МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «МИСиС»**

*ИНСТИТУТ* ИТАСУ

*КАФЕДРА АВТОМАТИЗИРОВАННЫХ СИСТЕМ УПРАВЛЕНИЯ*

*НАПРАВЛЕНИЕ* 09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

**Лабораторная работа №6**

**По дисциплине**: Технологии интеллектуального анализа данных

**На тему:** Прогнозирование временных рядов

**Студент: Костромин Д.А.**

**Группа: МИВТ 18-5-7**

**Преподаватель: Фомичева О.Е.**

**Оценка: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Дата защиты: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Москва, 2019

**Оглавление**

[**Введение** 3](#_Toc26225709)

[**1. Метрики оценки качества регрессии** 4](#_Toc26225710)

[**2. Анализ временных рядов** 6](#_Toc26225711)

[**2.1 Особенности задачи прогнозирования временных рядов** 6](#_Toc26225712)

[**2.2 Прогнозирование временных рядов на примере метеонаблюдений с 2009 по 2016 год** 6](#_Toc26225713)

[**Выводы** 14](#_Toc26225714)

[**Список источников** 15](#_Toc26225715)

# **Введение**

В работе рассматривается применение ансамблей регрессионных моделей для задач прогнозирования временных рядов на примере предсказания температуры основываясь на данных, зарегистрированных метеостанцией при биоинженерном институте Макса Планкта в Йене (Германия). Периодичность фиксации исходных данных составляет 10 минут [3].

При выполнении работы будут использованы модели машинного обучения из библиотеки scikit-learn для языка Python.

# **1. Метрики оценки качества регрессии**

Наиболее простым способом оценки качества регрессионной модели, особенно для несложных задач, является коэффициент детерминации .

Где,

x – признаки модели

y – случайная величина, ответы модели

условная по x-признакам дисперсия зависимой переменной, или дисперсия случайных ошибок ML-модели.

принимает значение от 0 до 1, чем ближе показатель к единице, тем выше обобщающая способность модели. Модель признается приемлемой для применения если вычисленный для неё коэффициент детерминации как минимум не меньше 0.5, достаточно хорошими считаются модели с коэффициентом детерминации от 0.8 и выше. Коэффициент детерминации прост в интерпретации, но не всегда достоверно показывает обобщающую способность модели [2].

Поэтому при принятии бизнес решений как правило учитываются и другие способы оценки качества регрессии, в частности средняя абсолютная и средняя квадратическая ошибки, а также визуализация предсказанных моделью значений [1].

В частности, MSE (Mean Squared Error) – среднеквадратическое отклонение определяется предсказанной величины от ее истинного значения, определяется как:

Вариацией среднеквадратического отклонения является RMSE (Root Mean Squared Error) – корень из среднеквадратического отклонения.

Основным отличием RMSE от MSE является более значительная величина штрафа за большие по модулю отклонения от правильных ответов.

# **2. Анализ временных рядов**

## **2.1 Особенности задачи прогнозирования временных рядов**

Временной ряд – это последовательность значений признака y вида , измеряемая через постоянные временные интервалы, фиксируемые с неизменной периодичностью.

Задача прогнозирование временного ряда – на основании исторических сведений о данных (например измерений температуры за определенный промежуток времени) предсказать значение температуры в будущем.

Главная особенность задачи прогнозирования временных рядов и одновременно основное отличие от “классических” задач анализа данных (таких как классификация) – необходимость в наличии явной зависимости между архивными значениями ряда и его будущем поведении [4].

При анализе временных рядов стоит обратить внимание на компоненты, непосредственно образующие временной ряд:

* Тренд – долгосрочное направление изменение уровня ряда. В честности трендом можно назвать заявления о глобальном потеплении и изменении климата.
* Сезонность – циклические изменения ряда с постоянной периодичности, в частности сезонные колебания температуры в зависимости от времени года.
* Цикл – изменения ряда с динамической периодичность. Примером может послужить “одиннадцатилетний” цикл солнечной активности, чью периодичность является весьма условной, так как способна динамически изменяться, например продолжительность солнечного цикла в XVIII—XX веках колебалась от семи до семнадцати лет.
* Ошибка – случайная компонента временного ряда, не поддающаяся прогнозированию, фактически является шумом или выбросом.

