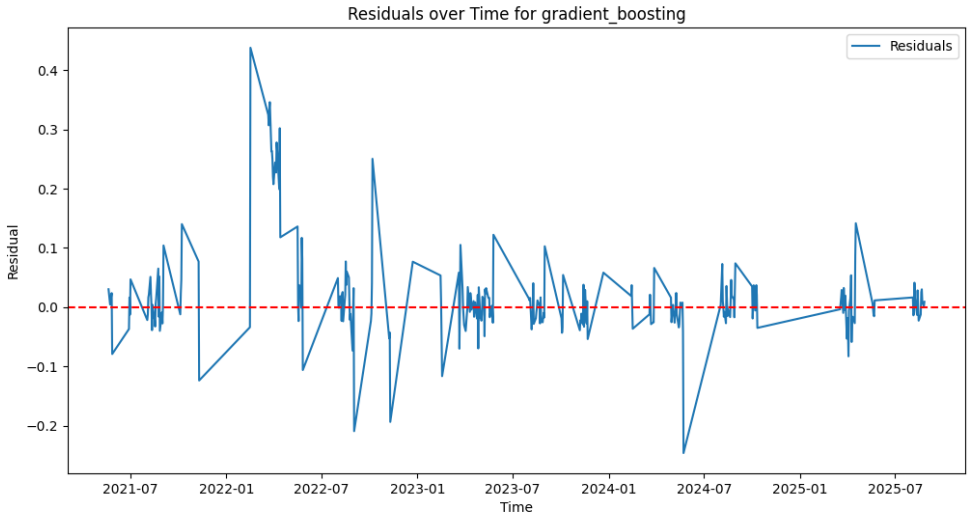

Рис. 5.9: Временная динамика остатков Gradient Boosting

Анализ остатков во времени показывает их равномерное распределение без выраженных паттернов, что свидетельствует об адекватности модели.

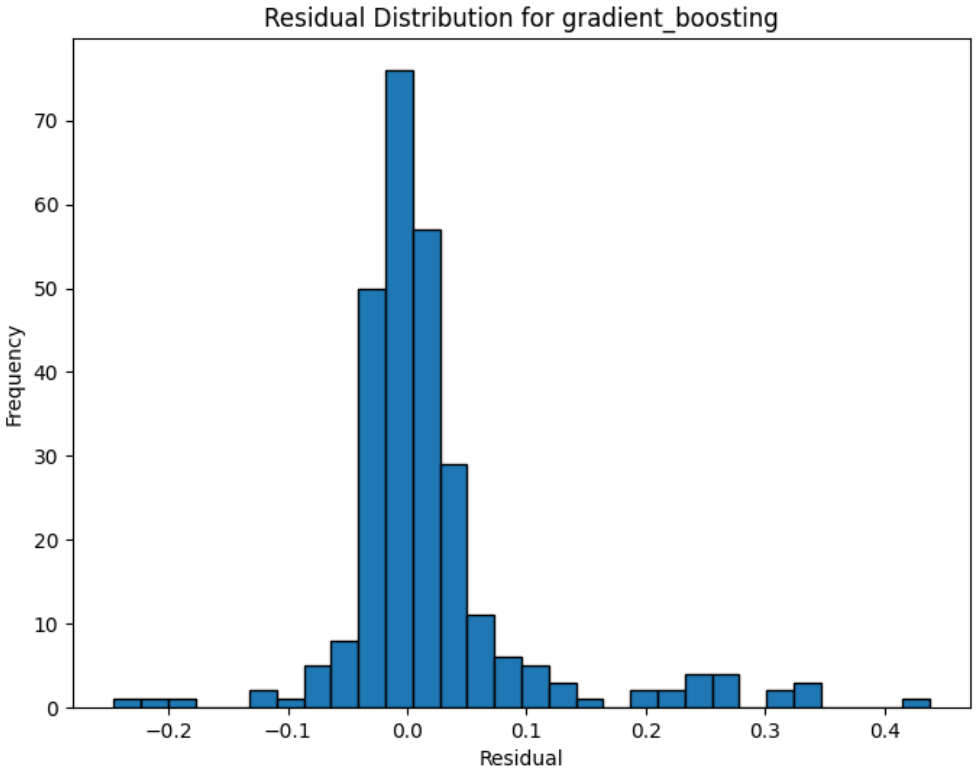

Рис. 5.10: Гистограмма распределения остатков Gradient Boosting

Распределение остатков близко к нормальному с легкой асимметрией, обусловленной наличием выбросов.

