МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математической кибернетики и компьютерных наук

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ДЕТЕКЦИИ АНОМАЛИЙ ВО ВРЕМЕННЫХ РЯДАХ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 451 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Шевцовой Варвары Антоновны

Научный руководитель
зав. каф. техн. пр.,
к. ф.-м. н., доцент _____ И. А. Батраева

Заведующий кафедрой
доцент, к. ф.-м. н. _____ С. В. Миронов

Саратов 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Анализ временных рядов представляет собой важное направление в области обработки данных и машинного обучения, предоставляя возможность предсказывать будущие значения на основе прошлых наблюдений. Одной из ключевых задач в этой области является обнаружение аномалий — неожиданных изменений или отклонений в данных, которые могут указывать на важные события или проблемы. Аномалии во временных рядах могут возникать по различным причинам, включая ошибки измерений, изменения внешней среды или критические события, и их своевременное обнаружение имеет решающее значение для принятия правильных управленческих решений.

Целью данной дипломной работы является исследование и сравнение различных методов обнаружения аномалий во временных рядах с использованием современных технологий машинного обучения. Основное внимание уделяется двум подходам: традиционному статистическому методу SARIMA и современным методам на основе рекуррентных нейронных сетей, таких как LSTM.

Для достижения поставленной цели были сформулированы следующие задачи:

- рассмотреть понятия временного ряда и аномалии во временных рядах, а также их классификацию и особенности;
- провести обзор существующих методов обнаружения аномалий в данных;
- подробно рассмотреть и реализовать метод SARIMA для обнаружения аномалий;
- исследовать и реализовать метод LSTM для обнаружения аномалий;
- применить эти методы к реальному набору данных;
- сравнить результаты и сделать выводы об эффективности построенных моделей.

1 Введение в анализ временных рядов

1.1 Понятие временного ряда

Временной ряд — это цепочка точек данных, наблюдаемых и регистрируемых в определенном временном порядке в течение определенного периода. Он представляет собой результат, полученный в результате мониторинга и отслеживания определенных событий или процессов. Пример временного ряда приведён на рис. 1.1.

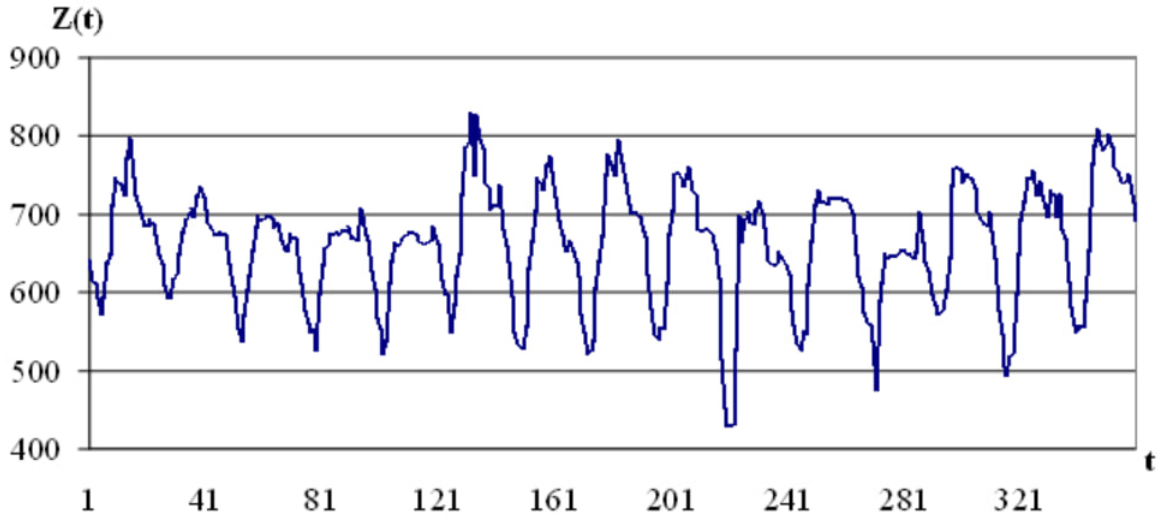

Рисунок 1.1 – Временной ряд цен на электроэнергию

Для многих технологических систем результаты мониторинга можно представить в виде временных рядов. Свойствами временного ряда являются:

- привязка каждого измерения ко времени его возникновения;
- равное расстояние по времени между измерениями;
- возможность из данных предыдущего периода восстановить поведение процесса в текущем и последующих периодах.