## **2.2 Прогнозирование временных рядов на примере метеонаблюдений с 2009 по 2016 год**

Исходные данные содержат метеоданные с 2009 по 2016 года с периодичностью в 10 минут. Данные характеризуются такими признаками как дата регистрации, атмосферное давление, влажность и так далее, Рисунок 1.

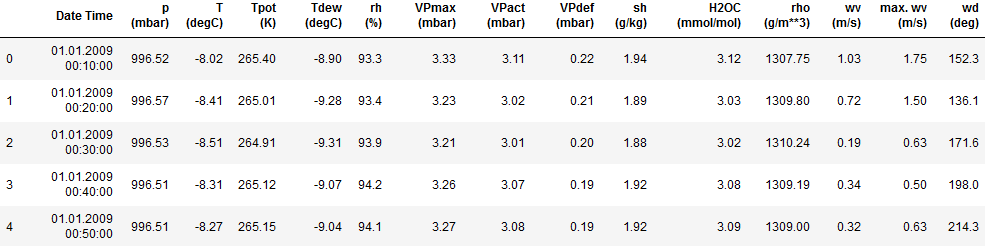


Рисунок 1 – первые 5 строк исходных данных

В рамках решения задачи будут рассматриваться только дата измерения и извлеченные из неё признаки (X) и значение температуры (y), являющееся целью предсказания, Рисунок 2.

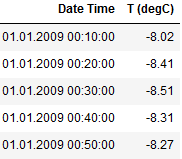


Рисунок 2 первые 5 строк данных для анализа

Распределение исходных данных можно визуализировать в виде графика, где по оси y находится значение температуры, а по оси x – номер наблюдения. По распределению можно сделать вывод о наличии явных сезонных колебаний, при этом имеется незначительный тренд на увеличение температуры. При этом в составе временного ряда преобладают измерения с положительной и высокой температурой, тогда как примеров близкой к нулю и отрицательной температуры существенно меньше, что способно затруднить прогнозирование низких температур, Рисунок 3

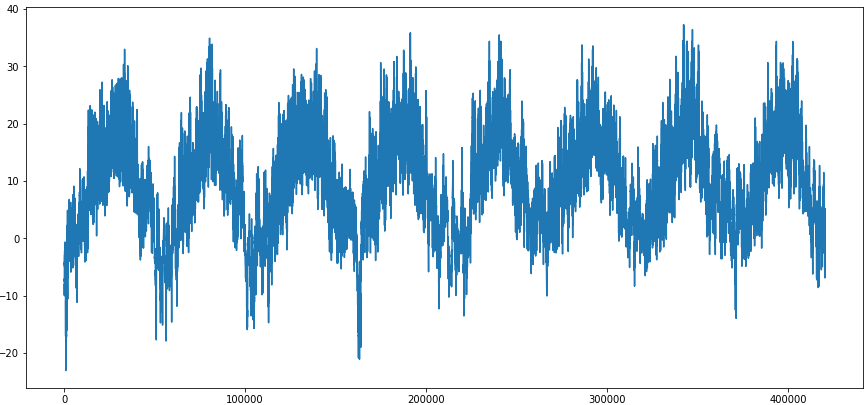


Рисунок 3 Визуализация распределения исходных данных

При построении моделей рассматривались данные в исходном виде, а также агрегированные как среднее значение температуры за определенный период – 1 час, 3 часа, 6 часов, 12 часов, 1 день, 3 дня, 7 дней, 14 дней.

Для прогнозирования временного ряда применялись ансамбли моделей из библиотеки scikit-learn – случайный лес и градиентный бустинг. Результаты моделей на тестовой выборке представлены в таблице 1.