5.3.4. Анализ ошибок прогнозирования

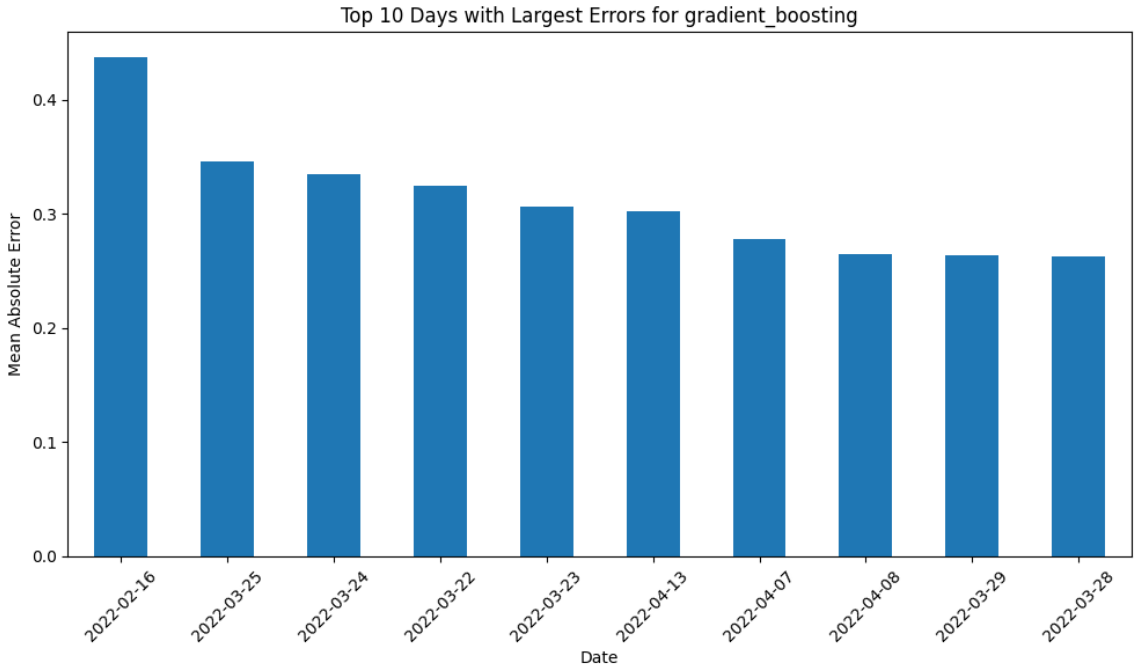

Рис. 5.11: Периоды с наибольшими ошибками прогнозирования

Наибольшие ошибки также сконцентрированы в первом квартале 2022 года.

5.3.5. Выводы

Модель Gradient Boosting (XGBoost) продемонстрировала умеренные результаты прогнозирования временного ряда T10YIE:

- **Умеренная объясняющая способность**: $R^2 = 0.7637$ на тестовой выборке
- **Приемлемая точность прогнозов**: MAE = 0.0470, MAPE = 1.87%
- **Способность к нелинейным зависимостям**: Может обнаруживать сложные нелинейные взаимосвязи
- **Неустойчивость к выбросам**: Хуже стабильность в экстремальных условиях по сравнению с линейной моделью

5.4. Реализация и анализ LSTM модели

5.4.1. Архитектура и принцип работы

Модель Long Short-Term Memory (LSTM) — тип рекуррентной нейронной сети, разработанный для работы с последовательными данными и долгосрочными зависимостями. Архитектура LSTM включает три типа gate, которые управляют потоком информации:

- Входной gate: Определяет, какая новая информация будет сохранена в состоянии ячейки
- Забывающий gate: Решает, какую информацию удалить из состояния ячейки
- Выходной gate: Определяет, какая информация будет передана на выход

5.4.2. Подбор гиперпараметров и обучение

Для настройки LSTM модели использовалась Optuna с 15 итерациями. Были оптимизированы следующие гиперпараметры:

