Вестник РУДН. Серия: Инженерные исследования
RUDN Journal of Engineering Research

2022;23(2):97–107

ISSN 2312-8143 (Print); ISSN 2312-8151 (Online)
journals.rudn.ru/engineering-researches

DOI 10.22363/2312-8143-2022-23-2-97-107
УДК 004.85

Научная статья / Research article

Современные аспекты применения искусственного интеллекта для прогнозирования стихийных бедствий на реках Российской Федерации (на примере реки Амур)

Н.Э. Александров[a] ✉, Д.Н. Ермаков[a,b], А.Е. Бром[c], И.Н. Омельченко[c], С.В. Шкодинский[a,d]

[a]Российский университет дружбы народов, *Москва, Российская Федерация*
[b]АО «НИИ „Полюс“ имени М.Ф. Стельмаха», *Москва, Российская Федерация*
[c]Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет), *Москва, Российская Федерация*
[d]Московский государственный областной университет, *Мытищи, Российская Федерация*
✉ 1042210208@rudn.university

История статьи
Поступила в редакцию: 10 марта 2022 г.
Доработана: 28 мая 2022 г.
Принята к публикации: 4 июня 2022 г.

Ключевые слова:
управление катастрофами, предсказание паводков, река Амур, машинное обучение, линейная регрессия, нейронная сеть, градиентный бустинг

Аннотация. Среди всех наблюдаемых природных стихийных бедствий катастрофы, связанные с водой, наиболее частые и несут серьезную опасность для людей и социально-экономического развития. Для России наибольшую актуальность представляют речные паводки, важность борьбы с которыми, в частности на Дальнем Востоке, неоднократно подчеркивал президент РФ В.В. Путин. Изучено качество работы различных методов искусственного интеллекта по предсказанию речных паводков в бассейне реки Амур. Уникальность исследования заключается в том, что прежде подобных изысканий для этой реки не проводилось. Основная задача состояла в последующем практическом применении полученных результатов в системах прогнозирования паводков и управления их риском. С этой целью поиск наилучшего метода выполнялся среди широко используемых на рынке методов, обладающих богатым выбором вспомогательных решений: градиентный бустинг на деревьях, линейная регрессия без регуляризации и нейронные сети. В дизайне исследования сделан упор на достижение максимальной воспроизводимости результатов. В итоге наивысшее качество показал градиентный бустинг над деревьями в отечественной реализации CatBoost. Полученные результаты могут быть экстраполированы и на другие реки, сравнимые как по площади, так и по объему собранных данных.

Для цитирования
Александров Н.Э., Ермаков Д.Н., Бром А.Е., Омельченко И.Н., Шкодинский С.В. Современные аспекты применения искусственного интеллекта для прогнозирования стихийных бедствий на реках Российской Федерации (на примере реки Амур) // Вестник Российского университета дружбы народов. Серия: Инженерные исследования. 2022. Т. 23. № 2. С. 97–107. http://doi.org/10.22363/2312-8143-2022-23-2-97-107

Modern aspects of the use of artificial intelligence for predicting natural disasters on the rivers of the Russian Federation (using the example of the Amur River)

Nikita E. Aleksandrov[a], Dmitry N. Ermakov[a,b], Alla E. Brom[c], Irina N. Omelchenko[c], Sergey V. Shkodinsky[a,d]

[a]Peoples' Friendship University of Russia (RUDN University), *Moscow, Russian Federation*
[b]Polyus Scientific Research Institute, *Moscow, Russian Federation*
[c]Bauman Moscow State Technical University, *Moscow, Russian Federation*
[d]Moscow State Regional University, *Mytishchi, Russian Federation*
✉ 1042210208@rudn.university

Article history
Received: March 10, 2022
Revised: May 28, 2022
Accepted: June 4, 2022

Keywords:
disaster management, floods forecasting, Amur River, machine learning, linear regression, neural network, gradient boosting

Abstract. Among all observed natural disasters, water-related disasters are the most frequent and pose a serious threat to people and socio-economic development. River floods are the most relevant for the Russian Federation, and the importance of flood control, particularly in the Far East, was repeatedly stressed by Russian President Vladimir Putin. The quality of performance of various artificial intelligence methods on the task of predicting river floods in the Amur River basin was investigated. The uniqueness of the research lies in the fact that similar studies have not previously been conducted for this river. The main task of the work was the subsequent practical application of the obtained results in flood forecasting and risk management systems. For this purpose, the best method was searched among widely used methods on the market, which have a rich choice of auxiliary solutions: gradient tree binning, linear regression without regularisation and neural networks. The study design focus on achieving maximum reproducibility of the results. The gradient boosting over the trees in the domestic implementation of CatBoost showed the highest quality. The results of this work can be extrapolated to other rivers comparable in both area and volume of data collected.