Таблица 1 Результаты применения моделей на данных, агрегированных с различной периодичностью

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Период** | **Модель** | **(train)** | **(test)** | **MSE** | **RMSE** |
| 10 минут | Случайный лес | 0.769 | 0.648 | 22.68 | 4.762 |
| 10 минут | Градиентный бустинг | 0.8 | 0.685 | 20.29 | 4.505 |
| 1 час | Случайный лес | 0.776 | 0.655 | 22.129 | 4.704 |
| 1 час | Градиентный бустинг | 0.795 | 0.691 | 19.829 | 4.453 |
| 3 часа | Случайный лес | 0.784 | 0.665 | 21.301 | 4.615 |
| 3 часа | Градиентный бустинг | 0.798 | 0.703 | 18.886 | 4.346 |
| 6 часов | Случайный лес | 0.795 | 0.674 | 20.247 | 4.5 |
| 6 часов | Градиентный бустинг | 0.804 | 0.704 | 18.394 | 4.289 |
| 12 часов | Случайный лес | 0.815 | 0.689 | 18.582 | 4.311 |
| 12 часов | Градиентный бустинг | 0.815 | 0.706 | 17.557 | 4.19 |
| 1 день | Случайный лес | 0.819 | 0.708 | 15.993 | 3.999 |
| 1 день | Градиентный бустинг | 0.811 | 0.706 | 16.129 | 4.016 |
| 3 дня | Случайный лес | 0.864 | 0.768 | 9.964 | 3.157 |
| 3 дня | Градиентный бустинг | 0.862 | 0.771 | 9.819 | 3.134 |
| 7 дней | Случайный лес | 0.914 | 0.809 | 5.929 | 2.435 |
| 7 дней | Градиентный бустинг | 0.916 | 0.798 | 6.269 | 2.504 |
| 10 дней | Случайный лес | 0.941 | 0.754 | 6.571 | 2.563 |
| 10 дней | Градиентный бустинг | 0.946 | 0.768 | 6.189 | 2.488 |
| 14 дней | Случайный лес | 0.95 | 0.809 | 3.895 | 1.974 |
| 14 дней | Градиентный бустинг | 0.965 | 0.829 | 3.489 | 1.868 |

Дополнительные сведения о качестве моделей способна показать визуализация прогнозирования на тестовых данных, где синим цветом будут обозначены результаты наблюдений (правильные ответы), а пунктирной красной линией – прогноз, сделанный моделью.

Первыми оцениваются модели по исходным данным (интервал – 10 минут), модель “Случайный лес” показывает приемлемое, но далекое от совершенства качество – коэффициент детерминации равен 0.648. Но визуализация результата дает больше сведений, модель в целом успешно определяет динамику сезонных колебаний, но результат выглядит сильно усредненным. Более того, как и предполагалось существуют существенные проблемы с прогнозированием отрицательных температур, Рисунок 4.

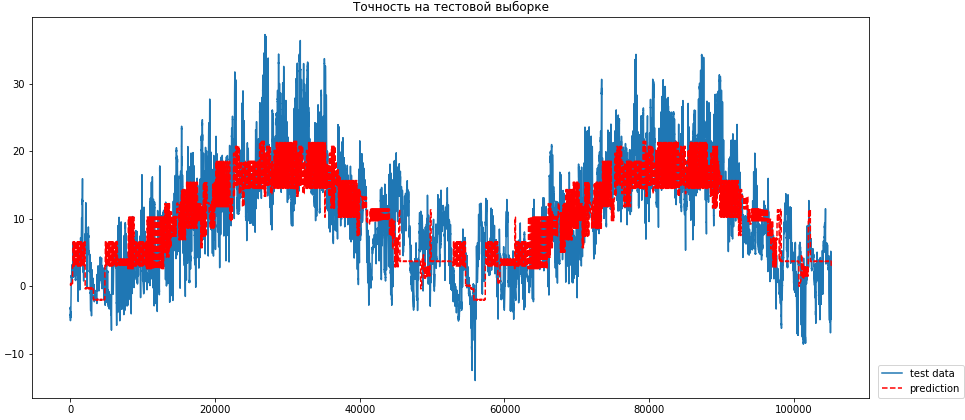


Рисунок 4 Визуализация прогнозирования с помощью случайного леса на исходных данных

Модель на базе градиентного бустинга показывает незначительно лучшее качество, сравнивая со случайным лесом, но сохраняет аналогичные недостатки относительно сильно усредненного результата и проблемы с предсказанием температуры ниже нуля, Рисунок 5.

