МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Тольяттинский государственный университет»

Кафедра «Прикладная математика и информатика»
(наименование кафедры полностью)

01.03.02 Прикладная математика и информатика
(код и наименование направления подготовки, специальности)

Компьютерные технологии и математическое моделирование
(направленность(профиль))

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (БАКАЛАВРСКАЯ РАБОТА)

на тему Исследование и реализация гибридного алгоритма прогнозирования временных рядов

Обучающийся А.С. Логинов
(Инициалы Фамилия) (личная подпись)

Руководитель д.т.н., доцент, С.В. Мкртычев
(ученая степень(при наличии), ученое звание (при наличии), Инициалы Фамилия)

Консультант к.т.н., доцент, М.В. Дайненко
(ученая степень(при наличии), ученое звание (при наличии), Инициалы Фамилия)

Тольятти 2025

Аннотация

Тема выпускной квалификационной работы – «Исследование и реализация гибридного алгоритма прогнозирования временных рядов».

Исследование и разработка методов прогнозирования временных рядов на основе комбинированных подходов представляет научную и практическую значимость.

Цель работы – разработка, реализация и тестирование гибридного алгоритма прогнозирования временных рядов, объединяющего статистические методы и машинное обучение.

Объектом исследования являются временные ряды из наборов данных NN3 и M4, содержащие сезонность, тренды и случайные компоненты.

Методы исследования включают STL-декомпозицию, преобразование Бокса–Кокса, методы экспоненциального сглаживания, алгоритмы машинного обучения (XGBoost, SVR) и оценку точности прогнозов (sMAPE, MASE).

Предметом исследования является гибридный алгоритм прогнозирования, основанный на STL-декомпозиции, экспоненциальном сглаживании и применении XGBoost и SVR для остаточной компоненты.

Практическая значимость работы заключается в создании программной реализации алгоритма, демонстрирующего высокую точность прогнозирования и конкурентоспособность по сравнению с традиционными методами, такими как AutoARIMA и Holt-Winters.

Работа состоит из введения, четырёх глав, заключения, приложения и списка литературы.

Бакалаврская работа содержит 48 страниц текста, 8 рисунков и 25 источников. В приложении представлены фрагменты кода.

Abstract

The title of the graduation work is «Development and Implementation of a Hybrid Time Series Forecasting Algorithm»

The study and development of forecasting methods based on hybrid approaches to time series analysis are of both scientific and practical relevance.

The aim of this work is to develop, implement, and evaluate a hybrid time series forecasting algorithm that combines statistical methods and machine learning.

The object of the research is time series from the NN3 and M4 datasets, which include seasonality, trends, and random components.

The research methods include STL decomposition, Box–Cox transformation, exponential smoothing techniques, machine learning algorithms (XGBoost, SVR), and forecast accuracy evaluation using sMAPE and MASE metrics.

The subject of the research is a hybrid forecasting algorithm based on STL decomposition, exponential smoothing, and the application of XGBoost and SVR to the residual component.

The practical significance of the work lies in the development of a software implementation of the algorithm, which demonstrates high forecasting accuracy and competitiveness compared to traditional methods such as AutoARIMA and Holt-Winters.

The work consists of an introduction, four chapters, a conclusion, an appendix, and a list of references.

The bachelor's thesis contains 48 pages of text, 8 figures, and 25 sources. The appendix includes fragments of code.

Оглавление

Введение 4
Глава 1 Постановка задачи исследования алгоритма 6
Глава 2 Обзор методов прогнозирования временных рядов 11
Глава 3 Методология исследования 23
Глава 4 Экспериментальная оценка 33
Заключение 39
Список используемой литературы 43
Приложение А Листинг (реализация программы) 46

Введение

Настоящая работа посвящена разработке и исследованию гибридного подхода к прогнозированию временных рядов с использованием декомпозиции данных, статистических методов и алгоритмов машинного обучения.

Предложенный метод включает разделение временного ряда на тренд, сезонную компоненту и остатки с помощью метода STL, прогнозирование каждой компоненты отдельно и последующее объединение результатов.

Для тренда и сезонности применяются методы экспоненциального сглаживания (Holt и Holt-Winters), а для остатков – алгоритмы градиентного бустинга (XGBoost) и машины опорных векторов (SVR) с предварительным извлечением признаков.

Тестирование проводится на наборах данных NN3 и M4, содержащих месячные временные ряды с выраженной сезонностью и трендами.

Оценка качества прогнозов выполняется с использованием метрик sMAPE и MASE.

Объектом исследования являются временные ряды, представленные наборами данных NN3 и M4. Предмет исследования – гибридный алгоритм прогнозирования, объединяющий декомпозицию, статистические методы и машинное обучение.

Цель работы – разработка, реализация и тестирование гибридного алгоритма прогнозирования временных рядов, а также сравнение его эффективности с традиционными методами.

Для достижения цели необходимо решить следующие задачи:

- сформулировать постановку задачи исследования и проанализировать методы прогнозирования временных рядов;
- изучить и проанализировать алгоритмы прогнозирования, включая статистические и комбинированные подходы;
- разработать и протестировать программу, реализующую

предложенный алгоритм.

Методы исследования включают декомпозицию временных рядов, экспоненциальное сглаживание, алгоритмы машинного обучения, а также оценку качества с использованием метрик sMAPE и MASE. Реализация выполняется с применением языков программирования высокого уровня.

Практическая значимость работы заключается в создании программного обеспечения, которое позволяет эффективно прогнозировать временные ряды с учётом их сложной структуры, что может быть использовано в экономике, логистике и других областях.

Работа состоит из введения, трёх глав, заключения и списка литературы.

Первая глава посвящена постановке задачи и анализу методов прогнозирования.

Вторая глава включает обзор и анализ алгоритмов.

Третья глава описывает программную реализацию и тестирование алгоритма на наборах данных NN3 и M4.

В заключении представлены основные результаты исследования.

Бакалаврская работа содержит 48 страниц текста, 8 рисунков и 25 источников. В приложении представлены фрагменты кода.

Глава 1 Постановка задачи исследования алгоритма

Прогнозирование временных рядов представляет собой одну из центральных задач в области анализа данных, находящую широкое применение в различных сферах человеческой деятельности, таких как экономика, производственные процессы, метеорология, энергетика, информационные технологии, а также социальные исследования. Данный подход позволяет выявлять закономерности в данных и использовать их для предсказания будущих значений, что крайне важно для планирования и принятия стратегических решений. Например, в экономике прогнозирование временных рядов помогает предсказывать спрос на товары, в энергетике – определять потребление электроэнергии, а в метеорологии – прогнозировать погодные условия, что способствует повышению безопасности и эффективности управления ресурсами.

Временной ряд – это последовательность числовых значений, зафиксированных в определённые моменты времени с фиксированным интервалом между наблюдениями.

Типичные примеры включают ежемесячные показатели продаж в розничной торговле, суточные метеорологические наблюдения (температура, осадки), биржевые котировки акций, а также данные о трафике в телекоммуникационных сетях. Основной целью прогнозирования временных рядов является предсказание будущих значений на основе анализа исторических закономерностей, что позволяет принимать обоснованные решения, оптимизировать использование ресурсов и минимизировать риски. Например, точный прогноз спроса на продукцию позволяет компаниям эффективно управлять запасами, а прогноз погоды помогает планировать сельскохозяйственные работы.

Временные ряды отличаются сложной внутренней структурой, которая включает несколько ключевых компонентов, определяющих их поведение:

– тренд: долгосрочное изменение уровня данных, отражающее устойчивый рост, снижение или стабильность. Например, рост продаж компании на протяжении нескольких лет может быть обусловлен расширением рынка;
– сезонность: периодические колебания, повторяющиеся через равные интервалы времени, такие как ежегодные изменения спроса на сезонные товары (например, увеличение продаж мороженого летом);
– случайные компоненты: непредсказуемые отклонения, которые не связаны с трендом или сезонностью и могут быть вызваны случайными событиями, такими как экономические кризисы или природные катаклизмы.

Анализ временных рядов направлен на решение ряда задач: описание динамики данных для понимания их поведения, объяснение причин изменений с целью выявления влияющих факторов, предсказание будущих значений для планирования, а также управление процессами на основе полученных прогнозов. В рамках данной работы основное внимание уделяется прогнозированию, которое рассматривается как процесс построения математических моделей, способных предсказывать значения временного ряда на заданном временном горизонте с высокой точностью. Это позволяет решать практические задачи, такие как планирование производства или управление энергопотреблением.

Существующие методы прогнозирования временных рядов можно разделить на две основные категории: статистические и основанные на машинном обучении. Статистические методы, такие как авторегрессионные модели интегрированного скользящего среднего (ARIMA) и экспоненциальное сглаживание (включая модели Holt и Holt-Winters), демонстрируют высокую эффективность при обработке временных рядов с выраженными линейными трендами и сезонными колебаниями. Эти методы опираются на предположение о стационарности или линейности данных, что делает их простыми и интерпретируемыми. Однако их возможности

ограничены при наличии сложных нелинейных зависимостей или нерегулярных изменений, таких как резкие скачки или аномалии. Методы машинного обучения, включая нейронные сети (например, рекуррентные нейронные сети), градиентный бустинг (XGBoost) и машины опорных векторов (SVR), обладают способностью моделировать сложные нелинейные закономерности. Тем не менее их применение к временным рядам сопряжено с рядом трудностей, таких как необходимость больших объёмов данных для обучения, сложность учёта временной структуры и высокая вычислительная нагрузка.