- **sequence_length**: Длина временного окна (10, 20, 30 дней)
- **batch_size**: Размер батча (16, 32)
- **hidden_size**: Размер скрытого состояния (50, 100 нейронов)
- **dropout**: Уровень dropout (0.2, 0.3)
- **learning_rate**: Скорость обучения (0.001, 0.01)

5.4.3. Результаты обучения

После оптимизации была получена следующая конфигурация:

- **sequence_length**: 10 дней
- **batch_size**: 32
- **hidden_size**: 50 нейронов
- **dropout**: 0.3
- **learning_rate**: 0.001

Относительно хорошие результаты на валидации ($R^2 = 0.8374$).

5.4.4. Анализ кривых обучения

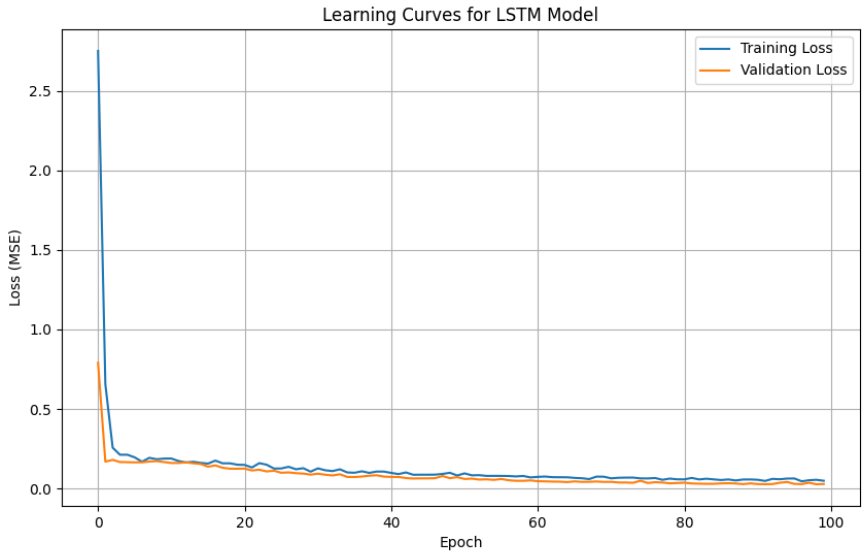

Рис. 5.12: Кривые обучения LSTM модели

Анализ кривых обучения показывает:

- Быструю сходимость на первых эпохах
- Стабильное снижение потерь на тренировочных данных

5.4.5. Визуализация результатов

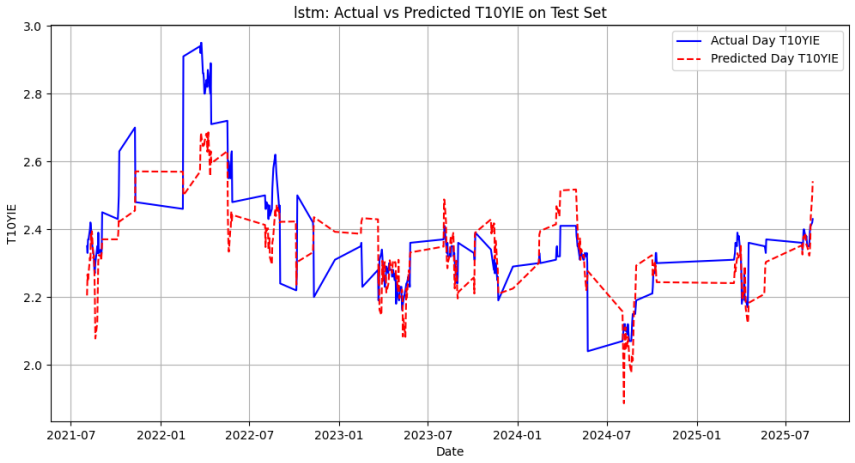

Рис. 5.13: Сравнение фактических и предсказанных значений (LSTM)

График предсказаний показывает, что модель не справляется с точным прогнозированием, особенно в периоды высокой волатильности.