1.2 Классификация временных рядов

Временные ряды можно классифицировать по-разному в зависимости от выбранного признака. Выделяют следующие группы:

1. Регулярные и нерегулярные.

   Регулярный временной ряд хранит данные для регулярно (равномерно) распределённых временных точек. *Нерегулярный временной ряд* хранит данные для последовательности произвольных временных точек.

2. Детерминированные и недетерминированные.

   Детерминированный временной ряд может быть выражен аналитическим

выражением в явном виде. В нем нет случайных или вероятностных аспектов. *Недетерминированный временной ряд* не может быть описан аналитическим выражением. Он имеет некоторый случайный аспект, который не позволяет описать его поведение в явном виде.

3. Стационарные и нестационарные.

   Временной ряд считается *стационарным*, если его статистические свойства, такие как среднее значение и дисперсия, не изменяются с течением времени. Временной ряд считается *нестационарным*, если его статистические свойства изменяются со временем, такие как тренд, сезонность или изменение дисперсии.

1.3 Анализ временных рядов

Анализ временных рядов — способ анализа последовательности точек данных, собранных за определенный промежуток времени. Он объединяет различные математические и статистические методы, направленные на выявление особенностей структуры временных данных и их прогнозирование.

Для анализа временных рядов используются различные математические и статистические методы. Ниже приведены некоторые из наиболее распространенных методов:

- Описательный анализ — визуализация данных с помощью графиков, гистограмм и диаграмм.
- Статистический анализ — вычисление основных статистических характеристик, таких как среднее значение, медиана, стандартное отклонение и коэффициент корреляции.
- Спектральный анализ — выявление цикличности и сезонности в данных, например, методы Фурье.
- Автокорреляционный анализ — выявление зависимости между текущими и предыдущими значениями временного ряда.
- Методы прогнозирования — различные методы прогнозирования, такие как методы временных рядов ARIMA (авторегрессия, интеграция, скользящее среднее), экспоненциальное сглаживание и машинное обучение.

1.4 Понятие и классификация аномалий в данных

Аномалии — это отклоняющиеся от нормы события или закономерности, которые не соответствуют ожидаемому прогнозу. Выявление аномалий важно в

широком спектре дисциплин, включая медицинскую диагностику, страхование и мошенничество с идентификацией личности, вторжение в сеть и дефекты программирования.

Аномалии обычно делятся на три типа.

1. *Точечные аномалии* представляют собой отдельные точки данных или наблюдения, которые значительно отличаются от остальных данных. Они являются изолированными случаями и могут возникать из-за ошибок в данных, редких событий или аномального поведения (рис. 1.2).

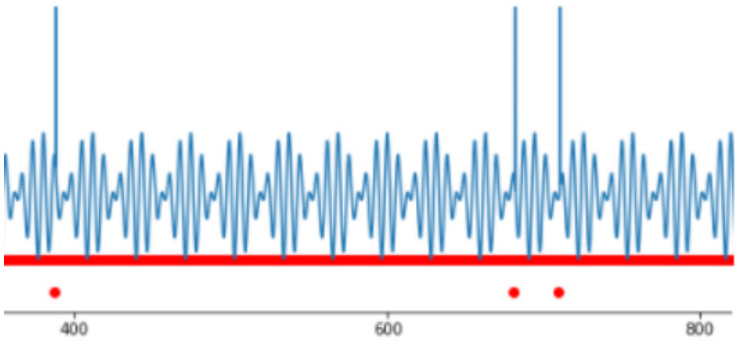

Рисунок 1.2 – Пример точечных аномалий

2. *Контекстные аномалии* представляют собой участки, когда отклонение от общего тренда или шаблона данных не может быть определено, не учитывая контекст окружающих данных. Другими словами, эти аномалии могут быть нормальными в одном контексте, но аномальными в другом.
3. *Коллективные аномалии* представляют собой группы или коллекции точек данных, которые вместе образуют аномальный паттерн. Они могут возникать из-за необычных событий или изменений в данных, которые влияют на несколько наблюдений одновременно (рис. 1.3).