Для преодоления ограничений, присущих отдельным подходам, в данной работе предлагается гибридная модель прогнозирования, которая сочетает в себе преимущества декомпозиции временного ряда с использованием статистических методов и современных алгоритмов машинного обучения. Разработанный подход включает следующие этапы:

- разделение временного ряда на тренд, сезонную компоненту и остатки с применением метода STL (Seasonal-Trend decomposition using Loess), который позволяет изолировать основные компоненты для дальнейшего анализа;
- прогнозирование тренда с использованием метода Holt, который эффективно учитывает линейные изменения уровня данных и обеспечивает точное предсказание долгосрочных тенденций;
- прогнозирование сезонной компоненты с применением метода Winter, специально адаптированного для моделирования периодических колебаний, таких как сезонные изменения спроса;
- извлечение признаков из остатков с последующим их моделированием с использованием алгоритмов градиентного бустинга (XGBoost) и машин опорных векторов (SVR), что позволяет учитывать нелинейные зависимости;
- линейное объединение прогнозов всех компонентов для получения итогового предсказания, что обеспечивает баланс между точностью

и устойчивостью модели.

Целью данного исследования является разработка гибридной модели прогнозирования временных рядов, которая обеспечивает более высокую точность по сравнению с традиционными статистическими методами, такими как ARIMA, и отдельно применяемыми моделями машинного обучения.

Для достижения поставленной цели были сформулированы следующие задачи:

- проведение сравнительного анализа предложенного гибридного подхода с классическими методами, такими как AutoARIMA и Holt-Winters, для оценки их эффективности;
- оценка эффективности методов машинного обучения, таких как XGBoost и SVR, при их интеграции в гибридную модель, с акцентом на способность моделировать остатки;
- сравнение декомпозиционного подхода с другими комбинированными методами прогнозирования, чтобы выявить преимущества и недостатки предложенной методологии.

Для тестирования разработанной модели были выбраны два набора данных: NN3, содержащий 111 временных рядов, и подмножество из 48 000 рядов, представленных в рамках соревнования M4. Оба набора данных включают месячные наблюдения, характеризующиеся выраженной сезонностью, трендами и случайными компонентами, что делает их подходящими для всесторонней оценки производительности модели. Основной горизонт прогнозирования составляет 18 временных точек, что соответствует долгосрочному предсказанию и позволяет оценить устойчивость модели на длительных периодах. Качество прогнозов оценивается с использованием двух метрик:

- sMAPE (симметричная средняя абсолютная процентная ошибка), которая выражает относительную ошибку в процентах и позволяет сравнивать модели на разных наборах данных;
- MASE (средняя абсолютная шкалированная ошибка), которая

сравнивает ошибку модели с ошибкой наивного прогноза, что делает её удобной для оценки точности;

- OWA (общая взвешенная средняя), адаптированная из соревнования M4, которая объединяет sMAPE и MASE для комплексной оценки модели относительно наивного прогноза.

Выводы по главе 1

Значимость работы заключается в разработке масштабируемого подхода, который может быть адаптирован к различным типам временных рядов, включая данные с разной степенью сложности и динамикой. Предложенная гибридная модель демонстрирует высокий потенциал для применения в задачах прогнозирования, требующих повышенной точности на длительных горизонтах, таких как планирование производства, управление энергопотреблением или прогнозирование экономических показателей. Структура данной работы организована следующим образом: вторая глава посвящена детальному обзору существующих методов прогнозирования, третья глава описывает разработанную методологию с акцентом на её теоретическое обоснование, четвёртая глава содержит результаты экспериментов и их анализ, а пятая глава подводит итоги исследования и определяет перспективные направления для дальнейших исследований, таких как интеграция глубокого обучения или адаптация модели к потоковым данным.

Глава 2 Обзор методов прогнозирования временных рядов

Прогнозирование временных рядов представляет собой процесс предсказания будущих значений на основе анализа исторических данных, собранных в хронологическом порядке. Основная задача заключается в идентификации закономерностей в данных и их использовании для построения моделей, аппроксимирующих зависимость между прошлыми и будущими наблюдениями. Временные ряды характеризуются наличием структурных компонентов, таких как тренд, сезонность и случайные колебания, что требует применения специализированных методов для достижения высокой точности прогнозов.

Прогнозирование временных рядов направлено на минимизацию ошибки предсказания посредством моделирования зависимости текущих значений от прошлых. Формально задача прогнозирования может быть определена как нахождение функции f, которая отображает исторические значения ряда $y1, y2, \ldots, yt$ на будущие значения $Yt + h$, где h – горизонт прогнозирования. Методы прогнозирования классифицируются по нескольким критериям:

- по типу подхода: качественные (основанные на экспертных оценках) и количественные (основанные на числовом анализе);
- по горизонту прогнозирования: краткосрочное (до 3 временных точек), среднесрочное (от 3 до 12 точек) и долгосрочное (более 12 точек);
- по типу данных: дискретные (наблюдения с фиксированным интервалом) и непрерывные (непрерывный поток данных).

В данной работе рассматриваются количественные методы прогнозирования, применяемые к дискретным временным рядам,

поскольку они обеспечивают автоматизацию анализа и подходят для обработки структурированных данных. На рисунке 1 представлен наглядный

пример прогнозирования временного ряда, где h – горизонт прогнозирования.

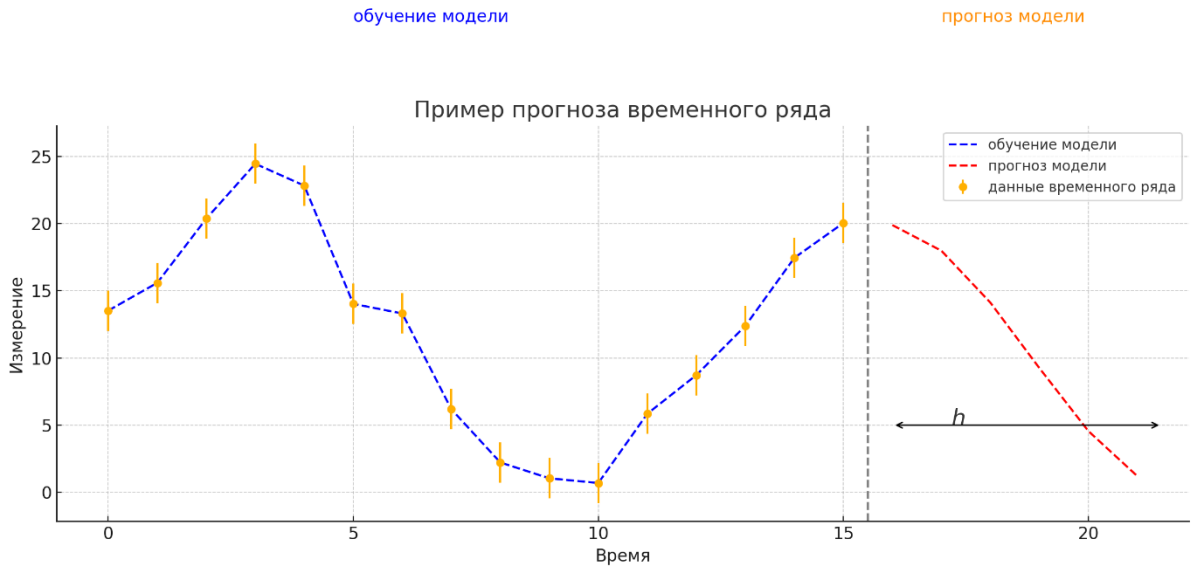

Рисунок 1 – Пример работы прогноза временного ряда

Статистические методы прогнозирования основаны на предположении, что будущее значение временного ряда может быть выражено как линейная функция его предыдущих значений и случайной ошибки.

Статистические методы являются традиционным инструментом для анализа временных рядов, представляя собой комбинацию функции их предыдущих значений и случайной ошибки, что включает различные подходы к обработке данных.

Авторегрессионные модели (AR): Авторегрессионная модель порядка p (AR(p)) представляет собой один из ключевых инструментов статистического анализа временных рядов, который описывает текущее значение ряда как линейную комбинацию p предыдущих значений, дополненную случайной ошибкой. Этот метод основан на предположении, что текущее значение временного ряда зависит от его собственных прошлых значений, что делает его особенно полезным для анализа стационарных рядов – тех, у которых среднее значение и дисперсия остаются постоянными со временем [3]. Модель AR(p) выражается уравнением (1):

$$y_t = \phi_1 y_{\{t-1\}} + \phi_2 y_{\{t-2\}} + \cdots + \phi_p y_{\{t-p\}} + \varepsilon_t y_t, \quad (1)$$

где ϕi – коэффициенты модели,

y_t – значение временного ряда в момент t;

$y_{\{t-1\}}$ – значения ряда в предыдущие моменты;

p – порядок авторегрессии;

ε_t – случайная ошибка.

Авторегрессионная модель относится к классу линейных моделей временных рядов и описывается уравнением, в котором текущее значение ряда представляется как линейная комбинация его прошлых значений с добавлением случайного шума [1].

Подробный процесс работы авторегрессионной модели представлен на блок-схеме, показанной на рисунке 2. Такие модели эффективны для стационарных рядов с выраженной автокорреляцией.

