

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»
Институт радиоэлектроники и информационных технологий–РТФ
Школа профессионального и академического образования

ДОПУСТИТЬ К ЗАЩИТЕ ПЕРЕД ГЭК

Директор ШПиАО
 Д.В. Денисов
(подпись) (Ф.И.О.)
« 03 » июня 2024 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ПРЕДСКАЗАНИЕ ПОГОДЫ НА ОСНОВЕ ДАННЫХ С МЕТЕОПОСТОВ
(НА ПРИМЕРЕ МЕТЕОСТАНЦИИ В Г. МИНСКЕ)

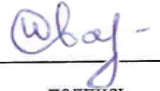
Научный руководитель: кандидат физико-математических наук, доцент
кафедры ИТиСУ ИРИТ-РТФ Юманова И.Ф.


подпись

Нормоконтролер: Огуренко Егор Владимирович


подпись

Студент группы: РИМ-220962 Иванов Дмитрий Владимирович


подпись

Екатеринбург
2024

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего
образования

«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РТФ
Школа профессионального и академического образования
Направление подготовки 09.04.01 Информатика и вычислительная техника
Образовательная программа 09.04.01/33.03 Инженерия машинного обучения

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студента Иванова Дмитрия Владимировича группы РИМ-220962.
(фамилия, имя, отчество)

1. Тема выпускной квалификационной работы ПРЕДСКАЗАНИЕ ПОГОДЫ НА
ОСНОВЕ ДАННЫХ С МЕТЕОПОСТОВ (НА ПРИМЕРЕ МЕТЕОСТАНЦИИ В
Г. МИНСКЕ)

Утверждена распоряжением по институту от «4» декабря 2023 г. № 33.02-05/298

2. Научный руководитель к. физ.-мат. наук, доцент Юманова Ирина Фарисовна

(Ф.И.О., должность, ученая степень, ученое звание)

3. Исходные данные к работе набор данных о погоде в г. Минске

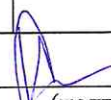
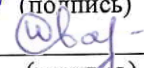
4. Перечень демонстрационных материалов презентация

5. Календарный план

№ п/п	Наименование этапов выполнения работы	Срок выполнения этапов работы	Отметка о выполнении
1.	Сбор данных для разработки модели МО	до 23.03.2024 г.	
2.	Выбор методов исследования	до 29.04.2024 г.	
3.	Применение методов исследования на наборе данных	до 20.05.2024 г.	
4.	Систематизация полученных результатов	до 24.05.2024 г.	

Научный руководитель И.Ф. Юманова
Ф.И.О.

Студент задание принял к исполнению 12.02.24
дата


(подпись)

(подпись)

6. Консультанты по проекту (работе) с указанием относящихся к ним разделов*

7. Допустить Иванова Дмитрия Владимировича к защите выпускной квалификационной
работы в экзаменационной комиссии

Директор ШПиАО


(подпись)

Д.В. Денисов
Ф.И.О.

РЕФЕРАТ

Магистерская диссертация 51 с., 23 рис., 3 табл., 40 источников.

Ключевые слова: машинное обучение, временные ряды, наивные предсказания, сезонные предсказания.

Объект исследования: технологический процесс для решения задачи по долговременному предсказанию температуры воздуха.

Предмет исследования: методы и алгоритмы автоматического построения прогноза температуры воздуха, поиск закономерностей.

Цель работы: исследование эффективности моделей для предсказания температуры воздуха.

В процессе исследования проводились: рассмотрение основных подходов построения моделей по краткосрочному и долгосрочному предсказанию температуры воздуха, настройка и реализация моделей для предсказания изменения температуры воздуха.

В работе продемонстрирована эффективность подхода использования машинного обучения в построении моделей для предсказания температуры воздуха.

Область практического применения: предложенный подход поможет таким сферам, как: сельское хозяйство, энергетика, строительство, транспорт, торговля и др. Снизит затраты на подготовку к сезонным мероприятиям и поможет точнее планировать финансовые потоки.

СОДЕРЖАНИЕ

РЕФЕРАТ.....	3
ВВЕДЕНИЕ	5
Глава 1 Анализ предметной области.....	9
1.1 Метеорологические исследования и прогноз погоды	9
1.2 Обзор аналогов	11
1.3 Подготовка и анализ данных	15
1.4 Оценка эффективности предсказания	17
2. Анализ данных	20
2.1 Описание датасета.....	21
2.2 Разведывательный анализ данных.....	23
2.3 Разложение временных рядов.....	29
Глава 3 Построение моделей и интерпретация результатов.....	31
3.1 Построение моделей.....	31
3.2 Интерпретация результатов	40
ЗАКЛЮЧЕНИЕ	44
БИБЛИОГРАФИЧЕСКИЙ СПИСОК	46

ВВЕДЕНИЕ

В развитии метеорологии прогнозирование погоды выделяется как одно из наиболее важных направлений [1]. Точность прогнозов имеет решающее значение для эффективного управления чрезвычайными ситуациями, смягчения последствий экономических потерь и содействия получению доходов в ключевых секторах, таких как энергетика, сельское хозяйство и транспорт [2].

Прогноз погоды основан на сборе данных о текущем состоянии атмосферы, таких как температура, влажность, атмосферное давление, скорость и направление ветра, а также на исторических данных. Эти данные вводятся в погодные модели, которые используются для прогнозирования будущих погодных условий [3].

В 1854 году была основана Британская метеорологическая служба, которая стала одной из первых организаций, занимающихся прогнозированием погоды [4]. К середине XX века появились первые компьютерные модели прогнозирования погоды. В последнее время метеорология полностью зависит от передовых компьютерных технологий и сетей наблюдений. Спутники и сенсоры позволяют наблюдать за погодой в реальном времени по всему миру [5].

Глобальное потепление оказывает большое влияние на жизнедеятельность человека [6]. Быстрое развитие технологий искусственного интеллекта, особенно в последние годы, может помочь в построении долгосрочного прогноза среднегодовой температуры воздуха. Модели машинного обучения учитывают текущие и исторические данные о погоде, атмосферных условиях, океанических течениях и других факторов, влияющих на изменение температуры воздуха. При помощи тонкой настройки моделей для предсказания температуры можно постоянно улучшать прогноз, чтобы он становился все более точным и надежным [7].

Цель исследования — построение модели для долгосрочного предсказания температуры воздуха при помощи машинного обучения, а также выбор оптимального периода наблюдений для работы модели.

Чтобы достичь поставленной цели следует решить следующие задачи:

1. Изучение методов машинного обучения и предметной области, которая связана с технологическими процессами, анализ оптимальных параметров для входных данных. Выявление эффективных методов для предсказания температуры воздуха.

2. Изучение и построение модели предсказания температуры воздуха, а также интерпретация результатов.

3. Написание программного кода обучения модели, корреляция признаков для отбора наиболее значимых данных.

4. Сравнение различных моделей, анализ полученных результатов и оптимизация параметров. Определение причин возникновения аномальных результатов на основе методов машинного обучения.

5. Интерпретация результатов. После оценки эффективности моделей необходимо проанализировать полученные результаты и интерпретировать их в рамках темы данной работы. Также будет произведено сравнение с классическими методами предсказания погоды, если на краткосрочный промежуток времени (до одной недели), результаты показывают высокую точность, то на долгосрочный промежуток результаты редко совпадают с реальными.

Методы исследования основываются на фундаментальных положениях анализа данных, теории смены климатических явлений, теории временных рядов, машинном обучении.

Объектом исследования является процесс создания модели машинного обучения для предсказания температуры воздуха.

Предметом исследования является создание модели, основанной на анализе временных рядов с метеопоста в г. Минске, для предсказания температуры воздуха.

Исследования по долгосрочному прогнозу температуры воздуха являются актуальными по нескольким причинам:

- Планирование и подготовка к сезонным изменениям: долгосрочные прогнозы помогают заранее подготовиться к предстоящим сезонным температурным изменениям, например, к наступлению холодного периода. Это позволяет лучше спланировать работу коммунальных служб, сельского хозяйства, энергетики и других отраслей.

- Адаптация к изменению климата: долгосрочные температурные тренды отражают происходящие климатические изменения. Анализ этих трендов важен для разработки стратегий адаптации к меняющимся климатическим условиям.

- Управление рисками и принятие решений: достоверные прогнозы температуры на сезонном и годовом уровне помогают лицам, принимающим решения, лучше оценивать климатические риски и принимать более обоснованные управленческие решения в различных сферах.

- Научные исследования: данные о долгосрочной динамике температуры необходимы для климатических моделей, изучения глобальных и региональных тенденций изменения климата, а также для оценки воздействия климатических изменений на окружающую среду и общество.

Структура диплома состоит из введения, главных разделов, заключения, приложения и списка использованных источников. Во введении определены цели и задачи исследования, а также их актуальность. Основные главы содержат информацию о методах обработки данных, их интерпретацию и оценку эффективности их прогнозирования. В заключении будут подведены итоги исследования и сделаны выводы об эффективности рассмотренных методов машинного обучения в области предсказания погоды. В списке использованной литературы будут перечислены все книги, научные статьи и другие источники информации, на которые ссылается данная работа.

Актуальность данной темы состоит в необходимости поиска алгоритмов долгосрочного прогноза погоды (квартал, год), так как краткосрочные прогнозы зависят от большого количества факторов (облачность, сила и направление ветра и др.). Долгосрочных прогноз является важной задачей для сельского хозяйства, коммунальных служб. Данная работа должна помочь сельскохозяйственным организациям помочь подготовиться к посевным и уборочным кампаниям, ведь надо четко понимать, в какой временной промежуток следует начинать засев полей, чтобы исключить вероятность заморозков и гибели будущего урожая. Коммунальные службы смогут составить смету для подготовки необходимого количества техники для уборки территории. Строительные предприятия смогут лучше спланировать график строительства, который зависит от применения строительных смесей в зависимости от погодных условий.

Научная новизна исследования состоит в создании долгосрочном прогнозе температуры воздуха с применением современных моделей машинного обучения.

Глава 1 Анализ предметной области

1.1 Метеорологические исследования и прогноз погоды

Вся жизнь в природе непрерывно связана с метеорологическими процессами, которые оказывают влияние на всю жизнедеятельность человека. Она охватывает множество областей, таких как: физическая метеорология, динамическая метеорология, синоптическая метеорология, климатология, аэрология. Без метеорологической информации не могут обойтись морской транспорт, авиация, военная сфера; от погодных условий зависит сельское хозяйство, коммунальные службы города, строительные предприятия, энергетические компании, торговля и т.д.

В метеорологических исследованиях измеряется большое количество данных (температура воздуха, почвы, воды, силы и направления ветра, продолжительности солнечной активности, количества осадков, облачности и др.), а, учитывая тот факт, что данные необходимо снимать каждые три часа [8] массив дневных данных увеличивается в восемь раз. Еще в XIX веке была обнаружена закономерность в появлении течения в Тихом океане под названием Эль-Ниньо [9], так же был доказан 11-летний цикл солнечной активности «цикл Швабе» [10].

На данный момент существует два основных научных подхода к прогнозированию погоды: синоптический и численный (гидродинамический).

Синоптический метод основывается на построении и анализе синоптических карт, где выделяются различные погодные явления, такие как циклоны, антициклоны и атмосферные фронты, каждому из которых соответствуют определенные типы погодных условий. Прогноз в данном случае имеет вероятностную природу, поскольку в его основе лежат закономерности, установленные математической статистикой. Для создания синоптических карт требуется сбор метеорологических данных из множества

источников, которые кодируются в соответствии с международным метеокодом КН-01.

Гидродинамический метод предполагает построение математических моделей атмосферных процессов и взаимодействия океана с атмосферой. Он основывается на применении уравнений гидро- и термодинамики, а также фундаментальных физических законов, таких как законы сохранения количества движения, массы и энергии, а также газовые законы. Прогнозирование погоды в этом случае осуществляется с помощью вычислительной техники. Использование методов машинного обучения сопряжено с некоторыми проблемами, которые стоит учитывать. В первую очередь необходимо качественно подготовить исходные данные (убрать пропуски и ошибочные данные). Также ошибки могут заключаться в обучении модели, которые исказят итоговый результат. К числу таких ошибок можно отнести отсутствие в результатах тренда и сезонности. Поскольку сейчас идет период глобального потепления [10] этот момент также должен отражаться.

Главной целью данной работы является поиск оптимального метода анализа временных рядов для получения наиболее точного прогноза на основе анализа данных датасета на 40000 строк. Будут различные методы машинного обучения, такие как: Sarima, Prophet, Sarimax.