For citation
Aleksandrov NE, Ermakov DN, Brom AE, Omelchenko IN, Shkodinsky SV. Modern aspects of the use of artificial intelligence for predicting natural disasters on the rivers of the Russian Federation (using the example of the Amur River). *RUDN Journal of Engineering Research*. 2022;23(2):97–107. (In Russ.) http://doi.org/10.22363/2312-8143-2022-23-2-97-107

Введение

Постановка проблемы. Среди всех наблюдаемых природных стихийных бедствий, катастрофы, связанные с водой, наиболее частые и представляют серьезную опасность для людей и социально-экономического развития. Согласно [1], в период с 1900 по 2006 г. всевозможные виды наводнений были ответственны за 30 % от общего числа стихийных бедствий, 19 % от общего числа погибших и 48 % от общего числа пострадавших. В этом же отчете утверждается, что природные катастрофы, связанные с водой, ответственны за 72 % от общего экономического ущерба, причиненного стихийными бедствиями, из которых 26 % – это наводнения. Во время селекторных совещаний с членами правительства президент РФ В.В. Путин неоднократно подчеркивал важность борьбы с паводками на Дальнем Востоке: «В зону возможного подтопления могут попасть до пяти тысяч населенных пунктов, это порядка 1,5 млн человек. Какие бы сюрпризы нам природа ни преподносила, мы, безусловно, должны быть готовы к любому варианту развития событий»[1]. В связи с изменением климата ожидается увеличение числа потерь от таких явлений. Таким образом, важно улучшать качество принятия решений при реагировании на наводнения.

Разработка систем прогнозирования и управления риском наводнений рекомендуется в качестве одной из мер подготовки к ним [2] по нескольким причинам. Во-первых, из-за неопределенности, связанной с силой, временем и местом наводнений,

[1] *Замахина Т.* Президент заявил о непростой обстановке с паводками и пожарами // Российская газета. 2021, 31 марта. URL: https://rg.ru/2021/03/31/putin-zaiavil-o-neprostoj-obstanovke-s-pavodkami-i-pozharami.html?ysclid=l6j9qiq7yf690189806 (дата обращения: 12.05.2022).

зачастую невозможно полностью контролировать их, и, как следствие, абсолютная защита от этого явления не всегда возможна [3]. Во-вторых, традиционные методы управления риском наводнений в основном состоят из структурных мер защиты, таких как дамбы и плотины, изменяющих характеристики наводнения для уменьшения пикового уровня воды и снижения масштаба разлива. Несмотря на то что структурные меры снижают риск наводнения, они не могут полностью устранить его. К тому же на практике данные меры защиты невозможно внедрить в некоторых областях: например, в отдаленные поселениях Сибири и на Дальнем Востоке. Также они могут приводить к нежелательным экологическим последствиям [4]. Следовательно, возведение структурных мер защиты не всегда целесообразно, в таких случаях предиктивные модели могут служить более простой в имплементации и дешевой альтернативой [5]. Можно заключить, что разработка и улучшение методов прогнозирования наводнений важны для решения задач управления и принятия решений при реагировании на паводки.

Данная работа фокусируется на поиске наилучшего метода машинного обучения для моделирования паводков на р. Амур, где они наносят значительный ущерб населению и экономике региона [6]. Исследование предпринято с целью улучшения методов прогнозирования паводков для последующего использования результатов исследования в решении задач управления при реагировании на паводки.

Существующие методы. Зачастую модели, прогнозирующие паводки, предсказывают будущий уровень воды или скорость потока. Классические методы, используемые в гидрологии, основаны на определении зависимостей между метеорологическими данными, характеристиками бассейна, субстрата и смоделированными целевыми значениями [7]. Существует множество подходов к моделированию природных процессов, имплементированных в виде гидрологических моделей. Такие модели могут быть основаны как на детерминистических, так и на стохастических подходах. Большинство моделей не учитывают специфику региона, а основаны на некоторых общих, характерных для любой реки принципах.

Для адаптации моделей к специфике региона прибегают к их калибрации [8]. Это позволяет снизить ошибку предсказания. Но калибрация сложных моделей может быть вычислительно слишком сложной. К тому же сложные модели требуют большого количества данных, которых может быть недостаточно или не быть совсем. Калибрация влечет и ряд прочих трудностей, делая применение таких моделей неэффективным в некоторых случаях [9].

Другой способ моделирования паводков – описание паттернов потоков воды с помощью дифференциальных уравнений. Недостаток таких методов заключается в нестабильности решений, вызванных накоплением ошибки и высокой вычислительной сложностью. К тому же такие модели могут быть трудно переносимы на другие реки, для которых им могут потребоваться дополнительные параметры.