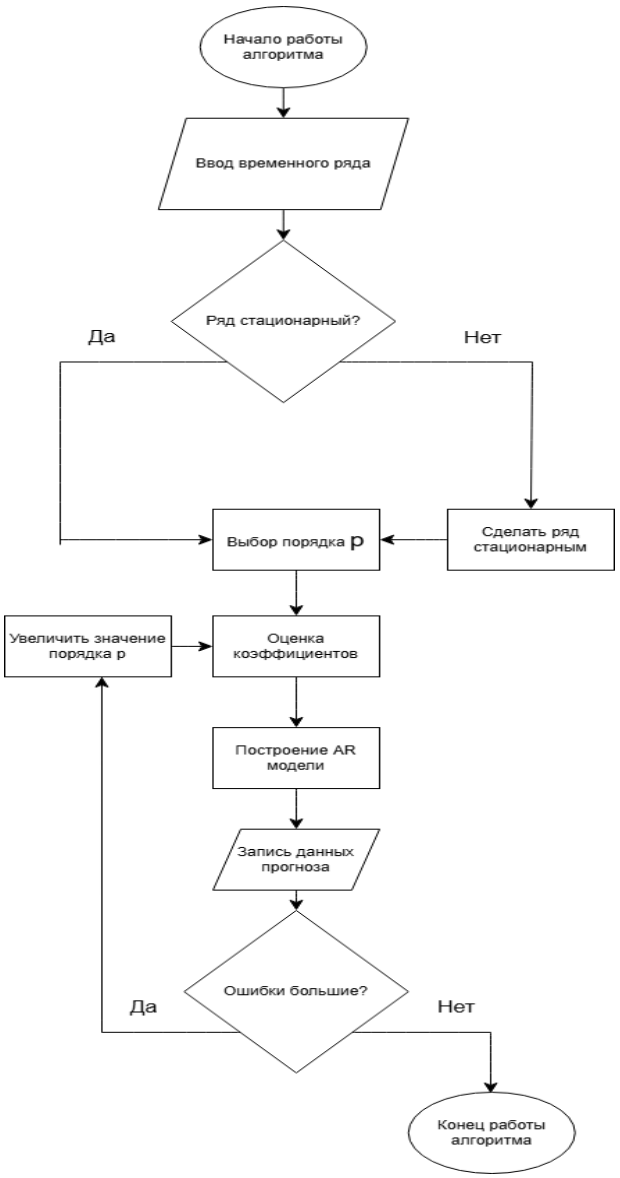

Рисунок 2 – блок схема работы авторегрессионной модели

Модель скользящего среднего порядка q (MA(q)): представляет собой один из базовых подходов к моделированию стационарных временных рядов. В данной модели текущее значение наблюдаемого процесса выражается как линейная комбинация текущего и предыдущих q случайных возмущений (ошибок прогноза), а также, при необходимости, среднего значения ряда. В отличие от авторегрессионной модели (AR), в которой текущее значение ряда моделируется на основе его предыдущих значений, модель MA ориентирована на учет влияния случайных возмущений, происходивших на предыдущих

шагах. Это делает MA-модель особенно полезной в случаях, когда присутствуют кратковременные случайные колебания, оказывающие значимое влияние на процесс, но не сохраняющиеся во времени [3]. Значение y_t временного ряда в момент времени t формируется как константа μ, дополненная текущим случайным шумом ϵ_t и суммой взвешенных предыдущих значений шума $\epsilon_{\{t-1\}}$, $\epsilon_{\{t-2\}}$, $\epsilon_{\{t-q\}}$, где веса определяются коэффициентами $\theta_1, \dots, \theta_q$. Таким образом, уравнение MA(q) принимает следующий вид (2):

$$y_t = \mu + \epsilon_t + \theta_1\epsilon_{\{t-1\}} + \theta_2\epsilon_{\{t-2\}} + \dots + \theta_q\epsilon_{\{t-q\}}, \quad (2)$$

где μ – константа,

$\theta_1, \dots, \theta_q$ – параметры модели скользящего среднего порядка q;

ϵ_t – ошибка (белый шум).

Модели MA подходят для обработки случайных колебаний.

Процесс работы модели скользящего среднего проиллюстрирован на блок-схеме, представленной на рисунке 3.

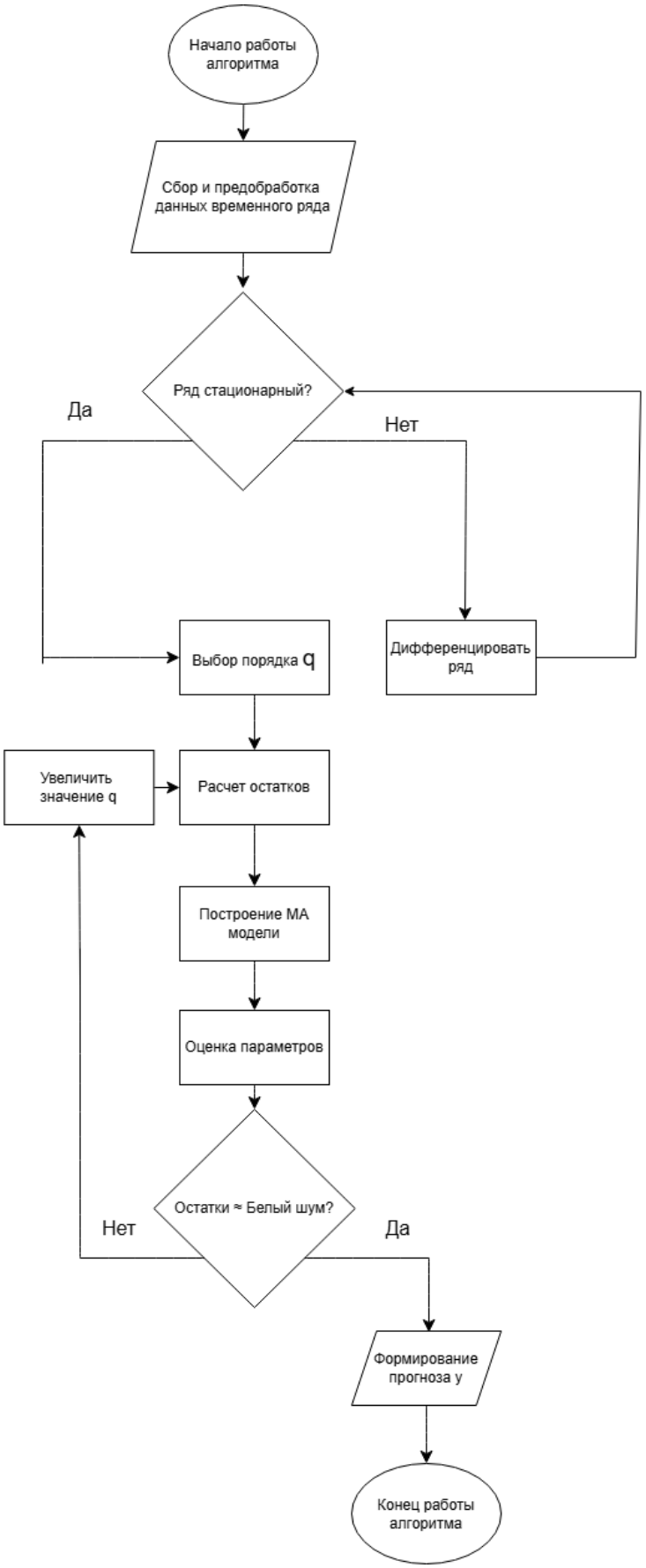

Рисунок 3 – блок-схема работы модели скользящего среднего

Модели ARIMA, модель авторегрессии интегрированного скользящего среднего (ARIMA(p,d,q)) объединяет авторегрессию, скользящее среднее и дифференцирование для устранения нестационарности. Параметр d определяет порядок дифференцирования, необходимый для приведения ряда к стационарному виду. Автоматические реализации ARIMA. AutoARIMA применяет алгоритмы оптимизации для выбора оптимальных значений параметров p, d, q, минимизируя информационные критерии, такие как критерий Акаике (AIC), критерий Байеса (BIC) или другие метрики, которые балансируют между сложностью модели и её способностью объяснять данные [12]. На рисунке 4 представлена блок-схема работы алгоритма ARIMA с возможностью использования autoARIMA.

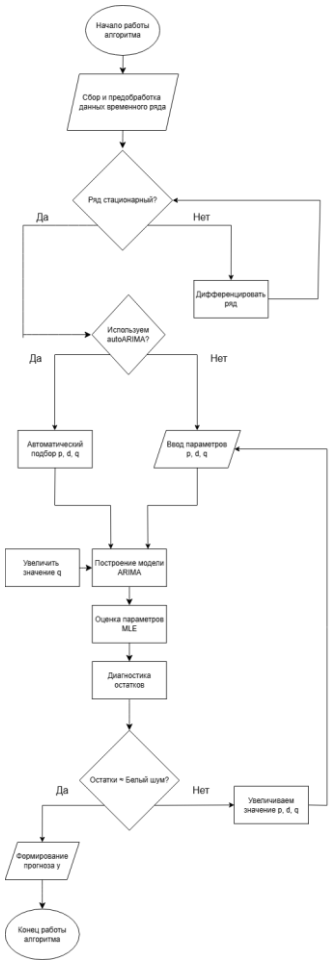

Рисунок 4 – Блок-схема работы алгоритма ARIMA

Экспоненциальное сглаживание представляет собой класс методов прогнозирования временных рядов, основанных на взвешенном усреднении прошлых наблюдений, при котором более свежим данным присваивается больший вес. Это делает методы экспоненциального сглаживания особенно эффективными для анализа и прогнозирования рядов с устойчивыми трендами и сезонными колебаниями. Их простота, вычислительная эффективность и высокая интерпретируемость обеспечили широкое распространение как в академической, так и в прикладной практике.

Существует несколько разновидностей методов экспоненциального сглаживания, каждая из которых адаптирована под определённый тип временного ряда [17]: Основные варианты включают:

– простое экспоненциальное сглаживание применяется к стационарным временным рядам, в которых отсутствуют явные тренды или сезонность. Прогноз $\hat{y}_{t+1}$ временного ряда на следующий момент времени $t+1$ формируется как взвешенная сумма текущего фактического значения y_t и предыдущего прогноза $\hat{y}_t$, где веса определяются параметром сглаживания α (3):

$$\hat{y}_{t+1} = ay_t + (1-a)\hat{y}_t, \quad (3)$$

где $\alpha \in [0,1]$ – параметр сглаживания,

$\hat{y}_{t+1}$ – прогноз на следующий момент времени;

$\hat{y}_t$ – предыдущий прогноз;

y_t – наблюдаемое значение.