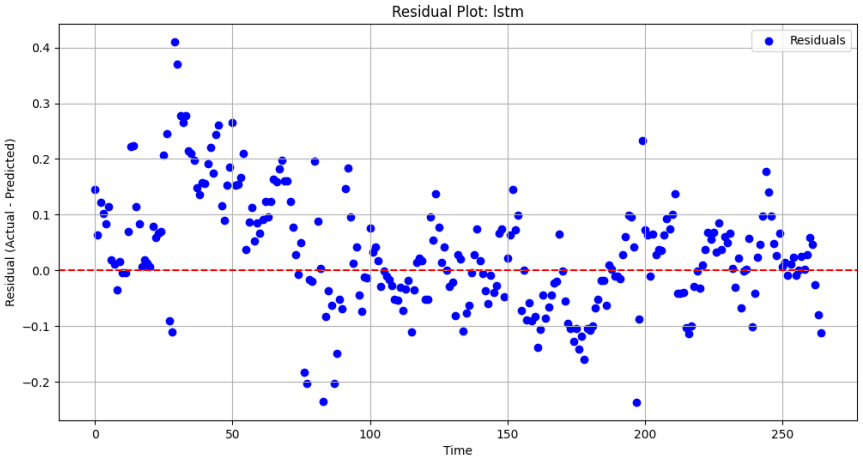

Рис. 5.14: Распределение остатков LSTM модели

Остатки демонстрируют значительную вариативность и систематические ошибки.

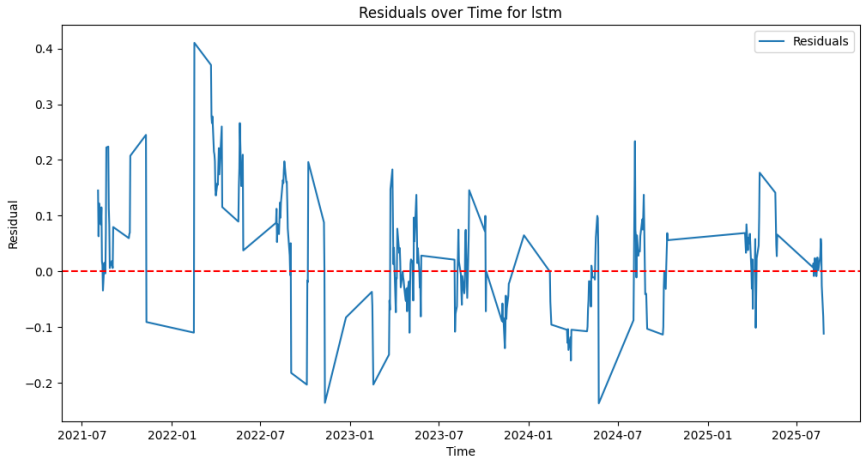

Рис. 5.15: Временная динамика остатков LSTM

Остатки показывают кластеризацию ошибок во времени, особенно в периоды кризиса рынка.