В моделировании паводков хорошо себя показывают методы машинного обучения, которым удается достигать высокой точности в этой задаче [10]. У них есть несколько важных практических плюсов: во-первых, очень развитая сопутствующая техническая экосистема [11], что существенно упрощает разработку систем на основе таких методов; во-вторых, наличие в России множества специалистов по машинному обучению[2], что дает возможность найти людей для создания промышленного решения на основе таких методов. В последние годы были созданы методы, позволяющие интерпретировать предсказания любых моделей машинного обучения. Например, метод SHAP [12], использующий подход из кооперативной теории игр и позволяющий проинтерпретировать отдельное предсказание. Из минусов этого подхода можно выделить необходимость большого количества данных для наиболее передовых моделей машинного обучения. На основе описанных доминирующих положительных факторов было решено сфокусироваться в данной работе именно на методах машинного обучения.

Научная новизна исследования заключается не только в определении наилучшего метода машинного обучения для предсказания паводков на р. Амур, но еще и в изучении работы метода CatBoost [13], созданного российской компанией «Яндекс», что актуально в условиях санкций и импортозамещения технологий.

Описание данных. Данные для моделирования и тестирования были предоставлены Сбербанком совместно с МЧС, Минприроды и Росгидрометом в рамках хакатона по разработке решений для предсказания паводков[3].

[2] Академия больших данных MADE и hh.ru составили портрет российского специалиста в сфере Data Science. 2020. URL: https://vk.company/ru/press/releases/10682/ (дата обращения: 04.03.2022).

[3] NoFloodWithAI: прогнозирование паводков на реке Амур. URL: https://github.com/sberbank-ai/no_flood_with_ai_aij2020 (дата обращения: 04.03.2022).

Река Амур является трансграничной, основная часть бассейна находится в пределах Российской Федерации. Для Амура характерна низкая водность в зимний период, небольшие половодья весной и неоднократные резкие подъемы воды во второй половине лета и в начале осени. Маловодные периоды сменяются годами большой воды [14]. В многолетнем режиме водного стока Амура отчетливо выражено чередование периодов пониженной и повышенной водности, каждый продолжительностью 10–15 лет [15]. Амур, по оценке гидрологов и исходя из истории наблюдений, вошел в очередной период высокой водности в конце 2000-х гг. Основываясь на гидрологической закономерности режима Амура, в ближайшие 5–7 лет следует ожидать сложную паводковую обстановку в течении Среднего и Нижнего Амура (наиболее сложная обстановка от слияния р. Сунгари и до Комсомольского района включительно). Наиболее крупномасштабные наводнения произошли в 2013 и 2019 гг., их причиной стали тропические циклоны, которые несли теплый влажный воздух, вызывали фронтальные разделы и сильные атмосферные осадки. В 2013 г. на значительной площади за 2–3 месяца сумма выпавших осадков превысила годовую, а местами и полуторагодовую норму. На рис. 1 представлена карта бассейна р. Амур.