– чем выше значение α, тем больше вес текущему значению и меньше – прошедшим значениям. Этот метод эффективен при сглаживании шумных рядов, не содержащих систематических изменений;

– метод Holt расширяет простое сглаживание за счёт включения компоненты линейного тренда. Он использует два уравнения – одно для оценки уровня ряда и второе для оценки тренда. Это позволяет

эффективно моделировать и прогнозировать ряды, демонстрирующие устойчивый рост или спад. Метод Хольта особенно полезен в экономических и производственных данных, где наблюдаются линейные изменения со временем;

– метод Winter (Holt-Winters) вляется дальнейшим расширением, включающим, помимо уровня и тренда, компоненту сезонности.

Статистические методы демонстрируют высокую эффективность при обработке данных с чётко выраженными закономерностями, однако их возможности ограничены при наличии сложных нелинейных зависимостей или нерегулярных изменений. методы экспоненциального сглаживания демонстрируют высокую точность и стабильность при прогнозировании временных рядов, обладающих чётко выраженными и устойчивыми компонентами. Однако они менее эффективны в случае, когда поведение ряда определяется сложными нелинейными зависимостями, структурными сдвигами или аномалиями. В таких ситуациях предпочтение следует отдавать гибридным подходам, которые сочетают в себе статистические и машинные методы.

Методы машинного обучения получили распространение благодаря способности моделировать нелинейные зависимости и сложные структуры данных.

Среди наиболее распространённых подходов можно выделить следующие:

– нейронные сети: нейронные сети, такие как многослойные перцептроны или рекуррентные сети, способны улавливать сложные временные зависимости. Однако их применение к временным рядам требует значительных вычислительных ресурсов и больших объёмов данных, а также тщательной настройки архитектуры [8];

– градиентный бустинг: алгоритмы градиентного бустинга, такие как XGBoost, строят ансамбль слабых предсказательных моделей (обычно деревьев решений), что позволяет эффективно

моделировать нелинейные зависимости. Для временных рядов требуется предварительная обработка данных, включая извлечение признаков [13];

- машины опорных векторов (SVR): машины опорных векторов для регрессии используют ядровые функции для отображения данных в пространстве высокой размерности, где выполняется линейная регрессия. SVR эффективны для небольших наборов данных, но чувствительны к выбору гиперпараметров [14].

Методы машинного обучения демонстрируют высокую производительность при наличии достаточного объёма данных, однако их прямое применение к временным рядам часто осложняется необходимостью учёта сезонности, трендов и временной структуры.

Комбинированные методы объединяют преимущества статистических подходов и машинного обучения для повышения точности прогнозирования.

К основным направлениям комбинированных методов относятся:

- декомпозиция временных рядов: разделение ряда на тренд, сезонность и остатки с последующим прогнозированием каждой компоненты отдельно. Метод STL (Seasonal-Trend decomposition using Loess) является широко используемым инструментом для этой цели, обеспечивая точное выделение компонентов [11];
- ансамблевые методы: объединение прогнозов нескольких моделей с использованием весовых коэффициентов или мета-обучения [16];
- гибридные модели: интеграция статистических методов для обработки структурированных компонентов (тренд, сезонность) с машинным обучением для моделирования остатков [5].

Декомпозиция временных рядов позволяет упростить задачу прогнозирования, поскольку каждая компонента характеризуется меньшей сложностью по сравнению с исходным рядом. Комбинированные подходы демонстрируют высокую эффективность в задачах, требующих учёта разнородных закономерностей. В таблице 1 представлены преимущества и

недостатки отдельных методов прогнозирования в гибридной модели.

Таблица 1 – Преимущества и недостатки отдельны методов прогнозирования модели

Метод	Преимущества	Недостатки
Декомпозиция STL	Разделяет ряд на тренд, сезонность и остатки, упрощая прогнозирование Устойчивость к выбросам благодаря локальной регрессии Loess Точное выделение компонентов	Чувствительность к выбору параметров (например, длины сезонного цикла) Не учитывает нелинейные зависимости в данных Может быть вычислительно сложной для длинных рядов
Экспоненциальное сглаживание(Holt)	Простота реализации и интерпретации Эффективно для рядов с линейным трендом Низкая вычислительная сложность Хорошо работает на коротких горизонтах прогнозирования	Ограниченная способность моделировать сложные нелинейные зависимости Не учитывает сезонность (в базовой версии Holt) Чувствительность к выбору параметров сглаживания
Экспоненциальное сглаживание(Holt-Winters)	Учитывает тренд и сезонность одновременно Простота настройки и применения Хорошая точность для рядов с выраженной сезонностью Быстрое выполнение даже на больших наборах данных	Не справляется с нелинейными закономерностями Может быть неустойчивым при наличии выбросов Требует ручной настройки параметров (α, β, γ) Ограниченная гибкость для сложных сезонных паттернов
AutoARIMA	Автоматический подбор параметров (p, d, q) с помощью критериев (AIC) Устраняет нестационарность через дифференцирование Хорошо работает с линейными рядами Устойчивость к трендам благодаря параметру d	Предполагает линейную зависимость, что снижает точность на нелинейных данных Не учитывает экзогенные переменные Высокая вычислительная сложность при большом переборе параметров

Продолжение таблицы 1

Метод	Преимущества	Недостатки
XGBoost	Высокая точность благодаря учёту нелинейных зависимостей Возможность работы с экзогенными переменными Устойчивость к шуму и выбросам Эффективно для прогнозирования остатков	Требует тщательной настройки гиперпараметров (глубина дерева и др.) Высокая вычислительная сложность на больших данных Может переобучаться на малых наборах данных
SVR	Хорошо моделирует нелинейные зависимости Устойчивость к выбросам благодаря ε-инсентивной функции потерь Гибкость в настройке через выбор ядра (например, RBF) Эффективно для прогнозирования остатков	Чувствительность к выбору ядра и гиперпараметров (C, γ) Высокая вычислительная сложность для больших наборов данных Требует нормализации данных перед применением Ограниченная интерпретируемость результатов

Выводы по главе 2

Основные проблемы прогнозирования временных рядов связаны с наличием сложных структур, ограниченным объёмом данных и необходимостью учёта внешних факторов. Статистические методы ограничены в обработке нелинейных зависимостей, тогда как методы машинного обучения требуют тщательной предварительной обработки данных. Комбинированные подходы, использующие декомпозицию и интеграцию различных методов, представляют перспективное направление для повышения точности прогнозов. В данной работе предлагается гибридный подход, основанный на декомпозиции временного ряда, прогнозировании тренда и сезонности методами экспоненциального сглаживания и моделировании остатков с применением алгоритмов машинного обучения. Такой подход направлен на преодоление ограничений отдельных методов и обеспечение высокой точности на различных горизонтах прогнозирования.

Глава 3 Методология исследования

Разработанная методология прогнозирования временных рядов основана на гибридном подходе, который интегрирует декомпозицию данных, статистические методы и алгоритмы машинного обучения для достижения высокой точности предсказаний. Основная идея метода заключается в том, чтобы разделить временной ряд на более простые компоненты – тренд, сезонность и остатки, – каждая из которых моделируется независимо с использованием подходов, наиболее подходящих для её характеристик, после чего результаты объединяются для получения итогового прогноза. Такой подход позволяет эффективно учитывать сложную структуру временных рядов, включая долгосрочные тенденции, периодические колебания и случайные отклонения, что делает его особенно полезным для прогнозирования данных с выраженной сезонностью и нелинейными закономерностями. В данном разделе подробно описываются этапы обработки данных, применяемые алгоритмы прогнозирования, метрики оценки качества, а также особенности программной реализации и настройки параметров.

Общая схема подхода:

- предварительная обработка данных с применением преобразования для стабилизации дисперсии. На этом этапе временной ряд подготавливается к дальнейшему анализу путём устранения значительных колебаний дисперсии, которые могут затруднять декомпозицию и моделирование. Для этого применяется преобразование Бокса-Кокса, которое позволяет привести данные к более стабильному виду, улучшая качество последующих этапов прогнозирования;
- декомпозиция временного ряда на тренд, сезонную компоненту и остатки с использованием метода STL;
- метод STL разделяет исходный ряд на три составляющие: тренд,

отражающий долгосрочные изменения; сезонность, описывающую периодические колебания; и остатки, представляющие случайные отклонения. Этот этап позволяет упростить задачу прогнозирования, выделяя компоненты с различной динамикой;

- прогнозирование тренда и сезонности с применением методов экспоненциального сглаживания. Для тренда используется метод Holt, который учитывает линейные тенденции, а для сезонной компоненты применяется метод Holt-Winters, способный моделировать периодические колебания с фиксированным циклом, например, 12 месяцев для месячных данных. Эти методы эффективны благодаря их способности адаптироваться к изменениям в данных с минимальными вычислительными затратами;
- извлечение признаков из остатков и их прогнозирование с использованием моделей машинного обучения. Остатки, как правило, содержат нелинейные и случайные зависимости, которые сложно моделировать традиционными методами. На этом этапе из остатков извлекаются признаки, такие как лаги или статистические характеристики, после чего применяются алгоритмы машинного обучения, такие как XGBoost и SVR, для их прогнозирования. Это позволяет уловить сложные закономерности, недоступные для статистических методов;
- линейное объединение прогнозов всех компонентов для получения итогового предсказания. После прогнозирования тренда, сезонности и остатков результаты суммируются, чтобы сформировать итоговый прогноз временного ряда. Такой подход обеспечивает точность за счёт разделения задачи на подзадачи и применения специализированных методов для каждой из них. Такой подход позволяет учитывать сложную структуру временных рядов, разделяя задачу прогнозирования на более простые подзадачи, каждая из

которых решается с использованием наиболее подходящего метода.