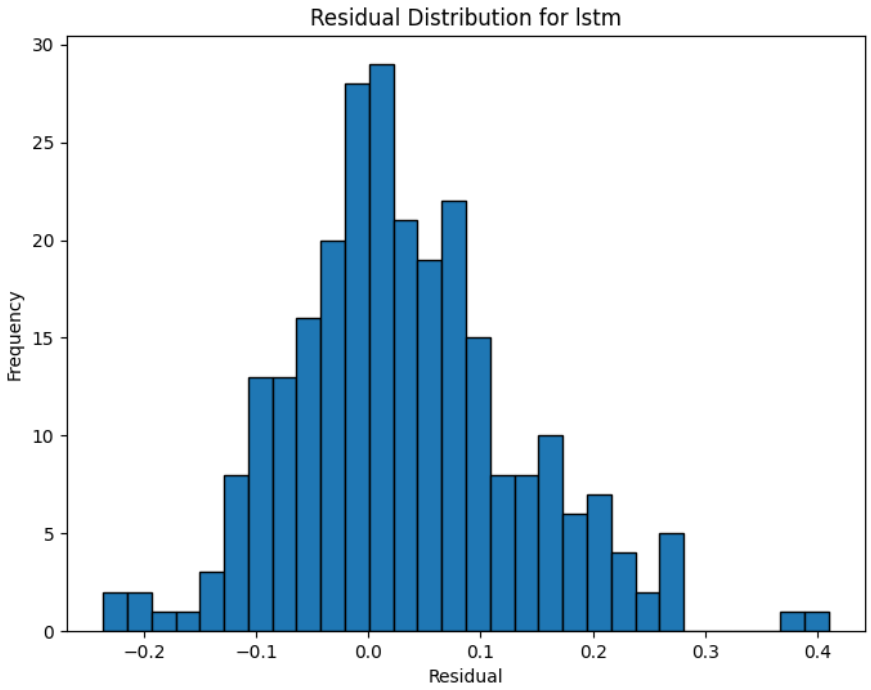

Рис. 5.16: Гистограмма распределения остатков LSTM

Распределение остатков близко к нормальному, но имеет асимметрию.

5.4.6. Анализ ошибок прогнозирования

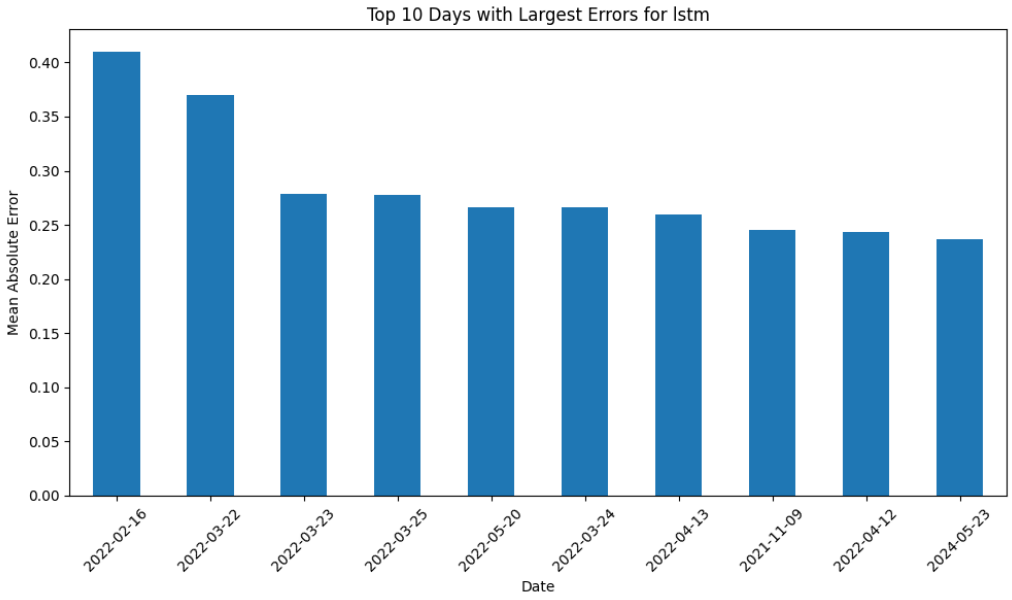

Рис. 5.17: Периоды с наибольшими ошибками прогнозирования

Наибольшие ошибки сконцентрированы в период конца 2021 - 2022.

5.4.7. Выводы

- **Низкая эффективность**: LSTM ($R^2 = 0.6332$ на тестовой выборке) показала худшие результаты по сравнению с линейной регрессией и XGBoost
- **Чувствительность к экстремальным событиям**: Наибольшие ошибки в периоды кризисов
- **Переобучение**: Значительный разрыв между validation(0.8024) и test(0.6332) результатами (для предотвращения использовались: dropout (0.3), ранняя остановка, регуляризация, что оказалось недостаточным)

Основные проблемы:

- Высокая волатильность временного ряда
- Ограниченная способность обобщать на тестовых данных
- Фокус на долгосрочных временных зависимостях не приносят дополнительной пользы
- Недостаточное количество данных
- Излишняя сложность архитектуры

5.5. Сравнительный анализ моделей

5.5.1. Обзор производительности моделей

Таблица 5.2: Сравнительные метрики качества моделей

Метрика	Линейная регрессия	XGBoost	LSTM
R^2	**0.9140**	0.7637	0.6332
MSE	**0.0026**	0.0072	0.0115
MAE	**0.0307**	0.0470	0.0810
RMSE	**0.0510**	0.0846	0.1072
MAPE	**1.29%**	1.87%	3.34%
SMAPE	**1.29%**	1.90%	3.38%