Таким образом, предсказание температуры является важным для многих отраслей, где погодные условия играют существенную роль. Например:

- Сельское хозяйство: точный прогноз температуры помогает предприятиям аграрного сектора планировать посевные кампании, определить оптимальные сроки для уборки урожая, контролировать посев и принимать меры для защиты посевов от серьезных природных катаклизмов, таких как град, засуха или наводнение;

- Энергетика: прогноз погоды важен для планирования производства электроэнергии и подготовке к осенне-зимнему периоду эксплуатации

отопления. Знание будущих погодных условий позволяет оптимизировать работу энергетических систем, нивелировать увеличение потребления электрической энергии и отопления;

– Транспорт: в первую очередь речь идет о гражданском транспорте и предупреждении о плохих погодных условиях, таких как гололед и туман. Предупреждение об этих условиях поможет дорожным организациям заранее принимать меры для обеспечения безопасности пассажиров и грузов, а также планировать маршруты и расписание. Во вторую очередь — морской и речной транспорт. Образование ледяного покрова на реках и морях оказывает большое влияние на грузоперевозки. В последние годы идет увеличение грузоперевозок по северному морскому пути, что часто сопряжено с использованием ледоколов. Долгосрочный прогноз температуры воздуха поможет лучше использовать гражданский атомный флот;

– Торговля: смена погоды имеет влияние на потребительское поведение и спрос сезонных товаров. Правильный прогноз погоды помогает розничным кампаниям управлять запасами и вовремя проводить закупки;

– Строительство: прогноз погоды важен для планирования и строительства объектов;

Это лишь некоторые примеры отраслей, где прогноз погоды имеет большое значение. Общая идея заключается в том, что знание будущих погодных условий позволяет принимать информированные решения, оптимизировать процессы и минимизировать риски, связанные с воздействием погоды на различные отрасли деятельности и общества, что положительно скажется на финансовой деятельности государства в целом.

1.2 Обзор аналогов

Обзор аналогов позволяет оценить существующие варианты и сделать научно-обоснованный выбор на основе информации о преимуществах, недостатках и особенностях каждого метода.

В работе «Использование анализа временных рядов в изучении многолетних температурных изменений» [7] целью исследования было рассмотрение методического подхода к анализу многолетних температурных рядов и выявление структурных составляющих для построения моделей.

Автором были собраны многолетние данные о температурном режиме в городе Псков. При помощи многомерной регрессии удалось найти тренд и циклические составляющие. Модель была построена при помощи программы Statistica 8.0, в которой была выполнена процедура сезонной декомпозиции.

Недостатком данного исследования является отсутствие информации о выбранном промежутке времени и общем количестве данных, на основе которых строилась модель. Несмотря на присутствие этапа обработки данных, не описан процесс анализа данных, отсутствует информация о том, какая работа проводилась с пропущенными значениями и данными (выбросами).

У автора данного исследования получилось создать модель для анализа временных рядов в изучении температурных изменений, которая описывает общие закономерности изменения временного ряда, не предсказывая экстремальные колебания температур.

В Европе прогнозами погоды занимается Европейский центр среднесрочных прогнозов погоды (ECMWF). Все оперативные прогнозы составляются с помощью интегрированной системы прогнозирования (IFS). Она включает в себя сложную систему усвоения данных и глобальную числовую модель системы Земли, а также вспомогательную инфраструктуру для предоставления прогностической продукции государствам-членам ECMWF.

Система ассимиляции данных объединяет последние наблюдения за погодой с недавним прогнозом для получения максимально возможной оценки текущего состояния системы Земли – так называемого анализа.

В основе модели системы Земли лежит модель атмосферы, соединенная с моделями других компонентов системы Земли, таких как

океан. Начиная с анализа, модель системы Земли затем используется для прогнозирования вероятного развития погоды и некоторых других аспектов системы Земли [11].

Летом 2023 года также начата разработка модели машинного обучения AIFS . Эта модель в настоящее время находится на стадии бета-тестирования, и пока доступны только экспериментальные прогоны.

Долгосрочный (сезонный) прогноз использует немного другую физику модели и основан на более старой версии IFS. Он работает с более низким разрешением (в настоящее время 36 км), каждый параметр которого глобально возмущен с использованием сингулярных векторов, стохастической физики, ансамбля ассимиляции данных (EDA) и некоторых поверхностных возмущений начального состояния (например, влажности почвы). Сезонный прогноз продлевается четыре раза в год (в феврале, мае, августе и ноябре) до 13 месяцев [12].

14 ноября 2023 года команда Google Deepmind представила свою модель предсказания температуры — GraphCast [13]. GraphCast — это современная модель искусственного интеллекта, способная делать среднесрочные прогнозы погоды с беспрецедентной точностью (0.25 °C на квадрат с расстоянием 28x28 км).

GraphCast может заранее предупреждать об экстремальных погодных явлениях, также с большой точностью предсказывать следы циклонов, определять атмосферные явления, связанные с риском наводнений, и предсказывать наступление экстремальных температур.

Особенность данной модели состоит в том, что она рассчитывает погоду без понимания физики атмосферных процессов. Инженеры обучали нейросеть на архивах погоды за период с 1979 по 2017 годы. Данный метод прогнозирует сотни погодных переменных на следующие 10 дней с разрешением 0.25°C по всему миру менее, чем за одну минуту.

Компания Huawei представила свою модель предсказания погоды под названием «Pangu-Weather». Система основана на глубоком обучении и

использует архитектуру 3D-трансформера. Ключевая особенность состоит в том, что сеть минимизирует функцию потерь, определяемую, как среднеквадратическая ошибка (RMSE) с взвешиванием по cos-широте, чтобы учитывать сферическую форму Земли, поскольку модель обучается на регулярной сетке широты и долготы [14].

На девятой интернациональной конференции молодых ученых была освещена тема «Прогнозирование температуры воздуха на метеостанции с использованием глубоких нейронных сетей» [15]. В своей работе автор использовал три модели глубоких нейронных сетей: многослойный перцептрон (MLP), сеть долговременной краткосрочной памяти (LSTM) и сверточную нейронную сеть (CNN) + LSTM [15]. Входные данные о погоде брались за временной промежуток с 2009 по 2019 гг. и включали такие параметры, как средняя скорость ветра, количество осадков, высоту снежного покрова, средняя, максимальная и минимальная температура воздуха (данные параметры автор выбрал на основе корреляции). Прогноз погоды на один день вперед показал неплохой результат: для модели CNN+LSTM автору удалось добиться точности в 97.42%, для моделей MLP и LSTM — 89.15% и 95.03% соответственно. Однако, при прогнозе погоды на 10 дней вперед, точность серьезно падает и составляет от 63.23% до 71.58%. В дальнейшем автор предполагает усложнение модели предсказания, добавления в датасет новых параметров, таких как атмосферное давление, относительную влажность, направление и скорость ветра.

В другой работе [16] индийские исследователи предложили использовать модель случайного леса. Входными параметрами для построения модели машинного обучения были выбраны такие показатели, как средняя и максимальная температуры, атмосферное давление, относительная влажность, скорость ветра, широта и долгота. В результате Decision Tree показало точность модели равную 89%, Multi Layer Perceptron — 90%, а RF-PSO — 97%.

В Республике Беларусь метеорологические прогнозы разрабатываются организацией «Белгидромет». Для составления прогнозов на 1-3 дня используются данные не только с наземных метеостанций, но и из всей толщи атмосферы до высоты 30 км. Для краткосрочного прогноза на ближайшие сутки требуется информация о климатических условиях по всей Европе, а для трехдневного прогноза - по всему Северному полушарию [17].

Численные расчеты будущего состояния атмосферы выполняет синоптик-специалист, который анализирует фактические и прогностические погодные карты у земной поверхности и на высотах. Он определяет источники воздушных масс, их циркуляцию в барических системах, выявляет атмосферные фронты и прогнозирует их перемещение. Точность краткосрочных прогнозов достигает 98% [18].

В таблице 1 представлены модели предсказания температуры воздуха и их точность. Модели с ансамблевым методом лучше подходят для краткосрочных прогнозов температуры воздуха.

Таблица 1 – Сравнение точности моделей предсказания

Модель предсказания	Точность, %
CNN+LSTM	97.42
MLP	89.15
LSTM	95.03
Decision Tree	89
Multi Layer Perceptron	90

1.3 Подготовка и анализ данных

Подготовка и анализ данных играют важную роль при анализе временных рядов для предсказания температуры воздуха. Данный этап включает в себя очистку данных, которые будут использованы в работе.

На первом этапе были собраны данные о температуре воздуха на метеостанции, расположенной в г. Минске, за 2011-2022 гг. Замер данных

производится ежедневно каждые три часа [8], в итоговом датасете получилось около 40 000 записей.

После сбора данных была произведена их очистка. Этот этап включал в себя удаление дубликатов, заполнение пропущенных значений, шумов и исправление ошибок в данных. Это может быть осуществлено с помощью различных методов, таких как замена пропущенных значений на среднее или медиану.

При работе с отсутствующими данными подход зависит от типа пропусков: они могут быть полностью случайными, случайными или неслучайными [19].

MCAR (Missing Completely at Random). Когда пропуски полностью случайные, это означает, что подвыборка наблюдений является репрезентативной выборкой из генеральной совокупности. Предположение о полной случайности пропусков - единственное допущение, которое можно проверить эмпирически.

MAR (Missing at Random). Случайные пропуски имеют место, когда вероятность пропуска зависит от значений других переменных, но не от самих пропущенных значений. В этом случае возможно смещение оценок параметров.

MNAR (Missing Not at Random). Неслучайные пропуски возникают, когда вероятность пропуска зависит от значений самих пропущенных данных. Такие пропуски обязательно вносят систематические ошибки в результаты анализа.

Методы обработки пропущенных данных делятся на традиционные и современные. К традиционным относятся удаление наблюдений, замещение средним и регрессионное замещение. К современным - оценка максимального правдоподобия, множественное замещение, селективная модель и модель смешанных паттернов [20].

Но, поскольку изменение температуры воздуха не всегда происходит линейно, то такие значения было решено удалить из датасета.

Было выполнено преобразование данных, которое включало в себя приведение всех значений к одному формату. Так, в колонке направления и скорости ветра было удалено значение порывов ветра и осталось только среднее значение данного показателя. Категориальные колонки были преобразованы в числовые (направление ветра, комфортность).

Для оценки и проверки эффективности моделей датасет был разделен на тренировочную и тестовую выборку в соотношении 70:30. Обучающий набор используется для обучения модели, а тестовый набор — для финальной оценки эффективности модели.

Визуализация данных помогла понять распределение данных, взаимосвязи между признаками и выявить аномалии или интересные закономерности. Это помогло принять решения о предобработке данных и выборе модели.

1.4 Оценка эффективности предсказания

Оценка эффективности моделей машинного обучения является важным этапом в процессе разработки и применения этих моделей. Существует несколько показателей и методов оценки, которые позволяют определить, насколько хорошо модель справляется с поставленной задачей. В рамках данной работы будут использованы следующие метрики:

Точность (Accuracy): Это наиболее распространенная метрика оценки моделей регрессии, которая измеряет долю правильных предсказаний модели относительно общего числа предсказаний. Однако точность может быть непоказательной, если данные не сбалансированы, то есть классы имеют разное количество примеров.

Кривая ROC (Receiver Operating Characteristic): Эта кривая используется для оценки моделей классификации. Она позволяет оценить компромисс между полнотой и специфичностью модели в зависимости от выбранного порога вероятности классификации.

Симметричная средняя процентная ошибка (sMAPE):

$$sMAPE = \frac{1}{H} \sum_{i=1}^H \frac{|y(h_i) - \bar{y}(h_i)|}{|y(h_i)| + |\bar{y}(h_i)|}$$

где:

H – количество наблюдений;

$y(h_i)$ – фактическое значение в момент времени h_i ;

$\bar{y}(h_i)$ – прогнозное значение в момент времени h_i .

Метрика MAPE (Mean Absolute Percentage Error) используется для оценки качества прогнозных моделей. Она показывает среднее абсолютное процентное отклонение предсказанных значений от истинных. Ее особенности:

- представляет собой среднее абсолютное процентное отклонение, что делает ее легко интерпретируемой;
- нормализует абсолютную ошибку относительно истинных значений, что позволяет сравнивать модели с разными диапазонами значений;
- может принимать значения от 0% до бесконечности. Это делает ее более чувствительной к отрицательным ошибкам (когда предсказание меньше истинного значения), чем к положительным ошибкам;
- становится неопределенной, если истинное значение равно нулю. Это ограничивает ее применение в задачах, где могут встречаться нулевые значения.

Метрика MSE (Mean Squared Error) широко используется для оценки качества прогнозных моделей. MSE вычисляет среднее квадратичное отклонение предсказанных значений от истинных. Ее особенности:

- является чувствительной к выбросам, поскольку квадратирует ошибки. Это означает, что большие ошибки оказывают значительно больше влияния на MSE, чем небольшие ошибки;

- не является масштабно-инвариантной, т.е. ее значение зависит от масштаба целевой переменной. Это затрудняет сравнение моделей с разными диапазонами значений;

- выражается в квадрате единиц измерения целевой переменной, что затрудняет ее интерпретацию. Однако, ее можно интерпретировать как средний квадрат ошибки;

- является несмещенной оценкой дисперсии ошибок прогноза.