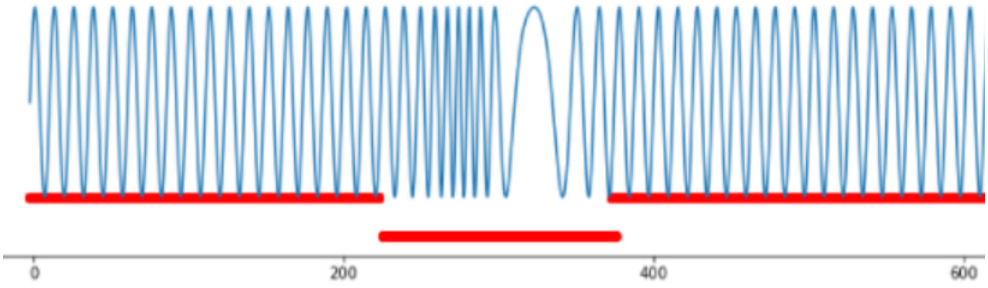

Рисунок 1.3 – Пример групповой аномалии

Поиск аномалий имеет огромное значение. Их наличие может затруднять понимание общей картины процесса или же вовсе свидетельствовать о возможном скором переходе устройства в аварийный режим.

2 Методы обнаружения аномалий во временных рядах

2.1 Классификация методов

Методы, используемые для обнаружения аномалий во временных рядах, обычно делятся на следующие категории:

- Методы на основе близости (proximity-based) — оценивают степень близости между точками данных или последовательностями точек.
- Методы на основе прогнозирования (prediction-based) — строят модели прогнозирования на основе исторических данных и сравнивают прогнозы с фактическими значениями.
- Методы на основе восстановления (reconstruction-based) — восстанавливают фрагменты данных для выявления аномалий..

Методы на основе прогнозирования (prediction-based) обладают преимуществами, поскольку способны учитывать временные зависимости данных и их изменения со временем. Они более гибкие, чувствительны к изменениям и могут адаптироваться к различным сценариям. Ниже рассмотрим некоторые из них подробнее.

2.2 Модель SARIMA

SARIMA (Seasonal AutoRegressive Integrated Moving Average) — универсальная и широко используемая модель прогнозирования временных рядов. Является расширением несезонной модели ARIMA, предназначенной для работы с данными, имеющими сезонный характер. SARIMA учитывает как краткосрочные, так и долгосрочные зависимости в данных, что делает ее надежным инструментом для прогнозирования. Она сочетает в себе концепции моделей авторегрессии (AR), интегрированной (I) и скользящей средней (MA) с сезонными компонентами.

Модель SARIMA можно представить следующим образом:

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right) \left(1 - \sum_{j=1}^{P} \Phi_j B^{sj}\right) (1 - B)^d (1 - B^s)^D y_t = \epsilon_t,$$

где y_t — значение временного ряда в момент времени t, ϵ_t — случайная ошибка в момент времени t, p, P — порядки авторегрессии и сезонной авторегрессии соответственно, ϕ_i, Φ_j — параметры авторегрессии и сезонной авторе-

грессии соответственно, B — оператор лага, который сдвигает временной ряд на один временной период назад (лаг 1), d, D — степени интегрирования, s — период сезонности.

2.3 Модель LSTM

LSTM (Long Short-Term Memory, долгая краткосрочная память) — это улучшенная версия обычной RNN, разработанная для облегчения улавливания долгосрочных зависимостей в последовательных данных. Такая перепроектировка RNN позволяет иметь состояние активации, которое также может выступать в качестве весов и сохранять информацию на больших расстояниях (в отличие от классической RNN), откуда LSTM и получила своё название. Архитектура представлена на рис. 2.1.

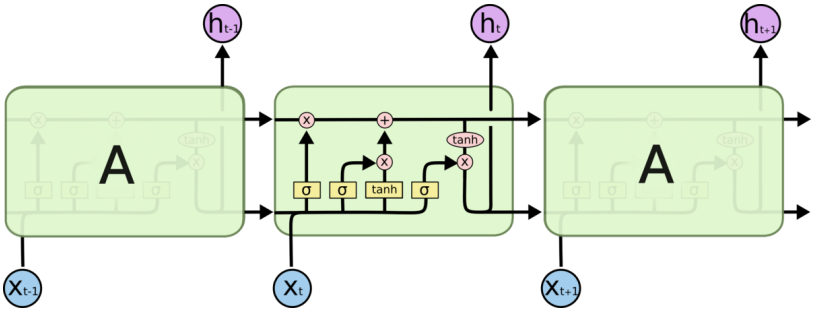

Рисунок 2.1 – Архитектура LSTM

Ключевыми компонентами LSTM-модуля являются **состояние ячейки** и **фильтры**, контролирующие его (входной, выходной и фильтр забывания). В процессе обучения состояние ячейки изменяется, информация добавляется или удаляется оттуда структурами, называемыми фильтрами. Фильтры контролируют поток информации на входах и на выходах модуля на основании некоторых условий. Они состоят из слоя сигмоидальной нейронной сети и операции поточечного умножения.