Для обеспечения стабильности временного ряда перед декомпозицией применяется преобразование Бокса-Кокса. Данное преобразование устраняет значительные колебания дисперсии, делая данные более пригодными для последующего анализа. Формула преобразования Бокса-Кокса, описывающая преобразованное значение $y^{(\lambda)}$ временного ряда, действует на исходное значение y и зависит от параметра λ, который определяет тип трансформации. Она пропорциональна разности $y^{\lambda} - 1$ и обратно пропорциональна самому λ в случае, когда $\lambda \neq 0$, а при $\lambda = 0$переходит в логарифмическую форму. Преобразование имеет следующий вид (4):

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, \lambda \neq 0 \\ ln(y), \lambda = 0 \end{cases}, \quad (4)$$

где $y^{(\lambda)}$ – преобразованное значение,

y – исходное значение ряда;

λ – параметр преобразования, подбираемый автоматически на основе максимизации логарифма функции правдоподобия.

После завершения прогнозирования выполняется обратное преобразование для возвращения значений в исходную шкалу [4].

Декомпозиция временного ряда осуществляется с использованием метода STL, который разделяет данные на три компоненты:

- тренд (T_t): отражает долгосрочное изменение уровня ряда, например, устойчивый рост или снижение;
- сезонная компонента (S_t): представляет периодические колебания с фиксированным циклом (например, 12 месяцев для месячных данных);
- остатки (R_t): случайные отклонения, не объяснённые трендом и сезонностью;

– Метод STL основан на локальной регрессии Loess, которая обеспечивает устойчивость к выбросам и точное выделение компонентов [11]. Временной ряд y_t (5) представляется как:

$$y_t = T_t + S_t + R_t, \quad (5)$$

где T_t – тренд,

S_t – сезонная компонента;

R_t – остатки.

Параметры декомпозиции, такие как длина сезонного цикла и степень сглаживания, задаются в зависимости от характеристик данных [4]. Для месячных рядов длина цикла устанавливается равной 12.

Каждая компонента временного ряда прогнозируется отдельно с использованием специализированных методов, оптимально соответствующих её свойствам.

Для моделирования тренда применяется метод Holt, представляющий собой вариант экспоненциального сглаживания с учётом линейного тренда [18]. Метод включает два уравнения:

Уравнение уровня (модель Хольта) (6):

$$l_t = \alpha y_t + (1 - \alpha)\left(l_{\{t-1\}} + b_{\{t-1\}}\right), \quad (6)$$

где l_t – уровень в момент t,

$b_{\{t-1\}}$ – тренд на предыдущем шаге;

α – коэффициент сглаживания уровня.

Уравнение тренда (модель Хольта) (7):

$$b_t = \beta\left(l_t - l_{\{t-1\}}\right) + (1 - \beta)b_{\{t-1\}}, \quad (7)$$

где l_t – уровень ряда,

b_t – тренд;

$\alpha, \beta \in [0, 1]$ – параметры сглаживания.

Прогноз на h шагов вперёд рассчитывается как (8):

$$\hat{y}_{t+h} = l_t + h \cdot b_t, \quad (8)$$

где $\hat{y}_{t+h}$ – прогнозируемое значение на h шагов вперёд,

l_t – уровень ряда на момент времени t;

b_t – оценка тренда на момент времени t;

h – горизонт прогнозирования.

Параметры α и β подбираются в диапазоне [0.2, 0.8] с использованием оптимизации по критерию минимизации ошибки на тренировочных данных [4].

Сезонная компонента моделируется с применением метода Winter, который расширяет экспоненциальное сглаживание для учёта периодических колебаний [17]. Метод включает три уравнения:

Уравнение уровня (модель Винтерса) (9):

$$l_t = a\frac{y_\tau}{s_{t-m}} + (1 - a)(l_{t-1} + b_{t-1}), \quad (9)$$

где s_{t-m}– сезонная составляющая m шагов назад.

Уравнение сезонности (модель Винтерса) (10):

$$s_t = \gamma\frac{y_t}{l_t} + (1 - \gamma)s_{t-m}, \quad (10)$$

где St – сезонная компонента,

m – длина сезонного цикла (например, 12 для месячных данных);

$\gamma \in [0, 1]]$ – параметр сглаживания сезонности.

И уравнение тренда (модель Хольта) (7). Прогноз на h шагов вперёд определяется как (12):

$$\hat{y}_{t+h} = (l_t + h \cdot b_t) \cdot s_{t+h-m}, \quad (12)$$

где s_{t+h-m} – прогноз сезонной компоненты.

Параметр γ оптимизируется в том же диапазоне [0.2, 0.8].

Остатки представляют собой стационарный ряд, содержащий случайные колебания, не объяснённые трендом и сезонностью. Для их моделирования применяется двухэтапный процесс:

- извлечение признаков: с использованием библиотеки tsfresh из остатков извлекаются статистические характеристики, включая среднее, стандартное отклонение, автокорреляционные коэффициенты, квантили и другие метрики. Это позволяет представить остатки в виде набора признаков, пригодных для машинного обучения [12];
- моделирование: на основе извлечённых признаков строятся две модели;
- градиентный бустинг (XGBoost): ансамблевый метод с параметрами: максимальная глубина деревьев – 5, количество итераций – 100. Модель оптимизируется по критерию минимизации среднеквадратичной ошибки;
- машины опорных векторов (SVR): модель с радиальным базисным ядром, гиперпараметры которой подбираются с использованием кросс-валидации.

Прогноз остатков на h шагов вперёд выполняется для каждой модели отдельно, после чего выбирается результат с наименьшей ошибкой на валидационной выборке.

Итоговый прогноз формируется путём линейного суммирования

прогнозов всех компонентов:

Объединение прогнозов (13):

$$\hat{y}_{t+h} = \hat{T}_{t+h} + \hat{S}_{t+h} + \hat{R}_{t+h}, \quad (13)$$

где $\hat{T}$ – прогнозы тренда,

$\hat{S}$– прогнозы сезона;

$\hat{R}$– прогнозы остатков компонент.

Прогнозы тренда, сезонности и остатков соответственно. Линейное объединение выбрано для минимизации дополнительных ошибок, связанных с использованием сложных весовых схем.

Для оценки точности прогнозирования используются две метрики:

– симметричная средняя абсолютная процентная ошибка (sMAPE). Метрика sMAPE, оценивающая качество прогноза для временного ряда, выражает относительную ошибку в процентах между фактическими значениями y_t и прогнозируемыми значениями $\hat{y}_t$, усредненную по всем точкам прогноза n (14). Она пропорциональна абсолютной разнице между фактическими и прогнозируемыми значениями и обратно пропорциональна сумме их абсолютных величин, с дополнительным масштабирующим коэффициентом:

$$\text{sMAPE} = \frac{1}{n}\sum_{t=1}^{n} \frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2} \cdot 100, \quad (14)$$

где y_t – фактическое значение,

$\hat{y}_t$ – прогнозируемое значение;

n – число точек прогноза.

Метрика выражает ошибку в процентах и устойчива к масштабу данных [4].

– средняя абсолютная шкалированная ошибка (MASE)

Метрика MASE, оценивающая точность прогноза модели для временного ряда, выражает среднюю ошибку модели $|y_t - \hat{y}t|$ в сравнении с ошибкой наивного прогноза, усредненную по всем точкам прогноза n. Она пропорциональна средней абсолютной ошибке модели и обратно пропорциональна средней абсолютной ошибке наивного прогноза, основанного на частоте данных f (15):

$$\text{MASE} = \frac{\frac{1}{n}\Sigma_{t=1}^{n}|y_t - \hat{y}t|}{\frac{1}{T-1}\Sigma_{t=2}^{T}|y_t - y_t - 1|}, \quad (15)$$

где в числителе – MAE (средняя абсолютная ошибка) модели;

в знаменателе – MAE (средняя абсолютная ошибка) наивной модели.

MASE позволяет сравнивать модели относительно простого бенчмарка [4].

Обе метрики применяются для оценки прогнозов на горизонте 18 точек, что соответствует долгосрочному прогнозированию Блок-схема гибридного алгоритма, представленная на рисунке 5, охватывает полный цикл работы с временными рядами, включая сбор данных, их предобработку, комбинированное моделирование и анализ результатов.