После выбора метрик и применения моделей, необходимо оценить их эффективность. Для этого можно использовать кросс-валидацию, которая позволяет оценить качество модели на различных выборках данных.

2. Анализ данных

Анализ данных является критически важной составляющей в машинном обучении и имеет огромное значение [21]. Вот несколько причин, почему анализ данных важен:

Качество и точность моделей: качество моделей машинного обучения напрямую зависит от качества исходных данных. Анализ данных позволяет исследовать, очищать, преобразовывать и подготавливать данные перед обучением модели. Это включает обработку выбросов, пропущенных значений, масштабирование и кодирование признаков, а также выбор наиболее информативных признаков. Чистые, хорошо структурированные данные обеспечивают лучшую основу для обучения моделей и повышают их точность и надежность.

Понимание данных: анализ данных позволяет получить представление о структуре, свойствах и особенностях данных. Это включает исследование распределений, корреляций, взаимосвязей и других характеристик данных. Понимание данных помогает выбрать наиболее подходящие модели и методы обучения, а также принимать обоснованные решения в процессе обработки и предварительной обработки данных.

Выявление закономерностей: анализ данных помогает обнаруживать скрытые закономерности, паттерны и взаимосвязи в данных. Это может быть полезно для создания новых признаков, выделения важных атрибутов и понимания сути решаемой задачи. Выявление закономерностей и паттернов позволяет строить более эффективные модели и улучшать их производительность.

Принятие обоснованных решений: анализ данных предоставляет информацию, которая помогает принимать обоснованные решения во время разработки моделей машинного обучения. Это может включать выбор наиболее подходящих алгоритмов и моделей, определение оптимальных гиперпараметров, оценку производительности моделей и сравнение

различных подходов. Анализ данных помогает снизить риски и принимать информированные решения на каждом этапе процесса машинного обучения.

2.1 Описание датасета

За период практики были собранные данные о метеорологических характеристиках и погоде в г. Минске за период с 2011 по 2023 год (рисунок 1). Итоговый датасет составил 37983 записи, которые включают в себя ежедневные измерения климатических явлений с периодичностью восемь раз в сутки (каждые три часа).

Архив погоды в Минске (город Минск, Беларусь)																	
Время (UTC), дата	Ветер (напр., м/с)	Видим.	Явления	Облачность	T (C)	Td (C)	f (%)	Te (C)	Tes (C)	Комфортность	P (гПа)	Po (гПа)	Tmin (C)	Tmax (C)	R (мм)	R24 (мм)	S (см)
00 1.01	З 2	7000 м	дымка (ливн. осадки, снег)	10/8 1020 м [Sc Ac trans]	-4.8	-5.8	93	-9	-9	по сезону	1010.6	982.5					
03 1.01	Ю 2	6000 м	дымка	9/9 840 м [Sc]	-5.1	-6.1	93	-9	-9	по сезону	1008.6	980.5			0.0		
06 1.01	Ю 3	10 км		10/8 870 м [Sc Ac trans]	-4.8	-6.1	91	-10	-10	по сезону	1003.9	976.0	-6.6		0.0		33
09 1.01	Ю 4-13 (12)	400 м	сильн. снег	10/10 240 м [Fmb Ns]	-4.4	-5.0	96	-11	-11	по сезону	998.2	970.5					
12 1.01	ЮЗ 4 (13)	8000 м	слаб. снег (метель)	10/10 300 м [Fmb Ns]	-1.3	-2.1	94	-7	-7	по сезону	993.8	966.5					
15 1.01	ЮЗ 4-10 (12)	1200 м	снег	10/10 210 м [Fmb Ns]	-0.8	-1.2	97	-7	-7	по сезону	992.0	964.9			4		
18 1.01	ЮЗ 2 (10)	1900 м	снег	10/10 210 м [Fmb Ns]	-0.2	-0.8	96	-4	-4	по сезону	991.5	964.4	-0.1		6		
21 1.01	ЮЗ 2	10 км	слаб. снег	10/10 450 м [Sc]	-0.2	-1.1	94	-4	-4	по сезону	991.3	964.2					
00 2.01	З 3	8000 м	слаб. снег	10/8 400 м [Sc As op]	-0.8	-1.5	95	-6	-6	по сезону	991.4	964.3					
03 2.01	З 2 (11)	10 км	слаб. снег	10/8 420 м [Sc As op]	-2.6	-4.2	89	-7	-7	по сезону	994.1	966.7			3		
06 2.01	З 3	10 км	слаб. снег	10/10 480 м [Sc]	-3.7	-7.0	78	-9	-9	по сезону	996.3	968.7	-3.7		3		35
09 2.01	З 2	6000 м	слаб. снег	10/10 480 м [Sc]	-4.6	-7.3	81	-9	-9	по сезону	999.8	971.9					
12 2.01	СЗ 1	4200 м	слаб. ливневой снег	10/10 270 м [Cb calv]	-4.7	-7.1	83	-7	-7	по сезону	1002.0	974.2					
15 2.01	З 2	4600 м	слаб. ливневой снег	10/10 400 м [Sc]	-4.7	-7.5	81	-9	-9	по сезону	1004.5	976.6			0.5		
18 2.01	СЗ 2	8000 м	слаб. снег (ливн. осадки)	10/8 1080 м [Sc Ac trans]	-6.1	-9.2	79	-10	-10	по сезону	1007.3	979.1	-3.4		0.5		
21 2.01	штиль 0	10 км	слаб. снег	10/9 1050 м [Sc]	-6.3	-10.2	74	-7	-7	по сезону	1010.3	981.9					

Рисунок 1 – Датасет температуры воздуха

Датасет состоит из 20 столбцов:

- 1. Время измерения данных: все измерения проводятся с интервалом в три часа, поэтому каждому дню соответствует восемь строк с данными.
- 2. Дата измерения: данный столбец помогает определить соответствие климатического сезона измеренным данным.

3. Направление ветра: показывает направление, с которого дует ветер.
4. Скорость ветра: указывает среднюю скорость ветра, измеренную за 10 минут, а также указывает на порывы ветра.
5. Видимость: характеристика прозрачности атмосферы и возможности различать зрением удаленные объекты, отделенные слоем воздуха той или иной мутности [8].
6. Явления: указаны атмосферные явления, наблюдавшиеся в срок или в последний час перед сроком.
7. Облачность: указана общая и нижняя облачность в баллах, высота нижней границы облаков в метрах, а также тип облаков на латинском языке.
8. Температура воздуха: текущая температура воздуха, зафиксированная ртутным термометром, установленным в психрометрической палатке.
9. Температура точки росы: температура, при которой водяной пар в воздухе достигнет состояния насыщения при понижении температуры.
10. Относительная влажность воздуха: текущая влажность воздуха, измеренная на высоте двух метров.
11. Эффективная температура: температура, которую ощущает одетый по сезону человек в тени. Характеристика душности погоды. При расчете учитывается влияние влажности воздуха и скорости ветра на теплоощущения человека.
12. Эффективная температура на солнце: температура, которую ощущает человек, с поправкой на солнечный нагрев. Характеристика знойности погоды. Ночью, в пасмурную погоду, а также при ветре 12 м/с и более поправка равна нулю.
13. Комфортность.
14. Атмосферное давление: атмосферное давление, приведенное к уровню моря

15. Атмосферное давление: атмосферное давление, представленное на уровне метеостанции.
16. Минимальная температура: минимальная температура воздуха на уровне 2 м над землей.
17. Максимальная температура: максимальная температура воздуха на уровне 2 м над землей.
18. Количество осадков: количество выпавших осадков за период времени, мм.
19. Количество осадков: количество выпавших осадков за 24-часовой период, выраженное в миллиметрах.
20. Снежный покров: высота снежного покрова, см.

2.2 Разведывательный анализ данных

На начальном этапе из таблицы были удалены одиннадцать столбцов. Столбцы, содержащие данные о минимальной температуре, максимальной температуре, количестве осадков, суточном количестве осадков и высоте снежного покрова, были исключены из-за большого числа отсутствующих значений. Эффективная температура (температура, которую ощущает одетый по сезону человек в тени), эффективная температура на солнце (температура, которую ощущает человек, с поправкой на солнечный нагрев) и температура точки росы (температура, при понижении до которой содержащийся в воздухе водяной пар достигнет насыщения) частично или полностью заменяет столбец температура воздуха. Столбцы явления и облачность удалены из-за пропусков и большого количества разнообразных категориальных данных, которые сложно сгруппировать.

В итоговом датасете остались восемь столбцов:

1. Дата (“timestamp”).
2. Направление ветра (“wind”).
3. Скорость ветра (“speed”).

4. Видимость (“visible”).
5. Температура воздуха (“t”)
6. Относительная влажность воздуха (“hum”)
7. Комфортность (“comfort”)
8. Давление (“p”)

В исходной таблице отсутствовал год наблюдений, поэтому ко дню и месяцу измерений, для лучшего понимания был добавлен год измерений. Направление ветра преобладает западное (рисунок 2), что говорит о сильной зависимости от Балтийского моря (расстояние 501.6 км).

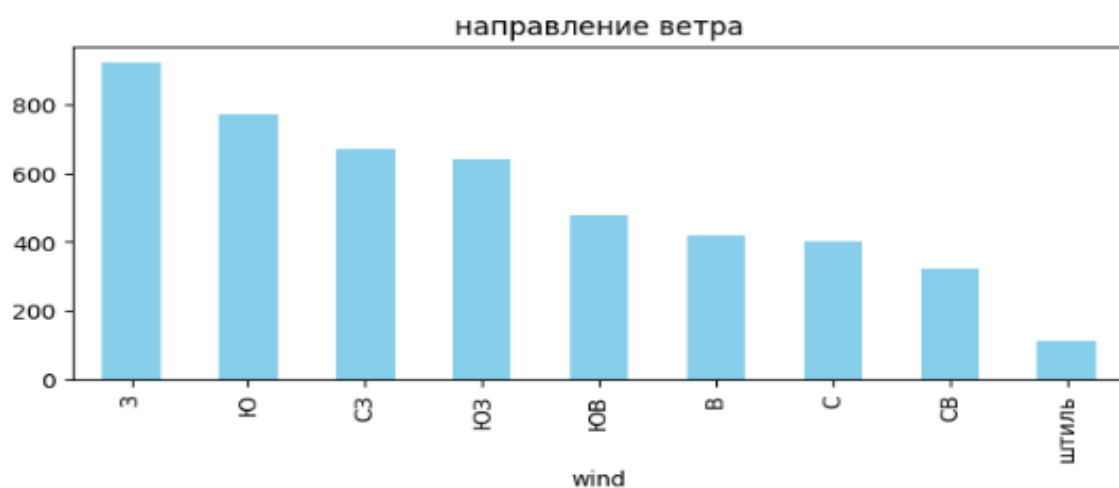


Рисунок 2 – График направления ветра

Преобладает небольшая, до 3 м/с скорость ветра, порывы случаются, но довольно редки (рисунок 3).

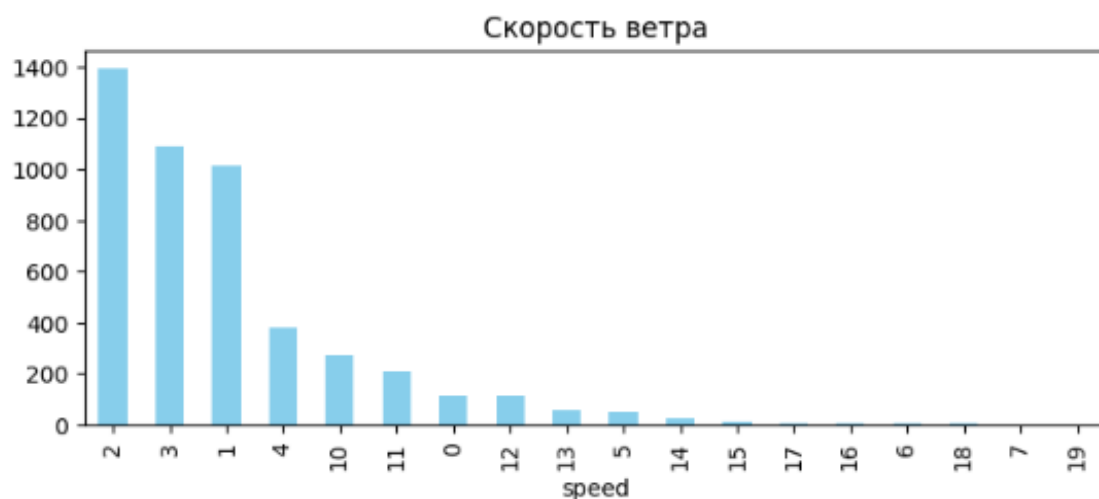


Рисунок 3 – График скорости ветра

Видимость характеризуется большими значениями, основная масса которых лежит в диапазоне от 10 до 20 км (рисунок 4).