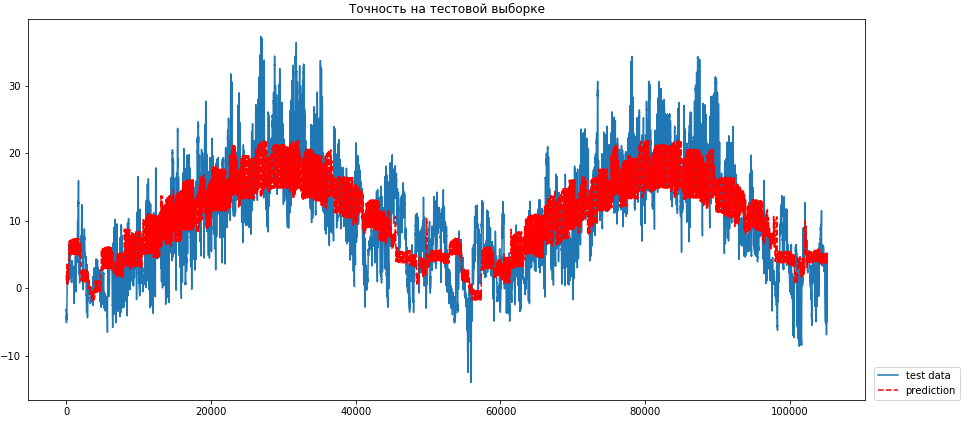


Рисунок 5 Визуализация прогнозирования с помощью градиентного бустинга на исходных данных

На следующем этапе рассматривается прогнозирование усредненной температуры с интервалом в 1, 3, 6 и 12 часов.

Случайный лес показывает несколько лучший результат, сравнивая с предсказаниями при интервале в 10 минут. Модель начинает предсказывать колебания температуры с меньшим усреднением, но сохраняется проблема с отрицательной температурой, Рисунок 6

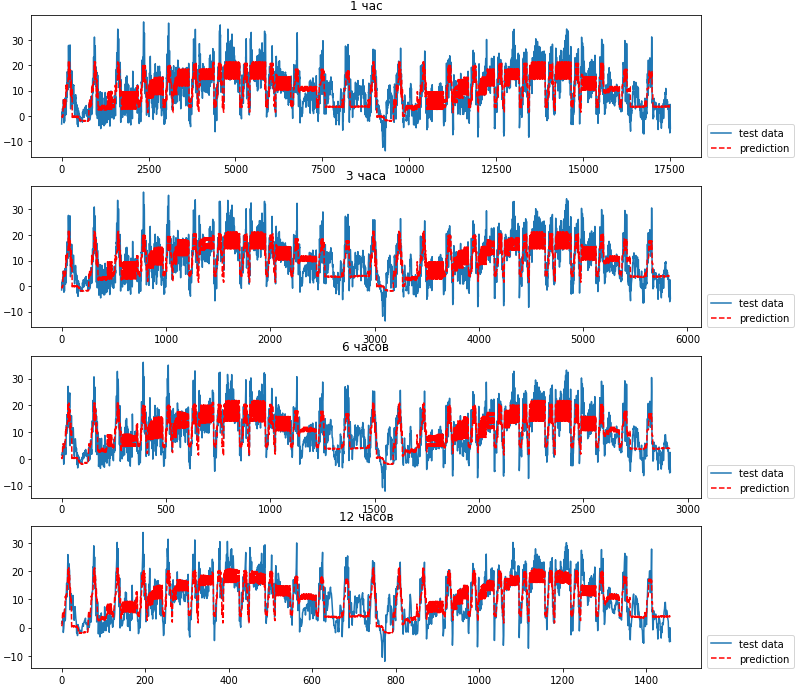


Рисунок 6 Прогнозирование температуры с интервалами в 1, 3, 6 и 12 часов (Случайный лес)

Градиентный бустинг показывает в целом лучшие результаты, графики показывает, что модель существенно лучше предсказывает колебания температуры, на что также указывают используемые для оценки качества метрики. Но, к сожалению, улучшение касается исключительно прогнозирования температуры выше нуля. Полностью сохраняется нулевая обобщающая способность модели для отрицательных значений, Рисунок 7.