Сигмоидальный слой возвращает числа в диапазоне [0; 1], которые обозначают, какую долю каждого блока информации следует пропустить дальше по сети. Умножение на это значение необходимо для пропуска или запрета потока информации внутрь и наружу памяти. Например, входной фильтр контролирует меру вхождения нового значения в память, а фильтр забывания контролирует меру сохранения значения в памяти. Выходной фильтр контролирует меру того, в какой степени значение, находящееся в памяти, используется при расчёте выходной функции активации.

3 Анализ данных

3.1 Описание датасета

Исследуемый набор данных Jena Climate Dataset содержит данные временных рядов погодных данных, записанных на метеостанции Института биогеохимии Макса Планка в Йене, Германия. Информация записывалась каждые 10 минут в течение нескольких лет. Этот датасет охватывает данные с 1 января 2009 года по 31 декабря 2016 года.

3.2 Предварительная обработка

После предобработки данных был построен временной ряд для показателей температуры воздуха в 2009 году (рис. 3.1) и было выявлено следующее:

— Тренд явно не выражен, визуально нельзя выделить общий восходящий или нисходящий тренд в течение года. Очевидно, это связано с общими погодными закономерностями на нашей планете.

— На графике заметны некоторые сезонные колебания. Например, можно заметить, что температура ниже в начале и конце года (зима), и выше в середине года (лето). Это соответствует сезонным изменениям температуры в умеренных климатических зонах.

— График показывает значительные колебания температуры, которые могут быть обусловлены как сезонными факторами, так и случайными выбросами. В данных присутствуют резкие изменения и скачки, что может указывать на присутствие случайных факторов.

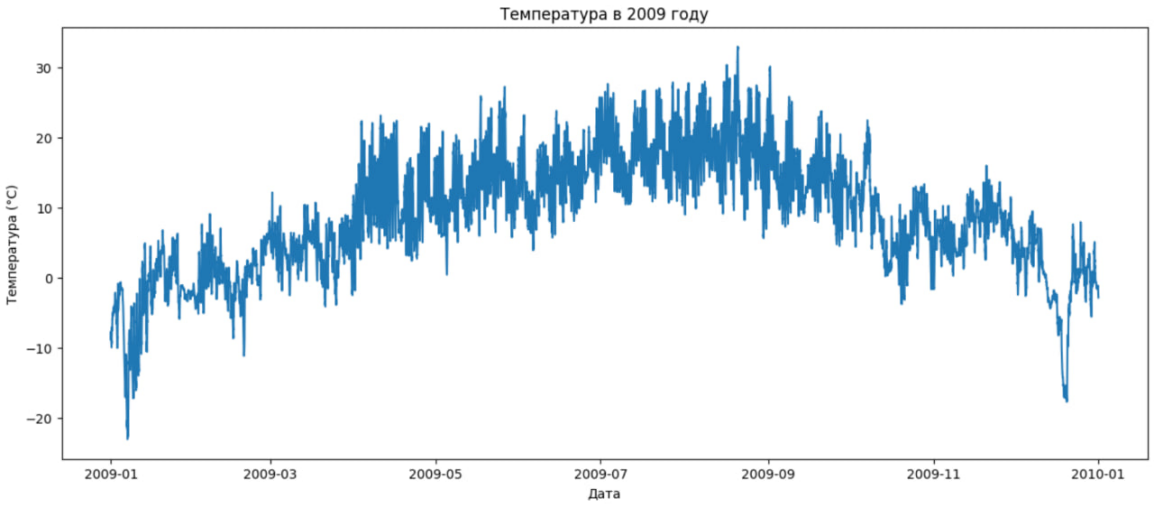

Рисунок 3.1 – Временной ряд для показателей температуры воздуха в Йене 2009 году

Рассматриваемый временной ряд, очевидно, является регулярным, так как временные точки равномерно распределены (данные записаны с постоянным интервалом времени между ними), и не является детерминированным, так как имеющуюся информацию нельзя представить в виде формулы. На основе результатов теста Дики-Фуллера был сделан вывод о том, что временной ряд является стационарным.