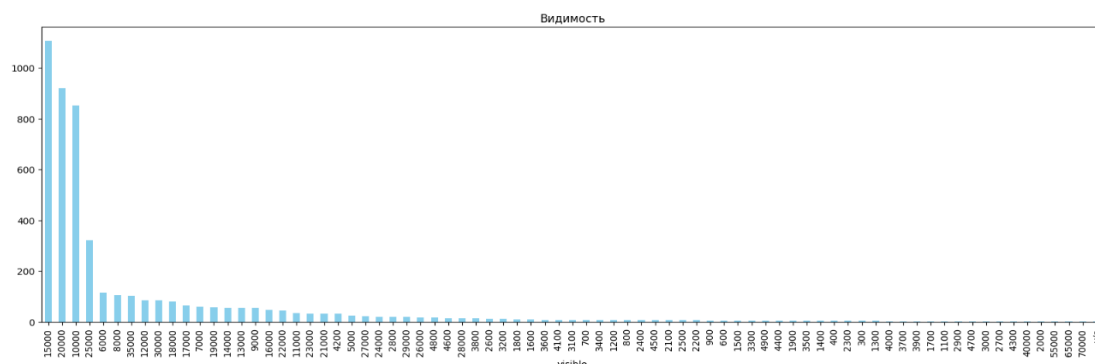


Рисунок 4 – График атмосферной видимости

Температура воздуха имеет ярко выраженную сезонность (рисунок 5), температура воздуха характерна для умеренно-континентального климата. Максимальные показатели температуры крайне редко превышают 32°C, а минимальные, особенно в последние годы, не опускаются ниже минус 20°C.

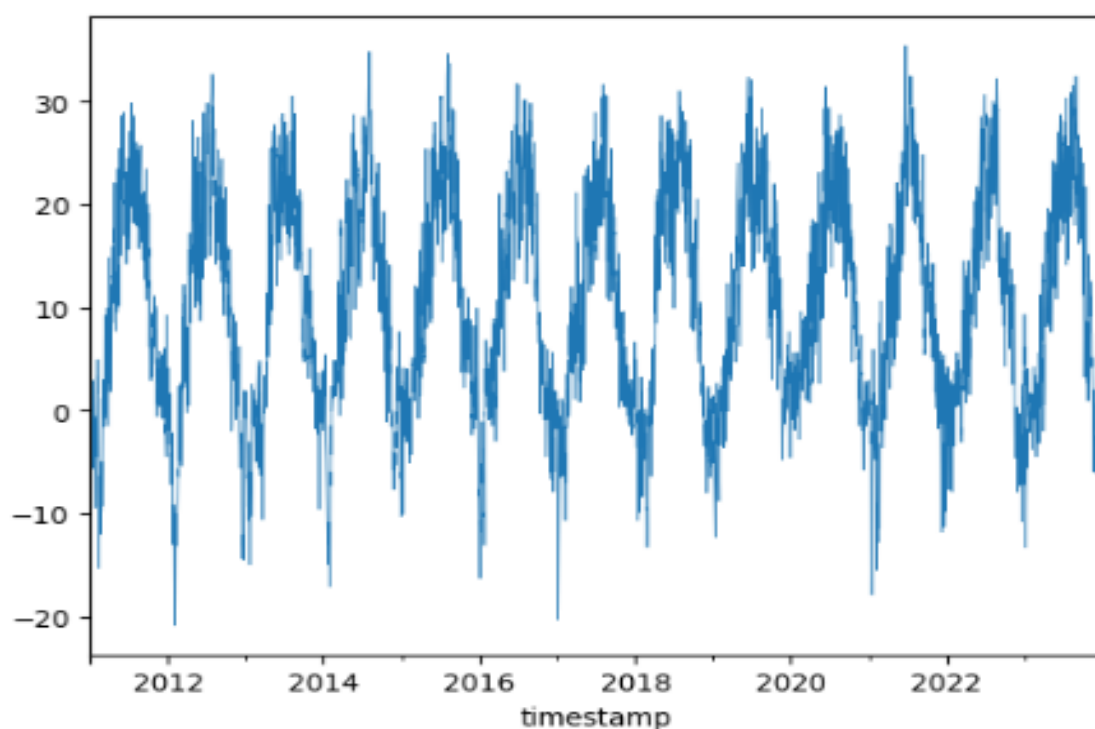


Рисунок 5 – График температуры воздуха

Относительная влажность воздуха является самым неоднозначным параметром. Первые две строчки занимают показатели влажности равные 48

и 92%, однако данные распределены достаточно равномерно. В распределении атмосферного давления нет каких-либо значений сильно выдающихся из средних (рисунок 6). Данный параметр можно сгруппировать в небольшое количество промежуточных значений.



Рисунок 6 – График относительной влажности воздуха

Под комфортностью понимается определение степени его благоприятности по отношению к организму человека [22]. Из рисунка 7 видно, что чаще всего температура соответствует сезону, какие-либо климатические аномалии (опасность, очень холодно, очень жарко), встречаются крайне редко.

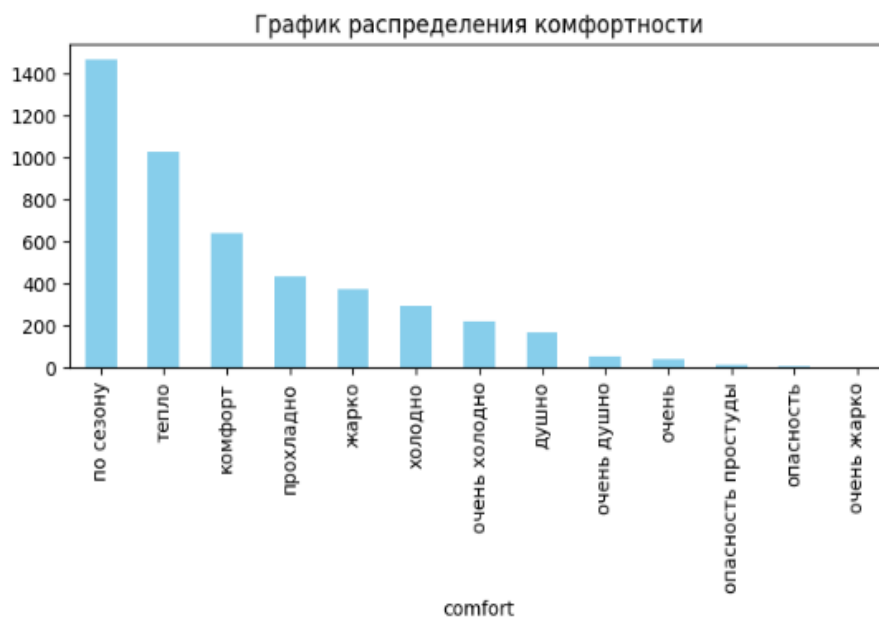


Рисунок 7 – График распределения комфортности

Обработка данных является неотъемлемой частью процесса создания модели машинного обучения. Она включает в себя предварительную обработку данных, удаление дублей, заполнение пропущенных значений, нормализацию, корреляцию и выбор значимых признаков [21].

На начальном этапе были проверены пропуски и дубликаты. Дубликатов в данных не было обнаружено, а пропусков было около 50. Поскольку пропущенных значений немного, а частота измерений довольно высока, было решено удалить эти значения из таблицы.

Далее все данные были разделены на категориальные и количественные. К числу категориальных были отнесены направление ветра и комфорт.

Затем категориальные колонки были закодированы в числовые значения, а столбец «видимость» был разбит на пять категорий. После этого был проведен корреляционный анализ (рисунок 8), которая позволяет определить степень линейной взаимосвязи между различными показателями и целевой переменной. Признаки с высокой корреляцией с целевой переменной могут быть выбраны для включения в модель.



Рисунок 8 – Корреляция параметров

Как видно из рисунка 9 корреляции относительно параметра температуры воздуха из первых пяти значений четыре занимают разные показатели комфортности климатических условий. Показатели «жарко», «душно», «очень жарко» свидетельствуют о том, что температура воздуха в момент метеорологических наблюдений находится выше среднестатистических значений.

```
t          1.000000
comfort    0.366112
visible     0.345921
year        0.034592
wind        0.005858
Pressure   -0.007402
speed       -0.011172
hum         -0.600428
Name: t, dtype: float64
```

Рисунок 9 – Корреляция параметров от параметра температуры воздуха

Стоит отметить, что самые часто встречаемые показатели комфортности «тепло» и «по сезону» имеют околонулевую и отрицательную корреляцию, это говорит о том, что данные параметры оказывают небольшое влияние на корреляцию. Большое влияние на корреляции имеет параметр атмосферная видимость (возможность визуально различать удаленные объекты, отделенные от наблюдателя слоем воздуха той или иной степени мутности. Характеризуется дальностью видимости, т. е. расстоянием, на котором наблюдаемый объект становится невозможно отличить от фона [23]), в значениях которой преобладают дни с большой видимостью. Основные же метеорологические параметры такие, как давление, относительная влажность, направление и скорость ветра, оказывают отрицательное влияние на модель.

2.3 Разложение временных рядов

Данные временных рядов могут отражать различные закономерности, и в таких случаях часто полезно разбить временной ряд на несколько компонент, каждая из которых будет представлять основные типы выявленных закономерностей. [24]. Одним из наиболее распространенных подходов к декомпозиции временного ряда является его разделение на компоненты, отражающие тренд, сезонность и остаточную составляющую (рисунок 10).

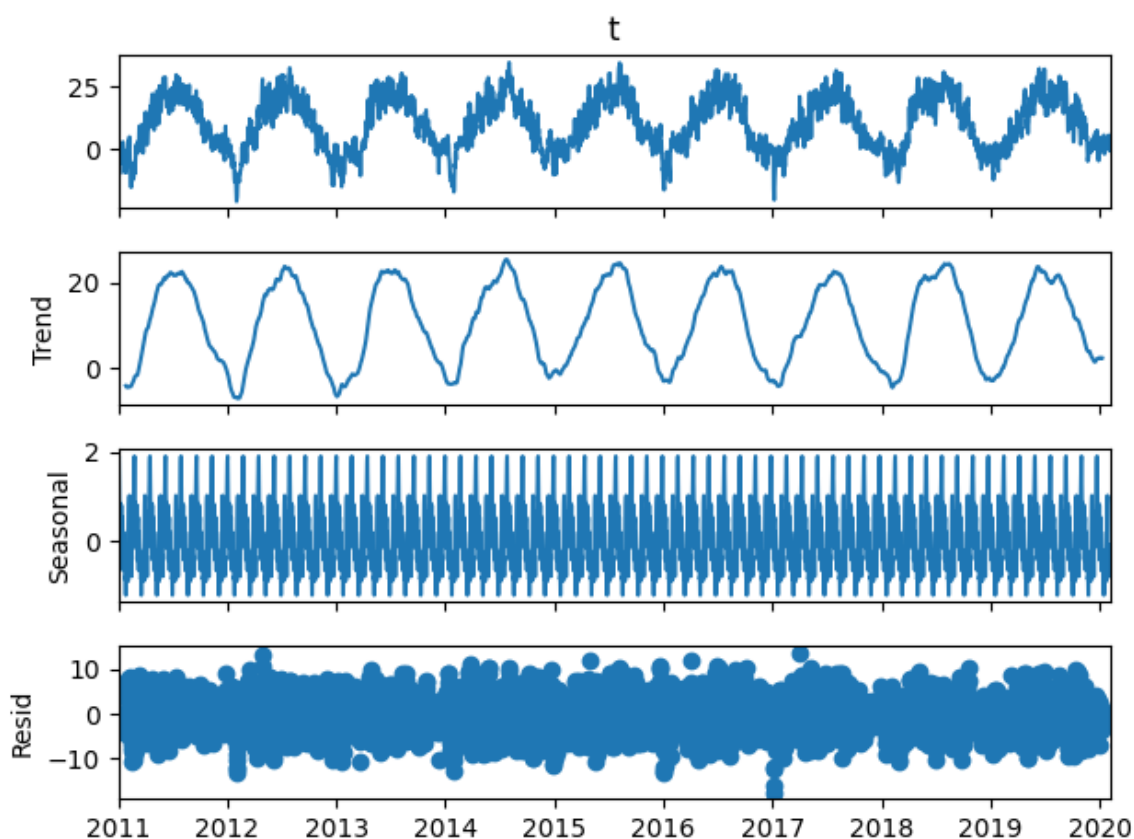


Рисунок 10 – Разложение временного ряда

В рамках пакета SKTime разложение можно выполнить при помощи объектов класса Deseasonalizer [25] и Detrender [26] (рисунок 11).