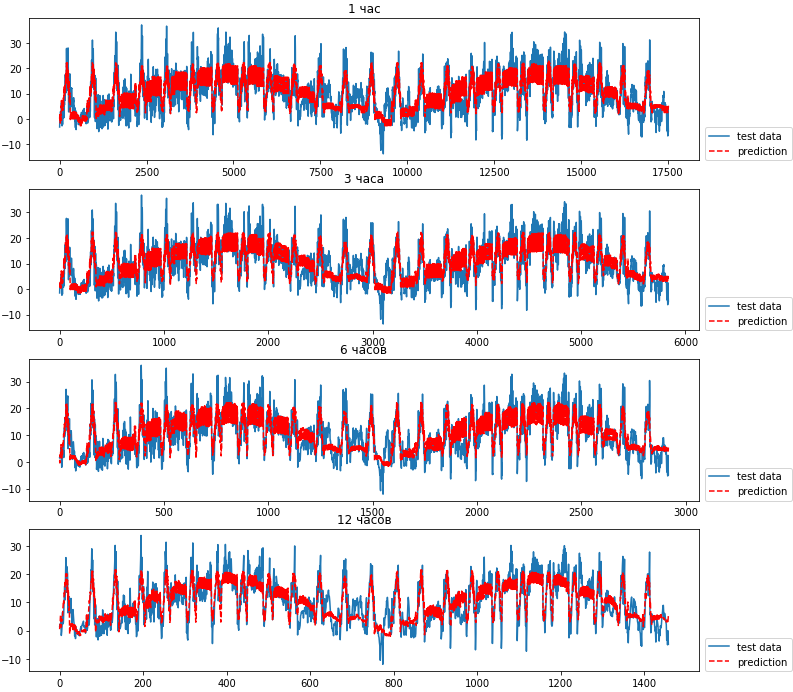


Рисунок 7 Прогнозирование температуры с интервалами в 1, 3, 6 и 12 часов (Градиентный бустинг)

На финальном этапе оценивается обобщающая способность моделей для предсказания средней температуры с интервалами в 1, 3, 7, 10 и 14 дней. Все метрики оценки качества показывают высокую обобщающую способность для подобных периодов. В частности, точность прогнозирования средней температуры для интервала в 14 дней согласно составляет 0.809 и 0.829 для случайного леса и градиентного бустинга соответственно, что говорит о хорошем качестве моделей. Также на хорошую обобщающую способность указывают низкие значения ошибки (RMSE и MSE), но визуализация позволяет получить более полную картину.

Визуализация модели на базе случайного леса показывает обратное. На графике заметно улучшение качества модели, которая существенно лучше предсказывает температурные колебания из тестовой выборки. Но эффект во многом обусловлен усреднением данных, так, например из-за усреднения данных для интервалов в 10 и 14 дней пропали отрицательные значения температуры, с предсказанием которых с самого начала были проблемы, что создало ложную иллюзию высокой обобщающей способности, рисунок 8.

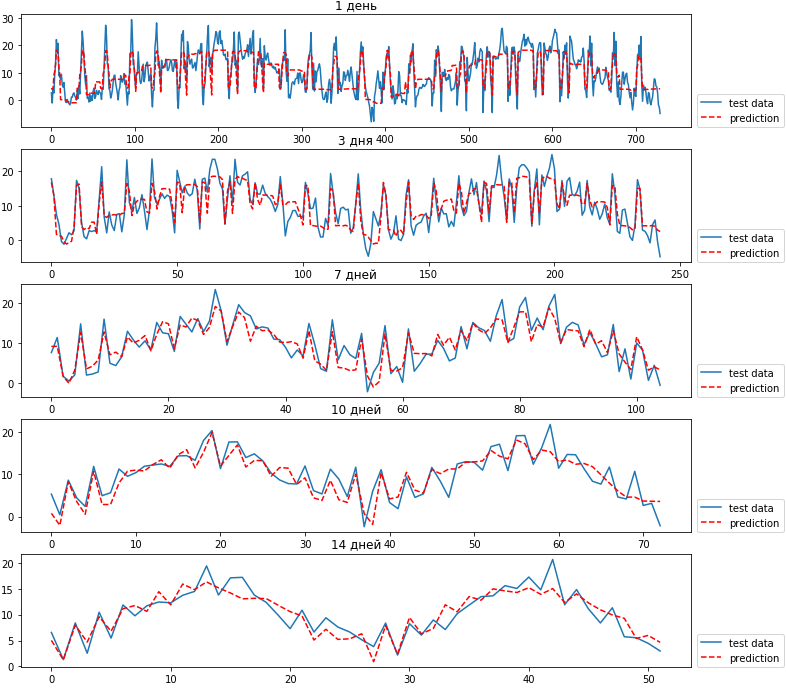


Рисунок 8 Прогнозирование усреднённой температуры с интервалами в 1, 3, 7, 10, 14 дней (Случайный лес)

Модель на база градиентного бустинга имеет аналогичные проблемы, так как они в первую очередь обусловлены исходными данными, рисунок 9.