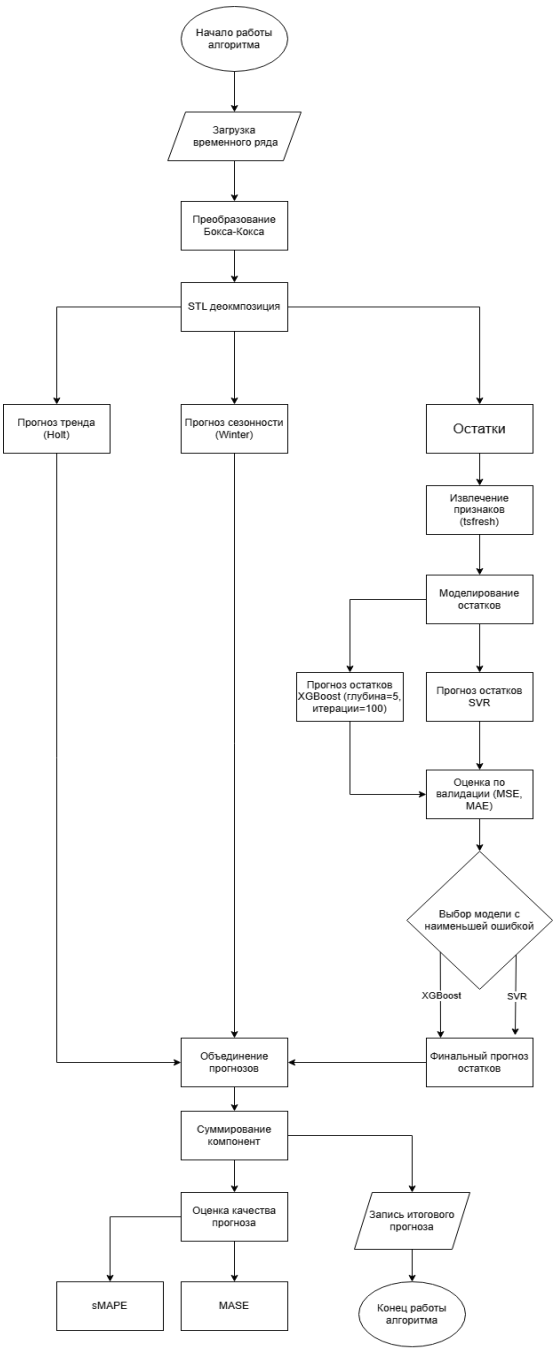

Рисунок 5 – Блок-схема работы гибридного алгоритма

Для тестирования предложенного подхода используются два набора данных:

– NN3: содержит 111 временных рядов, из которых подмножество из 11 рядов используется для предварительных тестов. Данные представляют месячные показатели с выраженной сезонностью и трендами;

– M4: подмножество из 48 000 месячных рядов, характеризующихся сложной структурой, включая тренды, сезонность и случайные компоненты.

Оба набора включают только положительные значения, что накладывает ограничения на применимость модели к рядам с отрицательными значениями. Данные разделяются на тренировочную и тестовую части, где тестовая часть охватывает 18 точек для оценки прогнозов.

Выводы по главе 3

Методология реализована на языке программирования Python с использованием следующих библиотек:

– Anaconda: управление зависимостями и среда для анализа данных.
– Statsmodels: реализация методов экспоненциального сглаживания.
– tsfresh: автоматическое извлечение признаков из остатков.
– stldecompose: выполнение декомпозиции STL.
– scipy: применение преобразования Бокса-Кокса.

Параметры моделей, такие как длина скользящего окна и гиперпараметры машинного обучения, оптимизируются на тренировочных данных для минимизации ошибки прогноза.

Глава 4 Экспериментальная оценка

Экспериментальная оценка разработанного гибридного подхода к прогнозированию временных рядов проводилась с целью проверки его эффективности в сравнении с традиционными статистическими методами и другими комбинированными подходами. В данном разделе описываются настройки экспериментов, используемые наборы данных, метрики оценки, полученные результаты и их анализ.

Основные параметры экспериментов включали:

- размер скользящего окна (16) для анализа данных определялся как:

$$w = h \cdot 2.5, \tag{16}$$

где h – горизонт прогнозирования (18 точек).

- параметры экспоненциального сглаживания (α, β, γ) для методов Holt и Winter подбирались в диапазоне [0.2, 0.8] с использованием оптимизации по критерию минимизации ошибки на тренировочных данных.

Для модели градиентного бустинга (XGBoost) устанавливались следующие параметры: максимальная глубина деревьев – 5, количество итераций – 100, оптимизация по среднеквадратичной ошибке.

Для машины опорных векторов (SVR) использовалось радиальное базисное ядро с гиперпараметрами, оптимизированными посредством кросс-валидации.

Качество прогнозов оценивалось с использованием двух метрик:

- sMAPE (симметричная средняя абсолютная процентная ошибка), выражающая относительную ошибку в процентах;
- MASE (средняя абсолютная шкалированная ошибка), сравнивающая ошибку модели с наивным прогнозом, использующим последнее

наблюдение.

Сравнение проводилось со следующими методами:

- AutoARIMA: автоматическая реализация модели ARIMA с подбором параметров.Holt-Winters;
- метод экспоненциального сглаживания с учётом тренда и сезонности;
- XGBoost и SVR без декомпозиции: Модели машинного обучения, применённые непосредственно к исходным данным.

На подмножестве из 11 временных рядов набора NN3 были получены результаты, представленные в Таблице 2. Оценка проводилась на горизонте прогнозирования 18 точек.

Таблица 2 – Производительность методов на подмножестве из 11 рядов NN3

Метод	sMAPE (%)	MASE	Стандартное отклонение sMAPE (%)	Стандартное отклонение MASE
Предложенная модель	14.17	0.912	12.09	0.394
AutoARIMA	18.15	0.906	14.22	0.412
Holt-Winters	21.22	1.495	16.87	0.673
XGBoost(без декомпозиции)	36.11	2.134	22.45	0.892

Предложенная гибридная модель продемонстрировала наименьшую ошибку по метрике sMAPE (14.17%) и конкурентоспособное значение MASE (0.912), превосходя AutoARIMA и Holt-Winters. Модель XGBoost, применённая без декомпозиции, показала значительно худшие результаты, что подтверждает необходимость предварительного разделения ряда на компоненты. Стандартное отклонение ошибок для предложенной модели также оказалось ниже, указывая на её стабильность.

Графическое представление прогнозов для отдельных рядов представлено на рисунке 6, где синей линией отмечены действительные

значения графика, а оранжевой предсказанные, ось x указывает на временные интервалы, а ось y - на значения переменной. Показало высокое соответствие предсказанных значений фактическим данным, особенно в условиях выраженной сезонности.

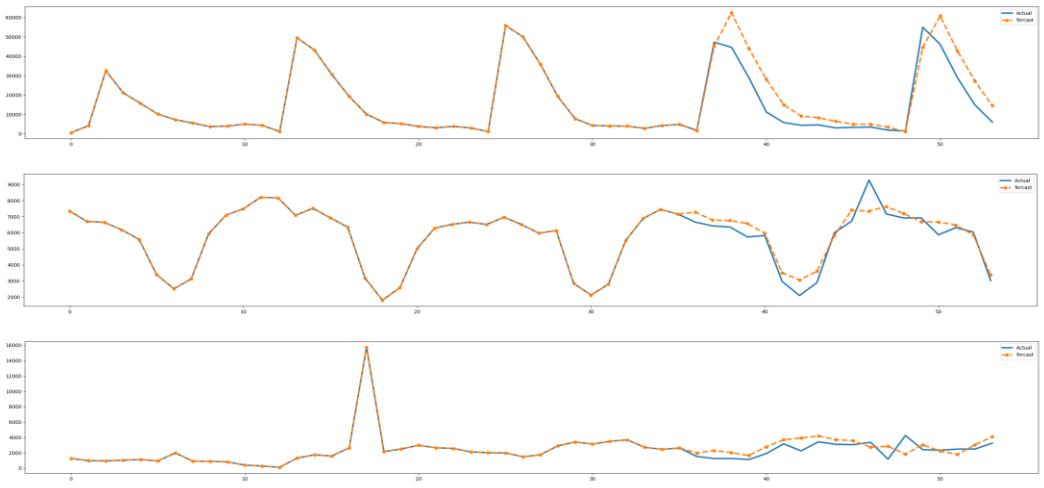

Рисунок 6 – Ряды 103, 104 и 110

На полном наборе из 111 временных рядов NN3 результаты представлены в Таблице 3.

Таблица 3 – Производительность методов на полном наборе NN3

Метод	sMAPE (%)	MASE	Стандартное отклонение sMAPE (%)	Стандартное отклонение MASE
Предложенная модель	16.28	1.300	13.45	0.521
AutoARIMA	16.98	1.228	14.67	0.498
Holt-Winters	88.05	3.214	25.33	1.127
XGBoost (без декомпозиции)	39.47	2.456	23.78	0.934

Предложенная модель показала близкие результаты к AutoARIMA по метрике sMAPE (16.28% против 16.98%) и незначительно уступила по MASE

(1.300 против 1.228). Однако она значительно превзошла Holt- Winters, который продемонстрировал высокие ошибки из-за неспособности эффективно моделировать сложные ряды. Модель XGBoost без декомпозиции вновь оказалась наименее точной, подчёркивая важность предложенного подхода. На рисунке 7 приведены графики, иллюстрирующие динамику ряда для разных периодов.