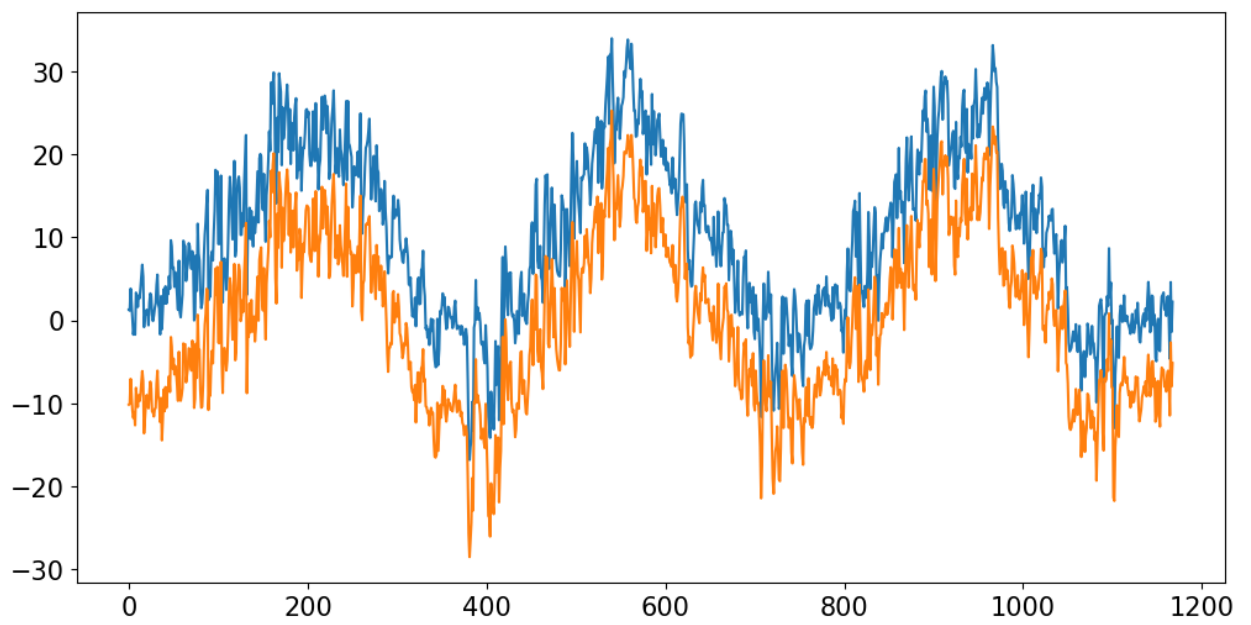


Рисунок 11 – Разложение временного ряда Deseasonalizer и Detrender

Глава 3 Построение моделей и интерпретация результатов

3.1 Построение моделей

В случае наивных прогнозов устанавливаются для всех прогнозов значение последнего наблюдения [24]. То есть:

$$\bar{Y}_{T+h|T} = Y_T$$

где:

$\bar{Y}_{T+h|T}$ - прогнозное значение в момент времени $T+h$, сделанное в момент времени T ;

Y_T - фактическое значение в момент времени T .

Наивный прогноз относится к базовым методам прогнозирования, которые не задействуют сложных моделей или статистических алгоритмов. Чтобы вычислить квантили прогнозирования сначала оценивается стандартная ошибка остатков прогнозирования в предположении о некоррелированных остатках. Затем дисперсия прогноза вычисляется путем умножения остаточной дисперсии на константу. Эта константа представляет собой корректировку смещения небольшой выборки и каждый метод (среднее, последнее, дрейф) имеет разные формулы для расчета константы [27].

```
SEASON = 365
fh = np.arange(y_test.size) + 1
#fh = ForecastingHorizon(y_test.index, is_relative=False)
# ПРЕДСКАЗАТЕЛЬ
forecaster = NaiveForecaster(strategy="mean", sp=SEASON)
forecaster.fit(y_train)
y_pred = forecaster.predict(fh)
plot_series(y_train, y_test, y_pred, labels=["y_train",
"y_test", "y_pred"])

print(f'SMAPE = {smape(y_pred.values, y_test.values):.3f}')
```

После проведения сравнительного анализа трех стратегий, наилучший результат показала стратегия «среднее» (рисунок 12). Однако, несмотря на лучшую метрику ($sMAPE = 0.606$) график проходит через средние значения параметров и учитывает максимальные и минимальные показатели температуры воздуха.

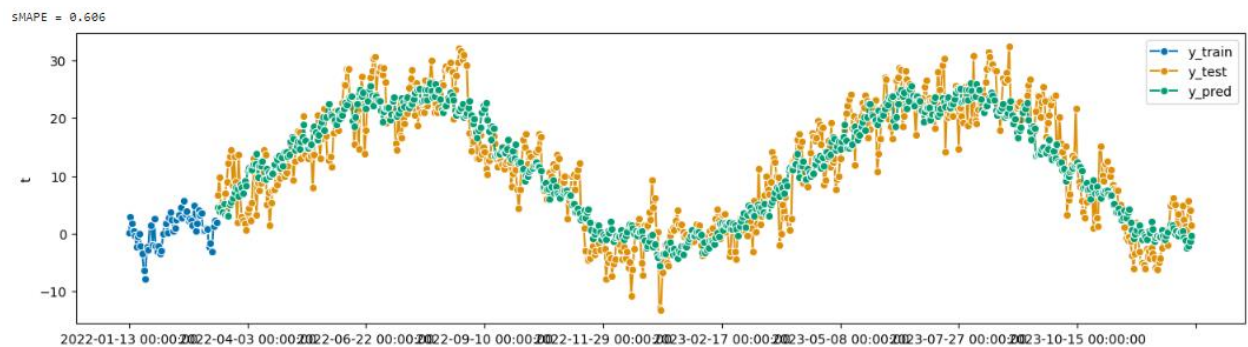


Рисунок 12 – Наивное предсказание

Экспоненциальное сглаживание является методом, при котором прогноз вычисляется как средневзвешенное прошлых наблюдений, при этом веса этих наблюдений экспоненциально уменьшаются по мере удаленности во времени. Т.е., чем моложе наблюдение, тем большим весом оно обладает [28].

```
SEASON = 365

fh = ForecastingHorizon(y_test.index, is_relative=False)
ses = ExponentialSmoothing()
forecaster = TransformedTargetForecaster(
    steps=[("deseasonalize", Deseasonalizer(model="additive",
sp=SEASON)),
    ("detrend",
Detrender(forecaster=PolynomialTrendForecaster(degree=1))),
    ("forecaster", ses)])
forecaster.fit(y_train)
y_pred = forecaster.predict(fh)
plot_series(y_train, y_test, y_pred, labels=["y_train",
"y_test", "y_pred"])
print(f'sMAPE = {smape(y_pred.values, y_test.values):.3f}')
```

Метод основан на допущении о том, что самое последнее наблюдение является единственным важным, а все предыдущие наблюдения не дают никакой информации на будущее. Экспоненциальное сглаживание можно интерпретировать как средневзвешенное значение, в котором весь вес приходится на последнее наблюдение (рисунок 13).

При использовании метода усреднения все будущие прогнозы равны простому среднему арифметическому наблюдаемых данных:

$$\bar{Y}_{T+h|T} = \frac{1}{T} \sum_{t=1}^T y_t$$

для $h = 1, 2, \dots$. Таким образом, метод усреднения исходит из предположения, что все наблюдения имеют одинаковую важность, и поэтому им присваиваются равные веса при формировании прогнозов.

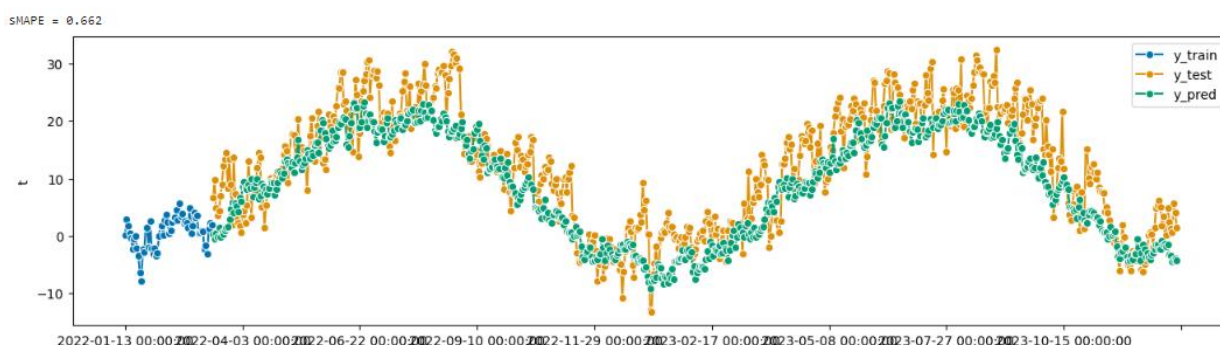


Рисунок 13 – Экспоненциальное сглаживание

Регрессия на основе k-ближайших соседей. Цель прогнозируется путем локальной интерполяции целей, связанных с ближайшими соседями в обучающей выборке.

Регрессия на основе ближайших соседей (рисунок 14) применяется в ситуациях, когда целевые переменные являются непрерывными, а не категориальными. Значение, назначаемое точке запроса, рассчитывается как среднее значение меток ее ближайших соседей.

Базовая регрессия ближайших соседей использует равномерные веса: то есть каждая точка в локальной окрестности вносит одинаковый вклад в

классификацию точки запроса. В некоторых случаях может быть полезно назначить веса точкам так, чтобы близлежащие точки вносили больший вклад в регрессию, чем удаленные. Это можно сделать с помощью параметра «weights» в модели. Значение по умолчанию назначает всем точкам одинаковые веса, но можно установить веса, обратно пропорциональные расстоянию от точки запроса. [29].

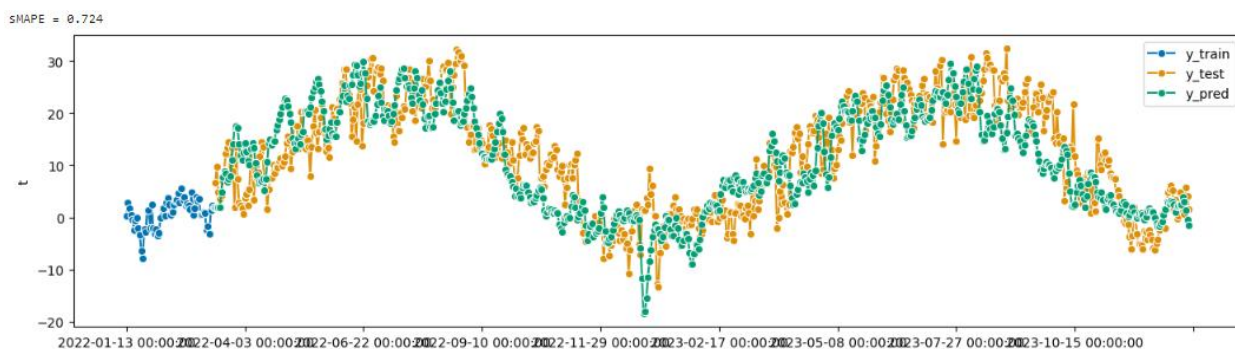


Рисунок 14 – Регрессия ближайших соседей

Метод экспоненциального сглаживания Холта-Винтерса применяется для прогнозирования временных рядов, которые проявляют как тренд, так и сезонные изменения. Методика Холта-Уинтерса состоит из следующих четырех методов прогнозирования, наложенных друг на друга (рисунок 15):

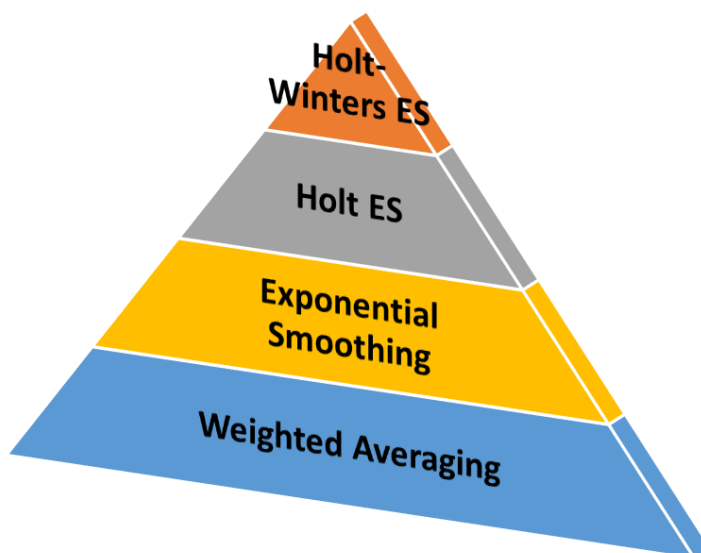


Рисунок 15 – Ключевые концепции, на которых основано экспоненциальное сглаживание Холта-Уинтерса

Средневзвешенные значения. Средневзвешенное значение — это просто среднее значение n чисел, где каждому числу присвоен определенный вес, а знаменатель — это сумма этих n весов. Веса часто назначаются в соответствии с некоторой функцией взвешивания. Широко используемыми функциями весов являются логарифмическая, линейная, квадратичная, кубическая и экспоненциальная. Усреднение как метод прогнозирования временных рядов обладает свойством сглаживать изменения исторических значений при расчете прогноза. Выбор подходящей функции весов позволяет прогнозисту определить, какие исторические значения будут иметь наибольшее влияние при расчете будущих значений временного ряда.