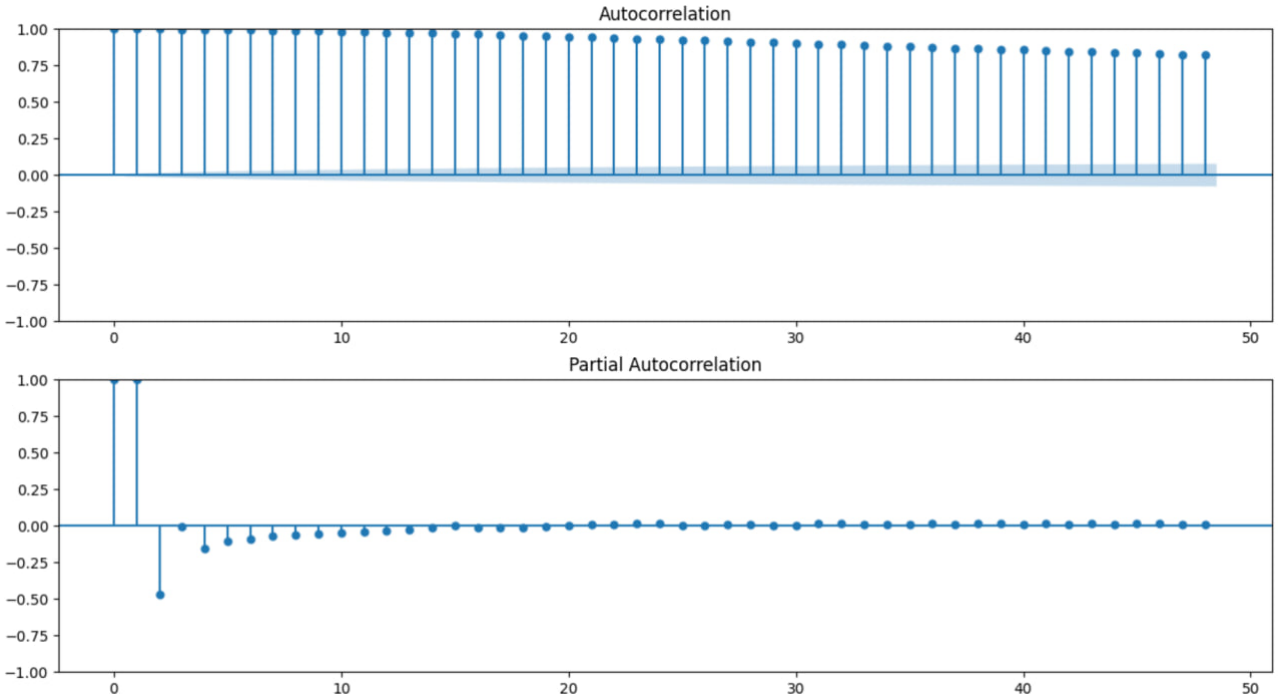

Рисунок 3.2 – Графики ACF и PACF для исследуемого временного ряда

График ACF показывает медленный убывающий тренд, что свидетельствует о сильной сезонности в данных. Значения автокорреляции остаются высокими на протяжении большого количества лагов, что указывает на необходимость учета сезонного компонента.

PACF имеет значительное значение на лаге 1 и резко падает на более высоких лагах, что указывает на возможное наличие AR компонента (авторегрессия) в модели. Также видно повторение периодичности, что снова указывает на сезонность в данных.

4 Реализация и сравнение методов обнаружения аномалий

4.1 Модель SARIMA

Для наилучшей работы требуется подобрать параметры для модели $SARIMA(p, d, q)(P, D, Q, s)$. В результате анализа свойств исследуемого временного ряда и графиков ACF и PACF были выбраны следующие параметры для модели SARIMA: $SARIMA(1, 1, 1)(1, 1, 1, 365)$.