- Линейная регрессия показала наилучшие результаты по всем метрикам
- Высокий $R^2 = 0.9140$ свидетельствует о преимущественно линейной природе зависимостей в данных
- Преобладают мгновенные корреляции над временными тенденциями
- XGBoost и LSTM показали худшие результаты, несмотря на большую сложность
- Обе модели демонстрируют склонность к переобучению

5.5.2. Анализ устойчивости к экстремальным событиям

Таблица 5.3: Сравнение ошибок в кризисные периоды (2021 - 2022)

Модель	Средняя ошибка	Максимальная ошибка
Линейная регрессия	0.152	0.390
XGBoost	0.189	0.438
LSTM	0.243	0.410

Наблюдения:

- Все модели ошибаются с прогнозированием в периоды кризисов
- Линейная регрессия показывает наибольшую устойчивость
- Сложные модели более чувствительны к выбросам

5.5.3. Интерпретируемость

- **Линейная регрессия**: Полная интерпретируемость, ясные коэффициенты
- **XGBoost**: Ограниченная интерпретируемость через feature importance
- **LSTM**: "Чёрный ящик", крайне ограниченная интерпретируемость

5.5.4. Зависимость от признаков

- Линейная регрессия: 18 признаков → оптимальный результат
- XGBoost: 18 признаков → умеренное качество (возможно, недостаточно данных)
- LSTM: 18 признаков → переобучение (требует больше данных или меньше признаков)

5.5.5. Общий вывод

Наиболее простая модель (линейная регрессия) показала наилучшие результаты для данного временного ряда. Это свидетельствует о том, что:

1. Основные зависимости в данных T10YIE имеют линейную природу
2. Сложные нелинейные и временные зависимости играют второстепенную роль
3. Регуляризация эффективнее справляется с переобучением, чем сложные архитектуры

Глава 6

Заключение

6.1. Ключевые выводы

1. **Превосходство простой модели**: Линейная регрессия с L2-регуляризацией показала наилучшие результаты ($R^2 = 0.9140$, MAPE = 1.29%), превзойдя более сложные методы машинного обучения, включая XGBoost ($R^2 = 0.7637$) и LSTM ($R^2 = 0.6332$). Это свидетельствует о преимущественно линейной природе исследуемых зависимостей.

2. **Критическая важность feature engineering**: Статистические признаки (квантили, скользящие статистики) и трендовые индикаторы оказались информативными.

3. **Эффективность методов отбора признаков**: Forward Selection продемонстрировал оптимальное соотношение производительности и эффективности, отобрав 18 наиболее значимых признаков с $R^2 = 0.9576$.

4. **Ограничения сложных моделей**: Нейросетевые подходы (LSTM) показали склонность к переобучению и требовали значительных вычислительных ресурсов, не обеспечивая при этом существенного улучшения качества.

5. **Уязвимость к экстремальным событиям**: Все модели демонстрировали наибольшие ошибки в периоды кризисов (2022).

6. **Значимость временных паттернов**: Сезонные и циклические компоненты играют важную роль в прогнозировании, однако краткосрочные временные зависимости оказались наиболее важными.

6.2. Направления для будущих исследований

1. **Расширенный feature engineering**:
    - Разработка специализированных финансовых признаков
    - Использование трансферного обучения с включением данных из смежных областей
2. **Учет экстремальных событий**:
    - Разработка методов обнаружения и обработки выбросов
    - Создание адаптивных моделей, способных быстро приспосабливаться к изменяющимся рыночным условиям
3. **Масштабирование и production**:

- Оптимизация моделей для работы в реальном времени
- Создание инфраструктуры для непрерывного обучения и адаптации моделей

6.3. Ссылки

6.3.1. Исходный код

- **Исходный код**: `https://github.com/ArinaArtiukevich/inflation_prediction`
- **Данные**: Федеральный резервный банк Сент-Луиса (FRED)
  - **Ряд**: T10YIE -- 10-Year Breakeven Inflation Rate
  - **Период**: 2 января 2003 года -- 27 августа 2025 года
  - **Ссылка**: `https://fred.stlouisfed.org/series/T10YIE`