Экспоненциальное сглаживание. Метод экспоненциального сглаживания (ES) прогнозирует следующее значение, используя средневзвешенное всех предыдущих наблюдений, при этом веса экспоненциально убывают от последнего значения к более ранним историческим данным. Когда вы используете ES, вы делаете решающее предположение, что недавние значения временного ряда гораздо важнее для вас, чем более старые значения. У метода ES есть два больших недостатка: его нельзя использовать, когда ваши данные демонстрируют тенденцию и/или сезонные колебания.

Экспоненциальное сглаживание Холта. Метод Holt ES можно использовать для прогнозирования данных временных рядов, имеющих тенденцию. Но Holt ES не работает при наличии сезонных изменений во временном ряду.

Экспоненциальное сглаживание Холта-Уинтерса: Holt-Winters ES модифицирует метод Holt ES так, что его можно использовать при наличии как тренда, так и сезонности [30].

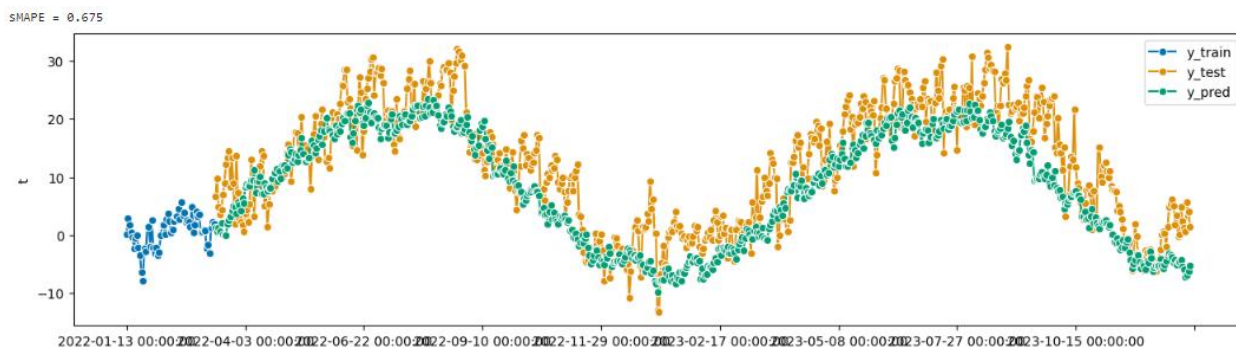


Рисунок 16 – Holt winters

Метод Prophet (пророк) основан на аддитивной модели (тренд + сезонность + праздники):

$$y(t) = g(t) + s(t) + h(t) + e(t)$$

$g(t)$ — тренд. Это логистическое уравнение для обработки нелинейного роста с насыщением, темпы роста которого зависят от пропускной способности [31].

Параметры, которые необходимо установить:

- рост, линейный или логистический тренд;
- точки изменения, список дат, определяющих изменение скорости роста (определяется автоматически моделью);
- `n_changepoint`, список точек изменений, определенных аналитиком;
- `Changepoint_prior_scale` меняет гибкость тренда: его увеличение сделает тренд более гибким.
- $s(t)$ означает сезонность. Временной ряд может иметь многопериодную сезонность, моделируемую рядом Фурье.

Параметром для определения сезонности являются:

- Ежегодная сезонность
- Еженедельная сезонность
- Ежедневная сезонность
- `Seasonity prior scale` меняет силу модели сезонности.

– $h(t)$ обозначает праздники и события. Пророки предоставляют список прошлых и будущих праздников и событий. Модель учитывает окна в эти дни, чтобы уловить последствия [32].

Параметры, которые необходимо установить:

- праздники, фрейм данных, содержащий дату праздников и событий;
- Holiday prior scale, меняет силу модели праздника.

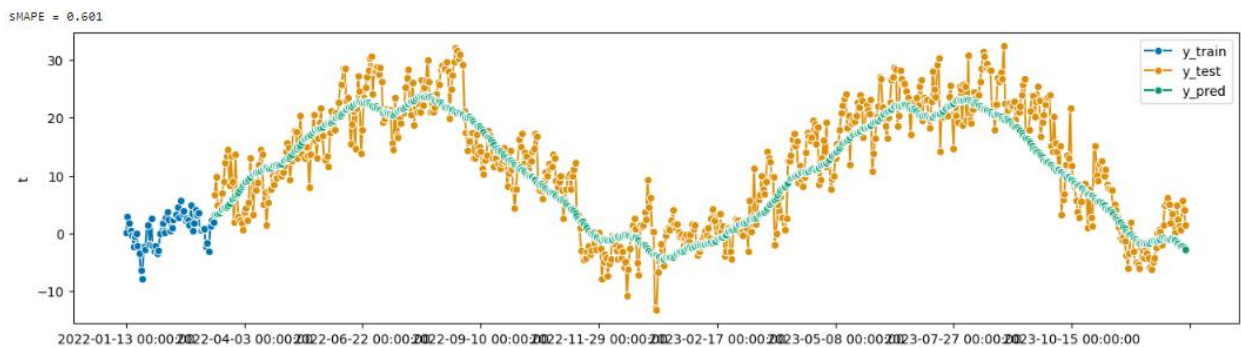
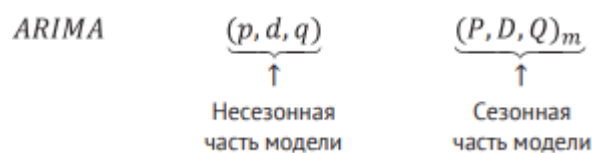


Рисунок 17 – Метод Prophet

Сезонная авторегрессия интегрированного скользящего среднего (SARIMA) является расширением ARIMA, которое поддерживает сезонные изменения [33]. Сезонная модель ARIMA строится путем добавления дополнительных сезонных компонентов в базовую модель ARIMA, которую мы рассматривали ранее. [24]. Это записывается следующим образом:



p : Порядок авторегрессии тренда.

d : Порядок разницы трендов.

q : Порядок трендовой скользящей средней.

P : Сезонный авторегрессионный порядок.

D : Порядок сезонных различий.

Q : Порядок сезонных скользящих средних.

m : количество временных шагов для одного сезонного периода.

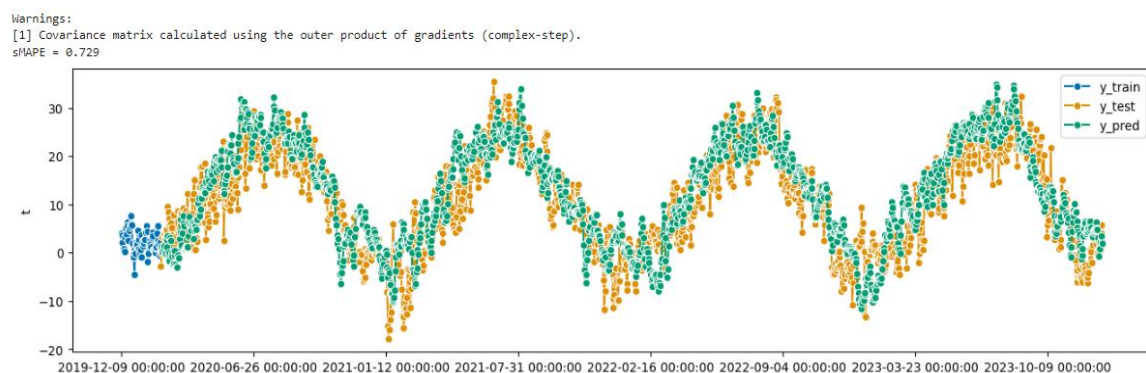


Рисунок 18 – Сезонная авторегрессия интегрированного скользящего среднего

При помощи автокорреляционной и частичной автокорреляционной функций (ACF и PACF соответственно) [34] подтвердим наличие сезонности в данных (рисунки 19, 20). Судя по ACF и PACF, мы видим сильную сезонность в наших данных.

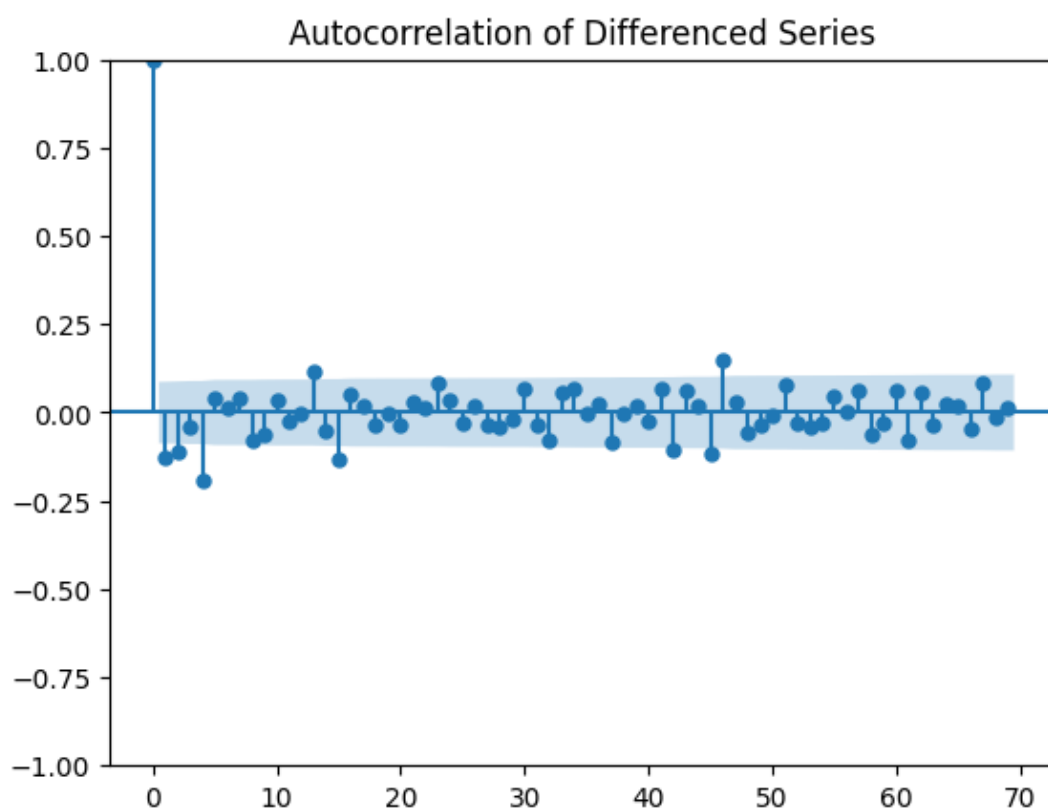


Рисунок 19 – Автокорреляционная функция

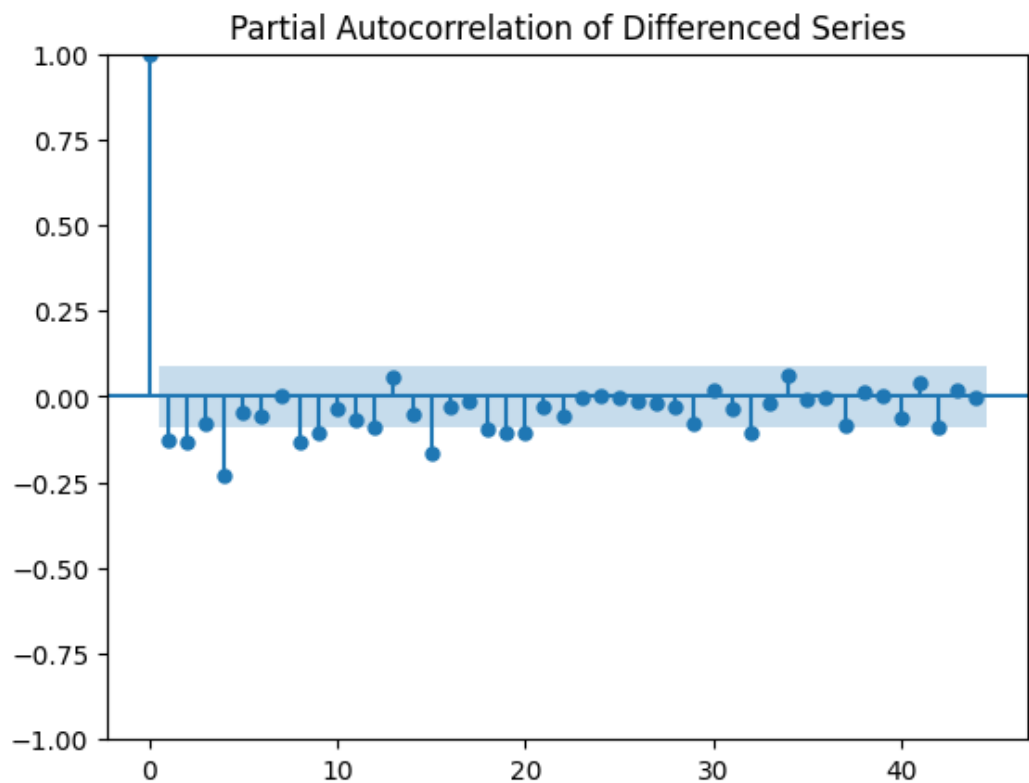


Рисунок 20 – Частная корреляционная функция

Одним из эффективных способов обнаружения выбросов в многомерных наборах данных является использование изолирующих лесов [35].