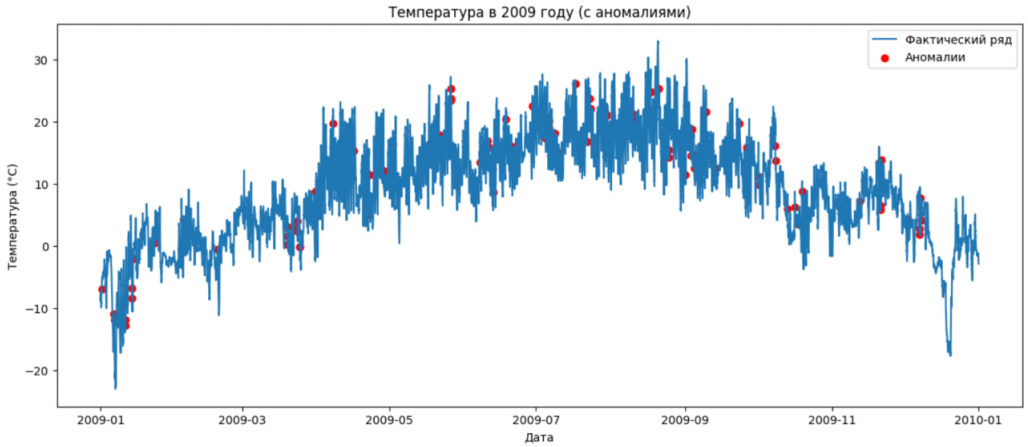

Рисунок 4.1 – Временной ряд для показателей температуры воздуха в 2009 году с предсказанными аномалиями с помощью SARIMA

Из графика видно, что аномалии распределены по всему году, но их количество и частота варьируются. Большинство аномалий происходит в периоды быстрых изменений температуры или резких колебаний, что может быть связано с необычными погодными условиями или событиями. Выявленные аномалии (красные точки) представляют собой данные, которые существенно отклоняются от модели. Эти аномалии скорее всего являются результатом экстраординарных погодных явлений.

Теперь построим такой же график, но для каждого месяца 2009 года. Это нужно для более детального анализа прогнозируемых аномалий.

Рассмотрим некоторые месяцы: февраль (рис. 4.2) и декабрь (рис. 4.3). Из графика для февраля видно, что модель успешно предсказывает некоторые плюсовые значения, что является аномалией, так как плюсовые значения температуры зимой не характерны для данного региона. Что касается декабря, модель отлично справилась с предсказанием аномально высокой температуры (примерно 9 декабря на графике).

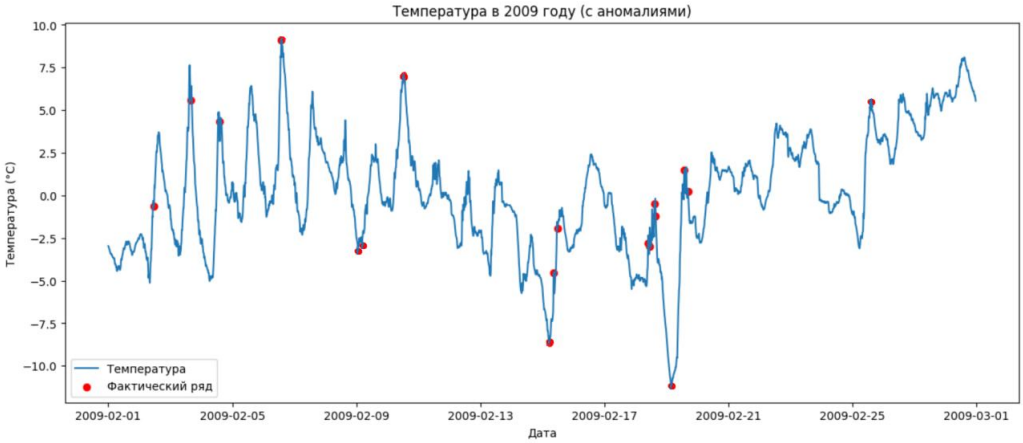

Рисунок 4.2 – Временной ряд для показателей температуры воздуха в феврале 2009 года с предсказанными аномалиями с помощью SARIMA

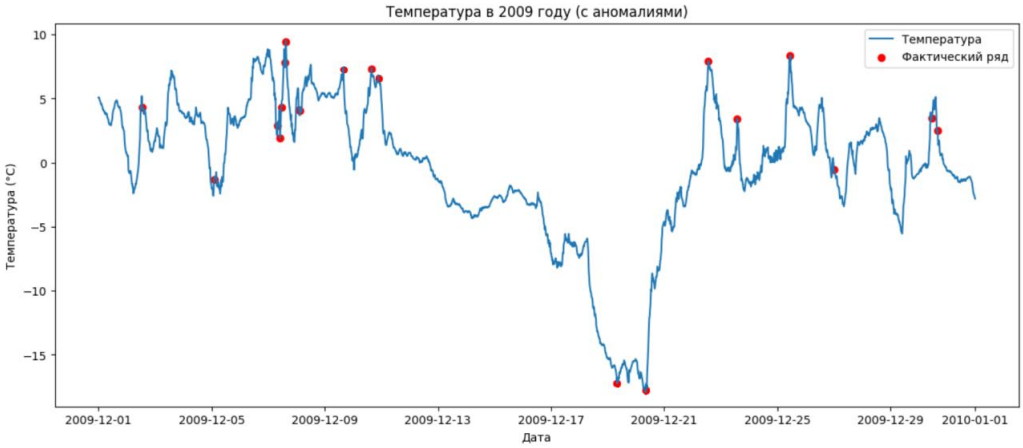

Рисунок 4.3 – Временной ряд для показателей температуры воздуха в декабре 2009 года с предсказанными аномалиями с помощью SARIMA