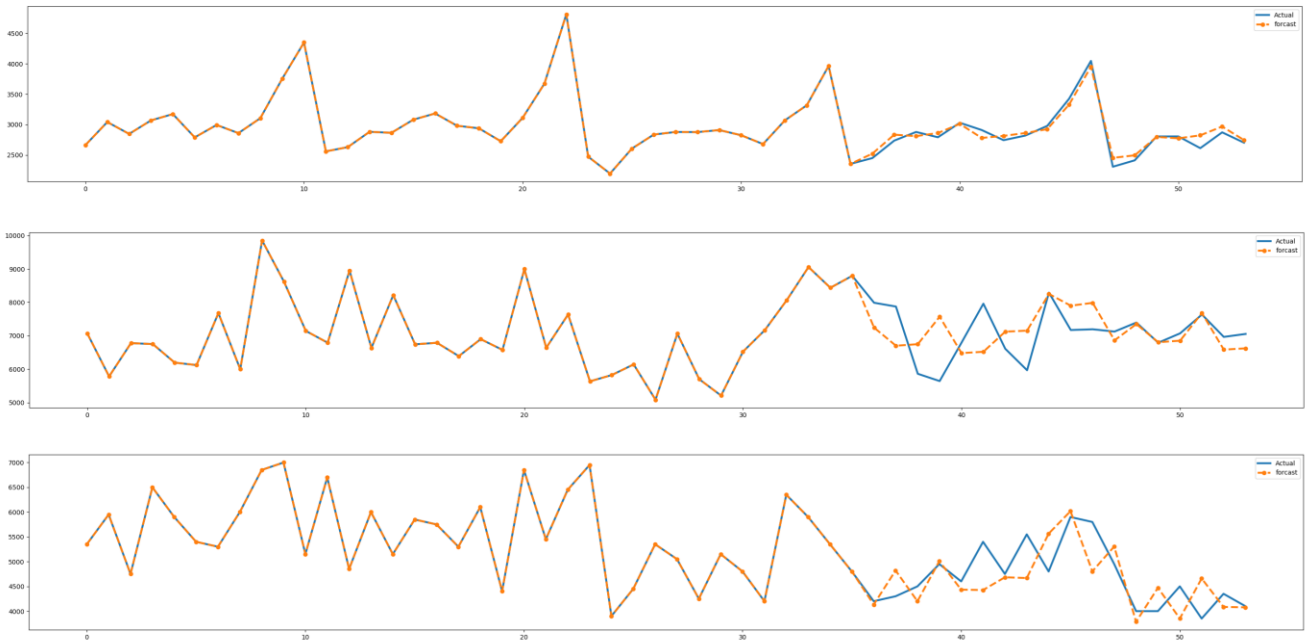

Рисунок 7 – Ряды 70, 73 и 7

На подмножестве из 48 000 рядов набора M4 оценка проводилась с использованием метрики OWA (Overall Weighted Average) (17), объединяющей sMAPE и MASE в единый показатель.

$$\text{OWA} = 0{,}5 \cdot \frac{\text{sMAPEmodel}}{sMAPEnaive} + 0.5 \cdot \frac{\text{MASEmodel}}{MASEnaive}, \quad (17)$$

Где $sMAPEnaive$ – значение sMAPE для базового наивного метода Naive2,

$MASEnaive$ – значение MASE для метода Naive.

Результаты ранжирования методов представлены в Таблице 4.

Таблица 4 – Ранжирование методов на наборе M4:

Метод	OWA
AutoARIMA	0.789
ETS	0.852
Предложенная модель	0.883
Holt-Winters	0.947

Предложенная модель заняла третье место по метрике OWA (0.883), уступив AutoARIMA и ETS, но опередив Holt-Winters. Учитывая сложность набора M4, включающего ряды с разнообразными характеристиками, полученный результат свидетельствует о высокой конкурентоспособности разработанного подхода.

Графический анализ прогнозов для отдельных рядов показал, что предложенная модель точно воспроизводит сезонные колебания и тренды, хотя в рядах с высокой случайной составляющей ошибки были выше. Ряды представлены на рисунке 8.

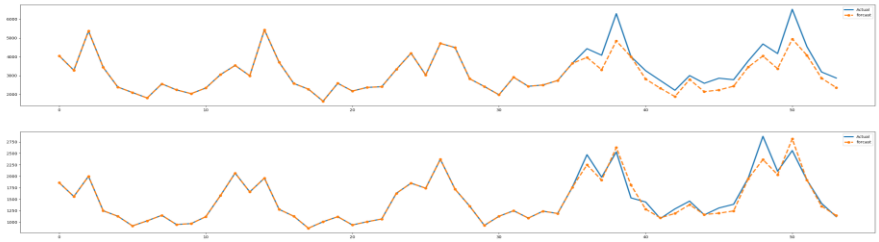

Рисунок 8 – Ряды 11 и 27

Полученные результаты подтверждают эффективность предложенного гибридного подхода. Основные преимущества включают:

– высокую точность прогнозирования за счёт декомпозиции ряда на компоненты, что упрощает задачу моделирования;

– эффективное использование методов машинного обучения для обработки остатков, позволяющее улавливать нелинейные зависимости;
– стабильность результатов, выраженную в низком стандартном отклонении ошибок.

Основные ограничения связаны с характеристиками наборов данных:

Короткая длина рядов в NN3 (менее 150 наблюдений) ограничивает возможности моделей машинного обучения.

Отсутствие отрицательных значений в данных снижает универсальность модели.

Фокус на месячных рядах не позволяет оценить производительность на дневных или годовых данных.

Эксперименты демонстрируют, что предложенный подход особенно эффективен для рядов с выраженной сезонностью и умеренными случайными колебаниями, что делает его перспективным для применения в реальных задачах прогнозирования.

Заключение

Проведённое исследование посвящено разработке и тестированию гибридного подхода к прогнозированию временных рядов, основанного на декомпозиции данных и интеграции статистических методов с алгоритмами машинного обучения. Разработанная модель включает разделение временного ряда на тренд, сезонную компоненту и остатки с использованием метода STL, прогнозирование тренда и сезонности методами экспоненциального сглаживания (Holt и Winter), а также моделирование остатков с применением градиентного бустинга (XGBoost) и машин опорных векторов (SVR) после извлечения признаков.

Экспериментальная оценка проводилась на наборах данных NN3 и M4, что позволило получить следующие выводы.

Разработанный гибридный подход продемонстрировал высокую эффективность в сравнении с традиционными методами прогнозирования. Основные результаты включают:

- повышенная точность прогнозирования: на подмножестве из 11 рядов набора NN3 предложенная модель достигла значения sMAPE 14.17% и MASE 0.912, превосходя AutoARIMA (sMAPE 18.15%, MASE 0.906) и Holt-Winters (sMAPE 21.22%, MASE 1.495). На полном наборе из 111 рядов модель показала sMAPE 16.28% и MASE 1.300, оставаясь конкурентоспособной по сравнению с AutoARIMA (sMAPE 16.98%, MASE 1.228);
- эффективность декомпозиции: разделение временного ряда на компоненты позволило упростить задачу прогнозирования, обеспечив точное моделирование тренда и сезонности с помощью методов экспоненциального сглаживания. Это подтверждается низким стандартным отклонением ошибок, особенно на рядах с выраженной сезонностью;
- вклад машинного обучения: применение моделей XGBoost и SVR к

остаткам после извлечения признаков с использованием библиотеки tsfresh позволило уловить нелинейные зависимости, что улучшило общую точность прогнозов. Сравнение с чистыми моделями машинного обучения (sMAPE 36.11% для XGBoost без декомпозиции) подчёркивает важность предварительной обработки данных;

- конкурентоспособность на больших наборах данных: на подмножестве из 48 000 рядов набора M4 предложенная модель заняла третье место по метрике OWA (0.883), уступив AutoARIMA (0.789) и ETS (0.852), но опередив Holt-Winters (0.947). Это свидетельствует о масштабируемости подхода для обработки сложных данных.

Полученные результаты отвечают на сформулированные задачи исследования:

Гибридный подход превосходит традиционные статистические методы в задачах долгосрочного прогнозирования, особенно для рядов с выраженной сезонностью.

Интеграция методов машинного обучения в гибридную модель обеспечивает прирост точности за счёт моделирования остатков, что невозможно при использовании только статистических методов.

Декомпозиционный подход демонстрирует сопоставимую производительность с другими комбинированными методами, сохраняя простоту реализации и интерпретируемость.

Несмотря на достигнутые результаты, предложенный подход имеет ряд ограничений, которые необходимо учитывать:

- ограниченная длина временных рядов: наборы данных NN3 и M4 содержат относительно короткие ряды (менее 150 наблюдений для NN3), что ограничивает возможности моделей машинного обучения, требующих большого объёма данных для эффективного обучения. Это проявилось в упрощённом прогнозировании остатков, где

использовалось повторение последнего предсказания;

- положительные значения данных: оба набора данных включают только положительные значения, что ограничивает применимость модели к рядам с отрицательными значениями. Преобразование Бокса-Кокса дополнительно усиливает это ограничение, требуя корректировки данных перед обработкой;
- фокус на месячных данных: тестирование проводилось исключительно на месячных временных рядах с сезонным периодом 12 месяцев. Производительность модели на дневных, недельных или годовых данных остаётся неизученной;
- отсутствие доменных знаний: модель не учитывает внешние факторы, такие как экономические события, праздники или другие контекстные переменные, которые могли бы повысить точность прогнозов;
- чувствительность к параметрам декомпозиции: метод STL предполагает аддитивную структуру ряда после преобразования Бокса-Кокса, что может быть неоптимальным для мультипликативных рядов без соответствующей обработки;
- эти ограничения указывают на необходимость дальнейших исследований для расширения применимости модели и устранения выявленных недостатков.

На основе анализа результатов и ограничений исследования предлагаются следующие направления для дальнейшей работы:

- тестирование на разнообразных наборах данных: проведение экспериментов с временными рядами различной периодичности (дневные, недельные, годовые) позволит оценить универсальность предложенного подхода. Исследование рядов с отрицательными значениями или мультипликативной структурой также расширит область применения модели;
- интеграция доменных признаков: включение внешних факторов,

таких как календарные события, экономические показатели или отраслевые характеристики, может улучшить точность прогнозов, особенно для рядов с высокой зависимостью от контекста;

- динамическая настройка параметров: разработка алгоритмов для автоматической адаптации параметров декомпозиции STL и моделей машинного обучения (например, длины скользящего окна или гиперпараметров XGBoost) повысит устойчивость модели к различным типам данных;
- усовершенствование прогнозирования остатков: в текущей реализации прогноз остатков упрощён из-за ограниченной длины рядов. Использование рекуррентных моделей или временных лагов для предсказания остатков на длительных горизонтах может повысить точность;
- оптимизация вычислительной эффективности: текущая реализация требует значительных ресурсов для извлечения признаков с помощью tsfresh. Исследование методов сокращения числа признаков или использования более лёгких алгоритмов позволит масштабировать подход для больших наборов данных;
- сравнение с современными методами: проведение сравнительного анализа с новыми подходами, такими как глубокие нейронные сети (например, LSTM или Transformer) или гибридные модели с байесовскими методами, позволит уточнить место предложенного подхода среди современных решений.