В этом подходе используются двоичные деревья для обнаружения аномалий, что приводит к линейной временной сложности и низкому использованию памяти, что хорошо подходит для обработки больших наборов данных. На рисунке 21 представлены аномальные показатели для временного ряда температуры воздуха.

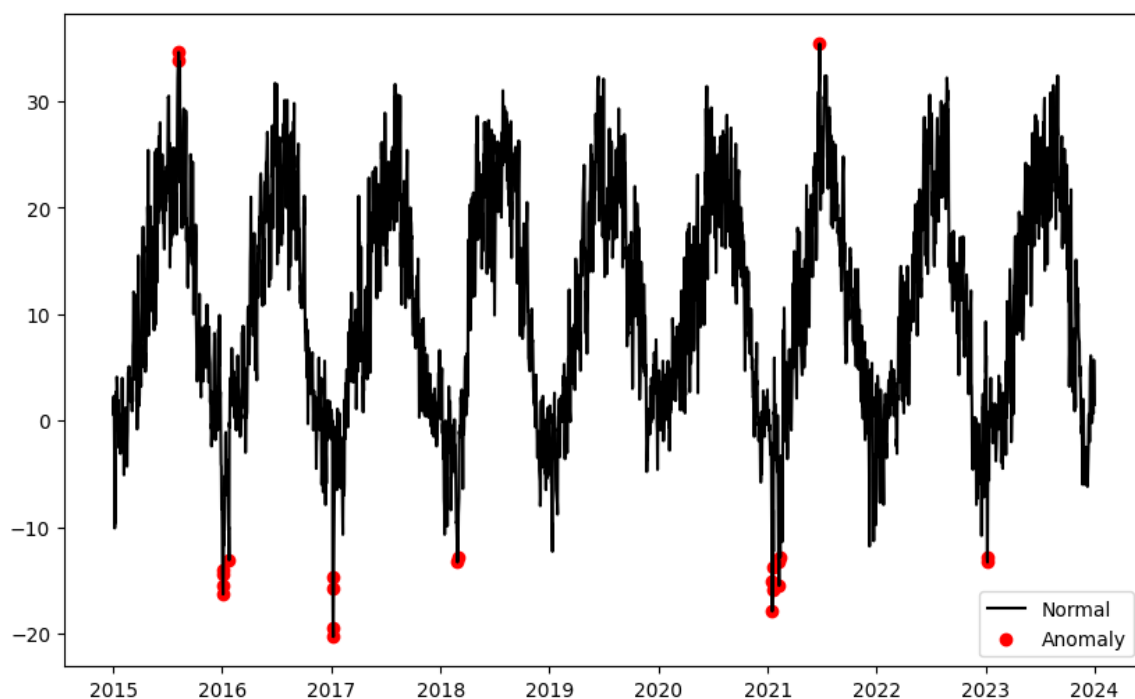


Рисунок 21 – Поиск аномалий при помощи IsolationForest

В изолирующем лесу случайно отобранные данные обрабатываются в древовидной структуре на основе случайно выбранных объектов [36]. Образцы, которые проникают глубже в дерево, с меньшей вероятностью будут аномалиями, поскольку для их изоляции требовалось больше разрезов. Точно так же образцы, которые попадают в более короткие ветви, указывают на аномалии, поскольку дереву было легче отделить их от других наблюдений [37].

3.2 Интерпретация результатов

По мнению [38] директора ФГБУ «Гидрометцентр России» Романа Вильфанда: «Фантасты были не правы в своих мечтах – максимальный период надежного прогноза погоды не может превысить 14 дней. Даже через 2-3 тысячи лет, при самом фантастическом развитии вычислительной техники этот срок увеличить невозможно, - заявил он. - Это связано с самой природой земной атмосферы».

Точность прогноза температуры воздуха на 5 дней варьируется в зависимости от нескольких факторов, включая используемую модель

прогнозирования, доступные данные и метеорологические условия. Прогноз погоды на длительный срок всегда связан с определенной степенью неопределенности, и точность его предсказания снижается с увеличением временного горизонта. В таблице 2 представлены данные о точности прогноза температуры воздуха ведущих мировых метеорологических служб за промежутки времени, равный пяти суткам.

Таблица 2 – Рейтинг глобальных прогностических моделей в 2022 году.

Точность прогноза на 120 часов [39]

Место	Модель	Страна	Точность, %	Разница с 2021 г.	Отставание от ECMWF, лет
1	ECMWF	Европа	90.53	0.60%	0
2	UKMO	UK	89.02	0.01%	6
3	BOM	Австралия	88.07	минус 0.08%	7
4	CMC	Канада	87.45	0.13%	9
5	ICON	Германия	87.30	0.40%	9
6	GFS	США	87.19	0.49%	10
7	JMA	Япония	87.05	0.38%	10
8	KMA	Корея	86.49	0.10%	10
9	CMA	Китай	81.43	нет данных	17
10	ПАВ	Россия	79.17	минус 0.16%	19

Из данной таблицы можно сделать вывод о том, что точность прогноза температуры воздуха падает с 98%, при прогнозе на до 3 суток, до 90% при прогнозе до 5 суток.

Когда речь идет о долгосрочном прогнозе температуры воздуха, то точность прогноза снижается по сравнению с более короткими прогнозами. Долгосрочный прогноз погоды не должен рассматриваться как точное предсказание будущей температуры. Он может быть полезным для общего представления о климатических тенденциях и сезонных изменениях, но конкретные значения температуры на долгосрочном горизонте могут быть менее точными.

Долгосрочные прогнозы температуры воздуха могут быть полезными при планировании потребления и производства электроэнергии, управлении запасами топлива и оптимизации поставок, подготовке к периодам высокого

спроса на энергию, планировании сельскохозяйственных работ, управление производством и хранением сельскохозяйственной продукции, планировании строительных работ с учетом погодных условий.

Как видно из таблицы 3, все использованные методы показали примерно одинаковый результат. Самая лучшая метрика получилась у модели Prophet и составила 0.601, что очень далеко от приемлемых результатов.

Таблица 3 – Сравнение методов предсказания

Метод	sMape	Mape	MSE	R ²
Наивное предсказание	0.652	1.426	27.723	0.732
Экспоненциальное сглаживание	0.626	1.409	35.86	0.647
Holt winters	0.692	1.561	39.089	0.615
Регрессия на основе k-ближайших соседей	0.892	1.087	33.572	0.67
Регрессия с использованием пайплайна	0.649	1.624	33.77	0.668
Prophet	0.601	1.089	19.605	0.807
Sarima	0.729	1.470	32.632	0.701

На рисунке 22 представлены доверительные интервалы.

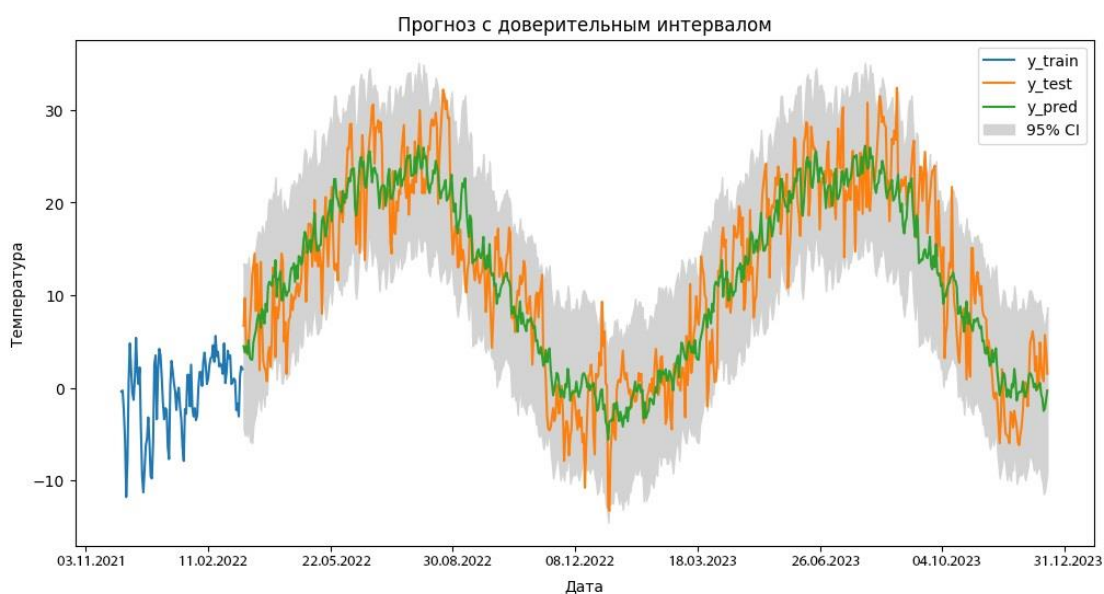


Рисунок 22 – График температуры с доверительным интервалом

Однако, из рисунка 23 видно, что модель можно настроить для предсказания перехода значений через отрицательные показатели и обратно. Если в модели наивного предсказания график смещался к средним значениям, то в модели с использованием экспоненциального сглаживания все предсказанные значения смещаются к минимальным значениям.

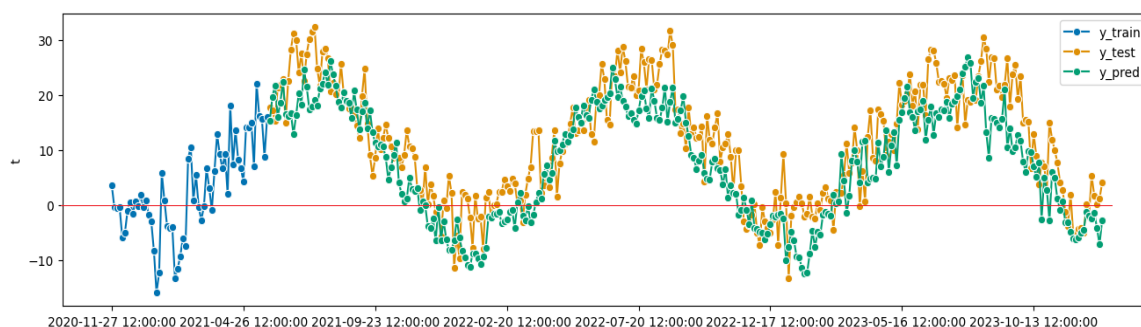


Рисунок 23 – График перехода модели через отрицательные значения

Учитывая, что многие ученые говорят о глобальном потеплении [6], связанных, в том числе, и с деятельностью человека, имеет смысл внесение новых показателей в исходные данные. Одним из таких показателей можно добавить коэффициент на деятельность человека или выбросы углекислого газа.

За период с 1990 по 2022 годы количество выбросов углекислого газа в атмосферу увеличилось в полтора раза [40]. Это говорит о том, что деятельность человека серьезно влияет на изменение климата и эти данные надо также учитывать при построении модели.

ЗАКЛЮЧЕНИЕ

В работе было проведено исследование эффективности долговременного предсказания температуры воздуха, выполненного при помощи методов анализа временных рядов, таких как: наивный, экспоненциальное сглаживание, Prophet, SARIMA.