Оценим качество прогнозов модели SARIMA с использованием сравнения фактических и прогнозируемых значений, а также расчета метрик оценки прогнозов. MSE равно 114.81. В контексте данной модели прогнозирования температуры это значение выше среднего, что указывает на некоторое расхождение между фактическими и прогнозируемыми значениями. MAE равно 17.78. Это значение также выше среднего для данного контекста. Значение R-квадрат равно 0.6. Это говорит о том, что модель объясняет значительную часть вариации целевой переменной, но есть еще место для улучшений в точности предсказаний.

На основании полученных результатов можно сделать вывод, что модель хорошо захватывает сезонные и несезонные компоненты данных о температуре за 2009 год и неплохо справляется с предсказанием ожидаемых и типичных аномалий. Однако модель не захватила всю структуру данных и требуется дообучение. Необходимо учесть тот факт, что для обучения модели требуется

большое количество времени и высокие мощности, обладая которыми есть возможность улучшить показатели.

4.2 Модель LSTM

Теперь перейдём к построению модели LSTM. Точно также загрузим данные и осуществим их предобработку, как и в случае с SARIMA. Далее необходимо нормализовать данные. Модель будет обучалась 5 эпох (это количество полных прохождений всех сетов), размер батча — 32 (это количество образцов, проходящих через модель за одну итерацию обновления весов).

Среднеквадратичная ошибка (MSE) равна 0.044, средняя абсолютная ошибка (MAE) равна 0.134. Эти значения довольно низкие, что говорит об очень хорошем результате обучения. Коэффициент детерминации (R^2) равен 0.999. Модель также очень хорошо объясняет вариацию целевой переменной и дает очень хорошие прогнозы.

В целом, эти метрики свидетельствуют о том, что модель обучена и оценена очень хорошо. Она демонстрирует высокую точность в прогнозировании температурных данных и обладает хорошей способностью к обобщению на новые данные.

Построим график, на котором отразим сам временной ряд и выявленные аномалии (такой же, какой был построен для модели SARIMA) (рис. 4.4).

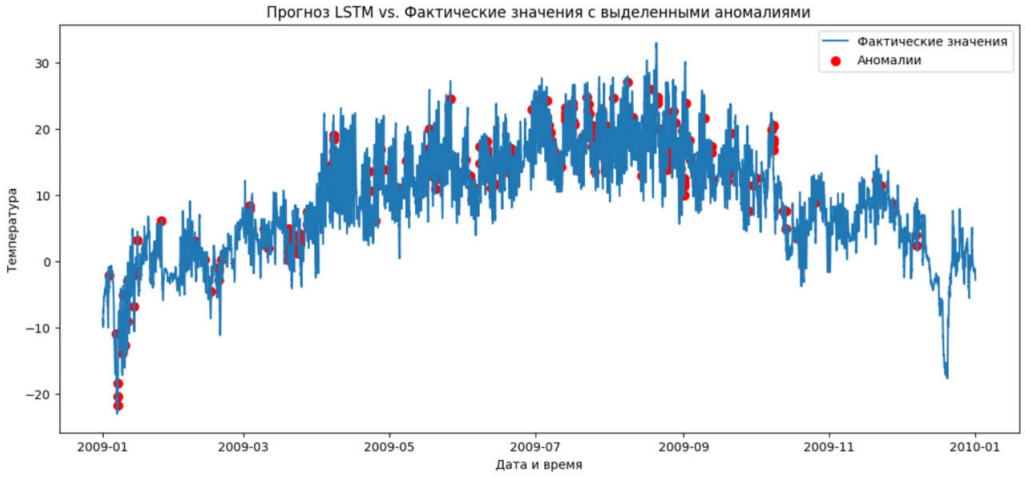

Рисунок 4.4 – Временной ряд для показателей температуры воздуха в 2009 году с предсказанными аномалиями с помощью LSTM

Рассмотрим также графики для февраля (рис. 4.5) и декабря (рис. 4.6).