Разработанный гибридный подход представляет практическую ценность для задач прогнозирования временных рядов в различных областях, включая экономику, энергетику и управление запасами. Простота реализации, основанная на доступных библиотеках Python, и высокая точность на наборах данных NN3 и M4 делают модель пригодной для использования в реальных приложениях. Масштабируемость подхода позволяет адаптировать его к различным типам данных, при условии устранения указанных ограничений.

Список используемой литературы

1. Андерсон Т. Статистический анализ временных рядов. – М.: Мир, 1976. – 756 с.
2. Бокс Дж., Дженкинс Г., Рейнсел Г. Анализ временных рядов: прогнозирование и управление. – М.: Мир, 1974. – 406 с.
3. Браун Р.Г. Методы сглаживания и прогнозирования. – М.: Статистика, 1964. – 286 с.
4. Броквелл П.Дж., Дэвис Р.А. Введение в анализ временных рядов и прогнозирование. – М.: Финансы и статистика, 2001. – 668 с.
5. Гаврилов Д.В. Прогнозирование временных рядов в задачах эконометрики и финансов. – М.: КНОРУС, 2020. – 272 с.
6. Гудвин П., Ло Д. Combining forecasts: A review and annotated bibliography // International Journal of Forecasting. – 2006. – Vol. 22, No. 4. – P. 637–645.
7. Дринкер Б. Support vector regression machines // Advances in Neural Information Processing Systems. – 1997. – Vol. 9. – P. 155–161.
8. Кендалл М., Стюарт А. Статистические выводы и связи. – М.: Наука, 1973. – 896 с.
9. Кутузов А.А., Орлова А.В. Временные ряды: от классики к нейросетям. – М.: URSS, 2023. – 366 с.
10. Лемке К., Габрис Б. Meta-learning for time series forecasting // Computational Statistics & Data Analysis. – 2010. – Vol. 54, No. 11. – P. 2632–2644.
11. Литвинова Г.Г., Зайченко Ю.П. Анализ и прогнозирование временных рядов с использованием нейросетевых моделей. – СПб.: БХВ-Петербург, 2011. – 304 с.
12. Макридакис С., Уилрайт С., Хайндман Р.Дж. Методы прогнозирования: количественные и качественные подходы. – М.: Статистика, 1989. – 312 с.

13. Мюллер А., Гвидо С. Введение в машинное обучение с Python: руководство для специалистов по анализу данных. – М.: ДМК Пресс, 2017. – 408 с.

14. Смил С. Deep learning for time series forecasting // IEEE Transactions on Neural Networks and Learning Systems. – 2019. – Vol. 30, No. 4. – P. 1003–1019.

15. Ткаченко П.А. Машинное обучение и анализ данных с использованием Scikit-Learn и TensorFlow. – М.: ДМК Пресс, 2021. – 512 с.

16. Хайндман Р.Дж., Атанасопулос Г. Прогнозирование: принципы и практика. – 2-е изд. – М.: ДМК Пресс, 2019. – 504 с.

17. Хасти Т., Тибширани Р., Фридман Дж. Элементы статистического обучения: интеллектуальный анализ данных, вывод и прогнозирование. – 2-е изд. – М.: ДМК Пресс, 2014. – 768 с.

18. Чатфилд К. Анализ временных рядов: введение. – 6-е изд. – М.: Финансы и статистика, 2004. – 424 с.

19. Эйткен А. On Bernoulli's theory of recurrent series and on inverse probability // Journal of the Royal Statistical Society. – 1942. – Vol. 105, No. 3. – P. 199–211.

20. Bergmeir C et al. A note on the validity of cross-validation for evaluating time series prediction // Journal of Statistical Software. – 2018. – Vol. 83, No. 5. – P. 1–26.

21. Chen T. XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. – P. 785–794.

22. Cleveland R.B. et al. STL: A seasonal-trend decomposition procedure based on loess // Journal of Official Statistics. – 1990. – Vol. 6, No. 1. – P. 3–73.

23. Hyndman R.J., Khandakar Y. Automatic time series forecasting: the forecast package for R // Journal of Statistical Software. – 2008. – Vol. 27, No. 3. – P. 1–22.

24. Makridakis S et al. M4 Competition: Results, findings, conclusion and

way forward // International Journal of Forecasting. – 2018. – Vol. 34, No. 4. – P. 802–808.

25. Taylor S.J. Forecasting at scale // Journal of Forecasting. – 2018. – Vol. 37, No. 1. – P. 83–97.

Приложение А

Листинг (реализация программы)

```
import numpy as np
import pandas as pd
from statsmodels.tsa.api import Holt
from scipy import stats
from sklearn.metrics import mean_squared_error
import warnings
warnings.filterwarnings('ignore')
from Modified_timeseries import smape, mase, get_decomposotion,
my_holt_trend, my_holtwinter_seas,
split_into_train_test_out_tsfresh, in_sampling_test,
my_gradientboosting, mlp_bench, my_LinearModel, my_xgboost,
my_SupportVector, get_inv_box_transform, get_all_stas_result,
get_ML_result

import time
import csv

start_time = time.time()

# Пути к данным
data_files = {
    'info': './Data//M4DataSet/M4-info.csv',
    'train': './Data//M4DataSet/Monthly-train.csv',
    'test': './Data//M4DataSet/Monthly-test.csv',
    'naive': './Data//M4DataSet/submission-Naive.csv',
    'holt': './Data//M4DataSet/submission-Holt.csv',
    'ets': './Data//M4DataSet/submission-ETS.csv',
    'damp': './Data//M4DataSet/submission-Damped.csv',
    'comb': './Data//M4DataSet/submission-Com-1.csv',
    'rnn': './Data/M4DataSet/submission-RNN.csv',
    'mlp': './Data//M4DataSet/submission-MLP.csv'
}

# Функция загрузки данных
def load_data(file_path, starts_with='M'):
    data = []
    with open(file_path) as f:
        next(f)
        for line in f:
            line = line.strip().split(",")
            line = [x.strip('"') for x in list(filter(None,
line))]
            if line[0].startswith(starts_with):
                data.append(line[:19] if 'submission' in
file_path else line)
    return data
```

```
# Загрузка всех данных
data_title = load_data(data_files['info'])
data_train = load_data(data_files['train'])
data_test = load_data(data_files['test'])
data_models = {key: load_data(data_files[key]) for key in
['naive', 'holt', 'ets', 'damp', 'comb', 'rnn', 'mlp']}

# Параметры
loop_start, loop_end, freq, monthly, fh, in_size = 30800, 31000,
1, 12, 18, 40
f =
open(f'./result/{fh}_{in_size}_{loop_start}_{loop_end}_M4_monthl
ydata_v2_200_samples.csv', 'a')

while loop_start < loop_end:
    if len(set([data_train[loop_start][0],
data_title[loop_start][0], data_test[loop_start][0]] +
[data_models[m][loop_start][0] for m in data_models])) == 1:
        train = np.array(data_train[loop_start][1:],
dtype="float")
        test = np.array(data_test[loop_start][1:],
dtype="float")

        # Расчёт метрик ошибок для всех моделей
        metrics = {}
        for model_name, model_data in data_models.items():
            pred = np.array(model_data[loop_start][1:],
dtype="float")
            metrics[model_name] = {'smape': smape(test, pred),
'mase': mase(train, test, pred, freq)}

        # Декомпозиция и прогнозирование
        to_transform = True
        data_log, lambda_, _ = stats.boxcox(train, alpha=0.05)
if to_transform else (train, None, None)
        if len(set(data_log)) == 1:
            to_transform = False
            data_log = train

        trend_, seas_, resid_ = get_decomposotion(data_log,
monthly)
        yhat_holt = my_holt_trend(trend_, fh)
        yhat_wn_holt = my_holtwinter_seas(seas_, fh)

        x_train, y_train, x_test =
split_into_train_test_out_tsfresh(resid_, in_size)
        id__ = in_sampling_test(x_train, y_train)

        pred_ = {
            1: my_gradientboosting,
```

```
                2: mlp_bench,
                3: my_LinearModel,
                4: my_xgboost,

                5: my_SupportVector
            }.get(id__, lambda *args: np.zeros(fh))(x_train,
y_train, x_test, in_size, fh)

            pred__ = np.vstack((yhat_holt, yhat_wn_holt,
pred_.reshape(len(pred_))))
            pred__ = get_inv_box_transform([sum(d) for d in
zip(*pred__)], lambda_) if to_transform else [sum(d) for d in
zip(*pred__)]

            smape_method = smape(test, pred__)
            mase_method = mase(train, test, pred__, freq)

            temp, temp_mase = get_all_stas_result(train, test, fh)
            temp_ML, temp_ML_mase = get_ML_result(train, test,
in_size, fh)

            # Запись результатов
            output = data_title[loop_start] + ["_"] + list(test) +
["_"] * 2 + list(pred__) + ["_"] * 2
            output += [smape_method, "_"] + [metrics[m]['smape'] for
m in metrics] + ["_"] + temp + temp_ML
            output += ["_"] * 3 + [mase_method, "_"] +
[metrics[m]['mase'] for m in metrics] + ["_"] + temp_mase +
temp_ML_mase

            csv.writer(f, dialect='excel').writerow(output)
            loop_start += 1
        else:
            print("Ошибка в naming convention")
            break

f.close()
print(f"--- {time.time() - start_time} seconds ---")
```