Основная идея заключалась в том, что климатические процессы обладают ярко выраженной сезонностью и повторяемостью природных явлений. Это имеет большое значение при анализе и прогнозировании долгосрочных тенденций в изменении температуры воздуха:

- Сезонность: температурные циклы в течение года, более высокие температуры летом и более низкие зимой. Смена времен года с характерными для них погодными условиями. Повторяемость сезонных экстремумов.

- Многолетняя повторяемость: наличие устойчивых климатических периодов и циклов (суточные, годовые, многолетние). Возможность прогнозирования на основе исторических данных. Выявление трендов и аномалий в рамках повторяющихся циклов.

- Привязка к географическим факторам: влияние широты, высоты над уровнем моря. Региональная специфика температурных режимов, учет микроклиматических особенностей.

Для решения задачи долгосрочного прогноза температуры воздуха были использованы модели анализа временных рядов, оценка эффективности которых показала результаты по sMAPE от 0.601 до 0.809.

Работа была проведена в два этапа. На первом этапе модель был использован полный датасет, содержащий 37983 записей. Этот датасет содержал записи за 13 лет наблюдений с периодичностью в три часа. Однако, после всех преобразований и анализа, оценка эффективности показала неудовлетворительные результаты. Ни одна модель не смогла определить цикличность и сезонность. По этому, на втором этапе, из исходного датасета

были использованы не трехчасовые показатели температуры, а ежесуточные. Далее на полученных данных были обучены модели.

Сравнение результатов работы моделей на двух наборах данных (полный и ограниченный) подтвердили эффективность использования ограниченного датасета. Точность работы на полном датасете оказалась в три раза хуже, чем на ограниченном.

Были изучены и построены модели предсказания температуры воздуха, произведена тонкая настройка параметров моделей, произведено сравнение моделей и интерпретированы результаты.

В результате исследования удалось достичь поставленной цели — построение модели для долгосрочного предсказания температуры воздуха при помощи машинного обучения, лучшей моделью для этого стала Prophet (таблица 3).

Таким образом, машинное обучение помогло в обработке больших объемов данных, их анализе и выполнении быстрых вычислений, что особенно важно для оперативного и точного прогноза погоды.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Alley, R. B. Advances in weather prediction / R.B. Alley, K.A. Emanuel, F. Zhang – Текст : непосредственный. // Science (New York, N.Y.). 2019. Т. 363. № 6425. – С. 342-344.
2. Increase of extreme events in a warming world | PNAS. – URL: <https://www.pnas.org/doi/10.1073/pnas.1101766108> (дата обращения: 20.03.2024) – Текст: электронный.
3. История прогноза погоды | Новости и события | Государственное учреждение «Белгидромет» — Республиканский центр по гидрометеорологии, контролю радиоактивного загрязнения и мониторингу окружающей среды. – URL: <https://www.belgidromet.by/ru/news-ru/view/istorija-prognoza-pogody-3558/> (дата обращения: 06.05.2024) – Текст: электронный.
4. Who we are – The Met Office is the national meteorological service for the UK. We provide critical weather services and world-leading climate science, helping you make better decisions to stay safe and thrive. – URL: <https://www.metoffice.gov.uk/about-us/who> (дата обращения: 22.03.2024) – Текст: электронный.
5. (PDF) Chaos and weather prediction January 2000. – URL: https://www.researchgate.net/publication/228552816_Chaos_and_weather_prediction_January_2000 (дата обращения: 17.05.2024) – Текст: электронный.
6. Причины и последствия изменения климата | Организация Объединенных Наций U. Nations. – URL: <https://www.un.org/ru/climatechange/science/causes-effects-climate-change> (дата обращения: 27.04.2024) – Текст: электронный.
7. Краткосрочное Прогнозирование Температуры На Основе Нейронных Сетей Кузнецов Б.Ф. – URL: <https://www.elibrary.ru/item.asp?id=37203146> (дата обращения: 20.03.2024) – Текст: электронный.

8. Хромов С.П., Петросянц М.А. Метеорология и климатология / Хромов С.П., Петросянц М.А. Московский государственный университет имени М.В. Ломоносова, 2012. – 584 с.– ISBN 978-5-211-06334-1. – Текст : непосредственный.

9. Вакуленко Н.В., Серых И.В., Сонечкин Д.М. Хаос и порядок в атмосферной динамике. Часть 3. Предсказуемость Эль-Ниньо / Вакуленко Н.В., Серых И.В., Сонечкин Д.М. – Текст : непосредственный. 2018. – С. 75-94.

10. АВАКЯН С.В. Роль Активности Солнца В Глобальном Потеплении / АВАКЯН С.В. – Текст : непосредственный. 2013. № 3. – С. 275-285.

11. Section 2.1.2.3 ENS Long-Range (Seasonal) Forecast - Forecast User Guide - ECMWF Confluence Wiki. – URL: <https://confluence.ecmwf.int/display/FUG/Section+2.1.2.3+ENS+Long-Range+%28Seasonal%29+Forecast> (дата обращения: 07.05.2024) – Текст: электронный.

12. Section 8.3 Long Range Output - Seasonal - Forecast User Guide - ECMWF Confluence Wiki. – URL: <https://confluence.ecmwf.int/display/FUG/Section+8.3+Long+Range+Output+-+Seasonal> (дата обращения: 07.05.2024) – Текст: электронный.

13. Learning skillful medium-range global weather forecasting | Science. – URL: <https://www.science.org/stoken/author-tokens/ST-1550/full> (дата обращения: 20.03.2024) – Текст: электронный.

14. Bouallègue, Z. B. The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context // Bulletin of the American Meteorological Society. 2024. Т. 1. The rise of data-driven weather forecasting. № аоп.

15. Roy, D. Forecasting The Air Temperature at a Weather Station Using Deep Neural Networks / D. Roy – Текст : непосредственный. // Procedia Computer Science. 2020. Т. 178. – С. 38-46.

16. Air temperature prediction using random forest and particle swarm optimization algorithm Yalla S J V Durga Bhavani, Venkata Raju Athili, Sirisha Balla, K. Ajita. – URL: <https://ijfans.org/uploads/paper/edf550957a0b2d2d58911de28230a40d.pdf> (дата обращения: 09.05.2024) – Текст: электронный.

17. Как составляется прогноз погоды в Белгидромете – погода by, погода в минске. – URL: <https://www.belgidromet.by/ru/news-ru/view/kak-sostavljaetsja-prognoz-pogody-v-belgidromete-2598> (дата обращения: 27.04.2024) – Текст: электронный.

18. Белгидромет: точность краткосрочных прогнозов погоды в Беларуси составляет 98 % - Минск-новости – 20 марта, 2024 - Прогнозы погоды Белгидромета на трое суток точны на 96–98 %. Об этом рассказала журналистам 20 марта начальник службы метеорологических. – URL: <https://minsknews.by/belgidromet-opravdyvaemost-kratkosrochnyh-prognozov-pogody-v-belarusi-sostavlyayet-98/> (дата обращения: 07.05.2024) – Текст: электронный.

19. Rubin, D. B. Inference and Missing Data / D.B. Rubin – Текст : непосредственный. // *Biometrika*. 1976. Т. 63. № 3. – С. 581-592.

20. Солодушкин С.И., Юманова И.Ф. Прикладная статистика: учеб. пособие / Солодушкин С.И., Юманова И.Ф. ; 2-е изд. Уральский федеральный университет, – 100 с.– ISBN 978-5-7996-3829-0. – Текст : непосредственный.

21. Akimov, A. A. Data preprocessing for machine learning / A.A. Akimov, D.R. Valitov, A.I. Kubryak – Текст : непосредственный. // Научное обозрение. Технические науки (Scientific Review. Technical Sciences). 2022. № №2 2022. – С. 26-31.

22. Витченко, А. Н. Геоэкологическая оценка комфортности климата крупных городов Беларуси / А.Н. Витченко, И.А. Телеш – Текст : непосредственный. 2011.

23. Атмосферная видимость • Большая российская энциклопедия - электронная версия. – URL: <https://old.bigenc.ru/physics/text/1838873> (дата обращения: 27.04.2024) – Текст: электронный.

24. Р. Хайндман, Д. Атанасопулос. Прогнозирование: принципы и практика / Р. Хайндман, Д. Атанасопулос ДМК Пресс, 2023. – 458 с.– ISBN 978-5-93700-151-1. – Текст : непосредственный.

25. Deseasonalizer — sktime documentation. – URL: https://www.sktime.net/en/stable/api_reference/auto_generated/sktime.transformations.series.detrend.Deseasonalizer.html (дата обращения: 17.05.2024) – Текст: электронный.

26. Detrender — sktime documentation. – URL: https://www.sktime.net/en/latest/api_reference/auto_generated/sktime.transformations.series.detrend.Detrender.html (дата обращения: 17.05.2024) – Текст: электронный.

27. NaiveForecaster — документация sktime. – URL: https://www.sktime.net/en/stable/api_reference/auto_generated/sktime.forecasting.naive.NaiveForecaster.html (дата обращения: 29.04.2024) – Текст: электронный.

28. Holt, C. C. Forecasting seasonals and trends by exponentially weighted moving averages / C.C. Holt – Текст : непосредственный. // International Journal of Forecasting. 2004. Т. 20. № 1. – С. 5-10.

29. 1.6. Nearest Neighbors – sklearn.neighbors provides functionality for unsupervised and supervised neighbors-based learning methods. Unsupervised nearest neighbors is the foundation of many other learning methods, notably m... – URL: <https://scikit-learn/stable/modules/neighbors.html> (дата обращения: 29.04.2024) – Текст: электронный.

30. Holt-Winters Exponential Smoothing – Holt-Winters Exponential Smoothing is used for forecasting time series data that exhibits both a trend and a seasonal variationRead more Holt-Winters Exponential Smoothing. – URL:

<https://timeseriesreasoning.com/contents/holt-winters-exponential-smoothing/>
(дата обращения: 16.05.2024) – Текст: электронный.

31. Tutorial: Time Series Forecasting with Prophet – Explore and run machine learning code with Kaggle Notebooks | Using data from Air Passengers. – URL: <https://kaggle.com/code/prashant111/tutorial-time-series-forecasting-with-prophet> (дата обращения: 09.05.2024) – Текст: электронный.

32. A gentle introduction to Time Series R. Pollastro – ARIMA and Prophet starter kit. – URL: <https://medium.com/quantyca/a-gentle-introduction-to-time-series-7e9f5418dad8> (дата обращения: 02.05.2024) – Текст: электронный.

33. Josef Perktold. statsmodels/statsmodels: Release 0.14.2.statsmodels/statsmodels / Josef Perktold, Skipper Seabold, Kevin Sheppard [и др.]. – [object Object], 2024.

34. Time Series: Interpreting ACF and PACF. – URL: <https://www.kaggle.com/code/iamleonie/time-series-interpreting-acf-and-pacf/notebook?scriptVersionId=80091228> (дата обращения: 09.05.2024) – Текст: электронный.

35. 2.7. Novelty and Outlier Detection – Many applications require being able to decide whether a new observation belongs to the same distribution as existing observations (it is an inlier), or should be considered as different (it is an ... – URL: https://scikit-learn/stable/modules/outlier_detection.html (дата обращения: 09.05.2024) – Текст: электронный.

36. IsolationForest example – An example using IsolationForest for anomaly detection. The Isolation Forest is an ensemble of “Isolation Trees” that “isolate” observations by recursive random partitioning, which can be represent... – URL: https://scikit-learn/stable/auto_examples/ensemble/plot_isolation_forest.html (дата обращения: 17.05.2024) – Текст: электронный.

37. sklearn.ensemble.IsolationForest – Examples using sklearn.ensemble.IsolationForest: IsolationForest example Comparing anomaly detection algorithms for outlier detection on toy datasets Evaluation of outlier detection estimators. – URL: [https://scikit-](https://scikit-learn.org/stable/auto_examples/ensemble/plot_isolation_forest.html)

learn/stable/modules/generated/sklearn.ensemble.IsolationForest.html (дата обращения: 09.05.2024) – Текст: электронный.

38. Почему долгосрочные прогнозы никогда не достигнут 100% оправдываемости | Официальный портал Северо-Западного УГМС, прогноз погоды по Санкт-Петербургу на трое суток. – URL: <http://www.meteo.nw.ru/articles/index.php?id=840> (дата обращения: 13.05.2024) – Текст: электронный.

39. ECMWF |. – URL: <https://wmo.cdnv.ecmwf.int/> (дата обращения: 09.05.2024) – Текст: электронный.

40. Выбросы CO₂ от сжигания топлива | Мировая статистика по данным CO₂ | Enerdata – Мировые данные по выбросам CO₂ от сжигания топлива. Карта и статистика по мировым выбросам CO₂. – URL: <https://energystats.enerdata.net/co2/emissions-co2-data-from-fuel-combustion.html> (дата обращения: 17.05.2024) – Текст: электронный.