Рисунок 4.5 – Временной ряд для показателей температуры воздуха в феврале 2009 года с предсказанными аномалиями с помощью LSTM

Рисунок 4.6 – Временной ряд для показателей температуры воздуха в декабре 2009 года с предсказанными аномалиями с помощью LSTM

На обоих графиках аномалии визуально совпадают с резкими изменениями температуры, что указывает на то, что модель правильно идентифицировала внезапные отклонения от тренда. Это свидетельствует о хорошей способности модели выявлять необычные события в данных.

В феврале модель обнаружила аномалии в различных точках, которые соответствуют экстремальным значениям температуры. В декабре также видны несколько аномалий, которые соответствуют резким падениям и подъемам температуры. Это показывает, что модель эффективно справляется с обнаружением как положительных, так и отрицательных аномалий. Также стоит отметить, что модель LSTM хорошо справилась с выявлением не только точечных, но и групповых аномалий, что также вытекает из графиков.

4.3 Интерпретация результатов

В ходе построения, анализа и сравнения моделей SARIMA и LSTM для выявления аномалий в климатических данных было обнаружено, что модель LSTM демонстрирует значительное превосходство по ряду важных показателей. LSTM показала лучшие результаты по точности прогнозов, что выражается в значительно меньших значениях среднеквадратичной ошибки (MSE) и средней абсолютной ошибки (MAE). Кроме того, модель LSTM показала гораздо более высокое значение коэффициента детерминации (R-squared), которое близко к 1. Это свидетельствует о том, что LSTM объясняет почти всю дисперсию исходных данных.

Важным аспектом, который также следует учитывать, является время обучения моделей. Обучение модели SARIMA заняло значительно больше времени по сравнению с LSTM и потребовало больше вычислительной мощности. Это связано с тем, что SARIMA требует оценки большого числа параметров и учет сезонных компонентов, что делает процесс обучения весьма ресурсоемким. В то время как модель LSTM, несмотря на свою сложность в архитектуре по сравнению с SARIMA, характеризуется более быстрой сходимостью и требует меньше времени на обучение. Тем не менее, необходимо отметить, что при наличии достаточных ресурсов возможно достичь лучших показателей для SARIMA.

Учитывая все вышеперечисленные факторы, можно сделать вывод, что в контексте ограниченности временных и вычислительных ресурсов модель LSTM является более подходящей и эффективной для задачи выявления аномалий в климатических данных. Она демонстрирует лучшие результаты по точности прогнозов, быстрее обучается и лучше справляется с сложными зависимостями в данных.

ЗАКЛЮЧЕНИЕ

В ходе данной работы была проведена комплексная исследовательская работа по сравнению методов обнаружения аномалий во временных рядах с использованием технологий машинного обучения, а именно моделей SARIMA и LSTM.

Целью работы было исследование и сравнение различных методов обнаружения аномалий во временных рядах. Для достижения этой цели было выполнено следующее:

- рассмотрены понятия временного ряда и аномалии во временных рядах, а также их классификацию и особенности;
- проведён обзор существующих методов обнаружения аномалий в данных;
- реализован метод SARIMA для обнаружения аномалий;
- реализован метод LSTM для обнаружения аномалий;
- методы применены к реальному набору данных;
- сделаны выводы об эффективности построенных моделей на основании полученных результатов.

SARIMA, будучи статистическим методом, хорошо справляется с регулярными и сезонными паттернами, однако может быть ограничена в способности обнаруживать сложные и нелинейные аномалии и требует очень много времени для обучения. LSTM, основанная на рекуррентных нейронных сетях, продемонстрировала высокую гибкость и способность выявлять сложные зависимости в данных, что позволяет более эффективно обнаруживать аномалии, особенно в данных с нелинейными зависимостями.

Сравнительный анализ показал, что выбор метода зависит от конкретных характеристик временного ряда и требований к точности и скорости обнаружения аномалий. В случае, где данные содержат сложные зависимости и важна высокая точность, методы на основе LSTM являются более эффективными.

Таким образом, данная работа вносит вклад в область анализа временных рядов и обнаружения аномалий, предоставляя сравнительный анализ двух подходов и рекомендации по их применению. Будущие исследования могут быть направлены на интеграцию и комбинирование различных методов для улучшения качества обнаружения аномалий и адаптации к специфическим требованиям и применены в области анализа данных и машинного обучения для задач мониторинга и предсказания временных рядов.