В результате можно сделать вывод о нецелесообразности прогнозирования усредненного значения температуры за интервал, превышающий 12 часов.

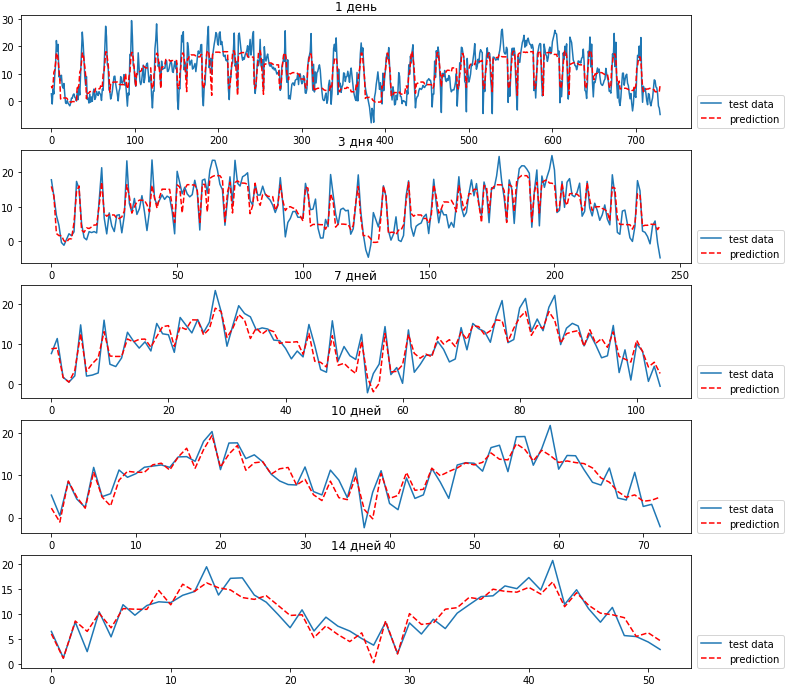


Рисунок 9 Прогнозирование усреднённой температуры с интервалами в 1, 3, 7, 10, 14 дней (Градиентный бустинг)

# **Выводы**

В результате выполнения работы настроены модели случайного леса и градиентного бустинга с целью прогнозирования временных рядов на примере данных с метеостанции Макса Плантка, размещенной в Йене.

Проверка моделей на тестовой выборке показывает в целом большую точность градиентного бустинга, при этом для всех моделей актуальная проблема с неэффективным предсказанием отрицательных значений температуры. С учетом данной проблемы и в целом приемлемой, но невысокой точности моделей наиболее точные прогнозы получаются для при предсказании средней температуры за интервал в 12 часов.

При этом прогнозирование с больший интервалом усреднения нецелесообразно, так как во первых предсказание средней температуры на следующие 10-14 дней практически лишено практического смысла и во вторых – при усреднении температуры за день и более из исторических данных пропаяют представленные в меньшем объеме температурные значения ниже нуля, что делает исходные данные недостоверными.

# **Список источников**

1. Андреас Мюллер, Сара Гвидо. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. — СПб.: Диалектика, 2017. — 480 с.: ил.
2. Коэффициент детерминации URL: -- <http://www.machinelearning.ru/wiki/>
3. Метеоданных, зарегистрированных метеостанцией при биоинженерном институте Макса Планкта URL: -- <https://www.kaggle.com/pankrzysiu/weather-archive-jena>
4. Coursera, специализация “Машинное обучение и анализ данных”, курс “Прикладные задачи анализа данных” временные ряды URL: -- https://www.coursera.org/lecture/data-analysis-applications/vriemiennyie-riady-EjNEV