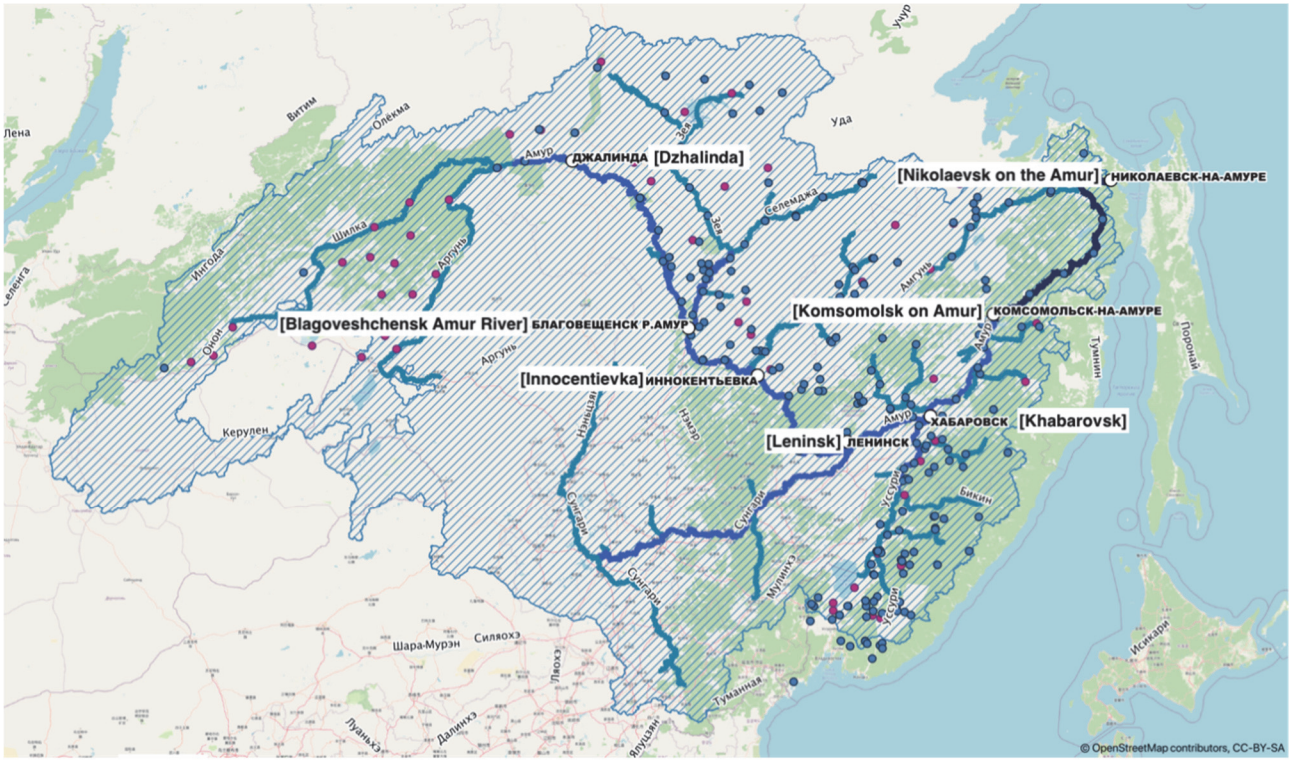

Рис. 1. Бассейн р. Амур и его основные притоки
Figure 1. The Amur River basin and its main tributaries

В наборе данных представлены наблюдения на 198 гидрологических постах сети Росгидромета за период с 1984 по 2018 г., содержащие данные об уровнях воды, расходах, температуре воды, наблюдения за поверхностью воды (становление ледостава, вскрытие). Сведения об уровнях воды описаны тремя величинами: минимальный, максимальный и средний уровни воды за день. На рис. 2 представлено несколько примеров данных по гидрологическому посту под идентификатором 5001. Отсутствие значений на графике означает пропуск.

Как видно из графиков, временной ряд уровня воды обладает сезонностью с периодом в один год, что полностью согласуется с контекстом задачи. Также видно, что в приведенном временном ряду имеются пропуски: например, их много с 1985 по 1990 г. на рис. 3. В 33 % наблюдений пропущены значения температуры воды, в 41 % – нет данных по потреблению воды, менее чем в 1 % – отсутствуют данные по уровню воды.

Задача – определить модель машинного обучения для предсказания уровня воды на 10 дней вперед.

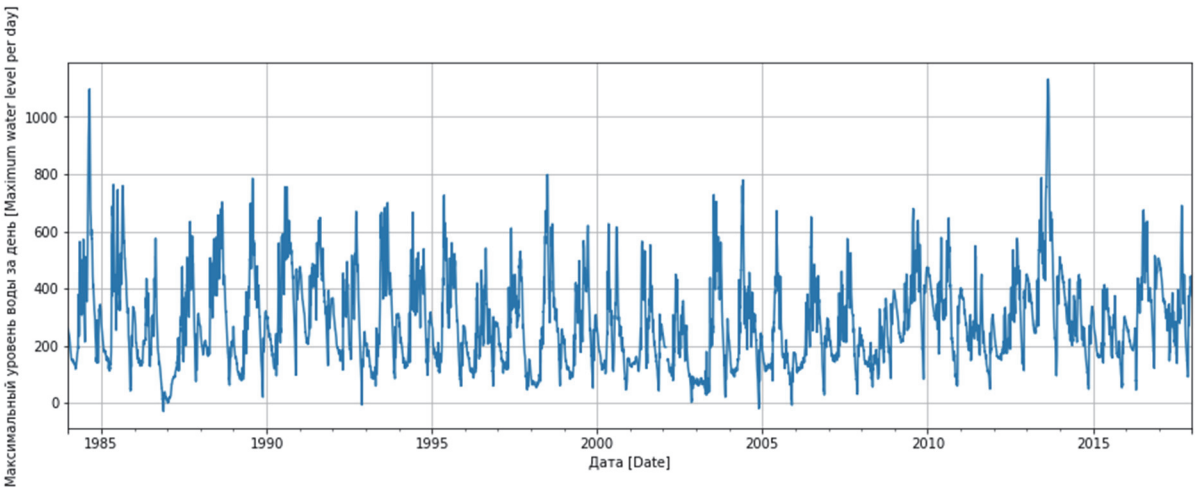

Рис. 2. Дневные значения максимального уровня воды для датчика 5001
Figure 2. Daily values of the maximum water level for the sensor 5001

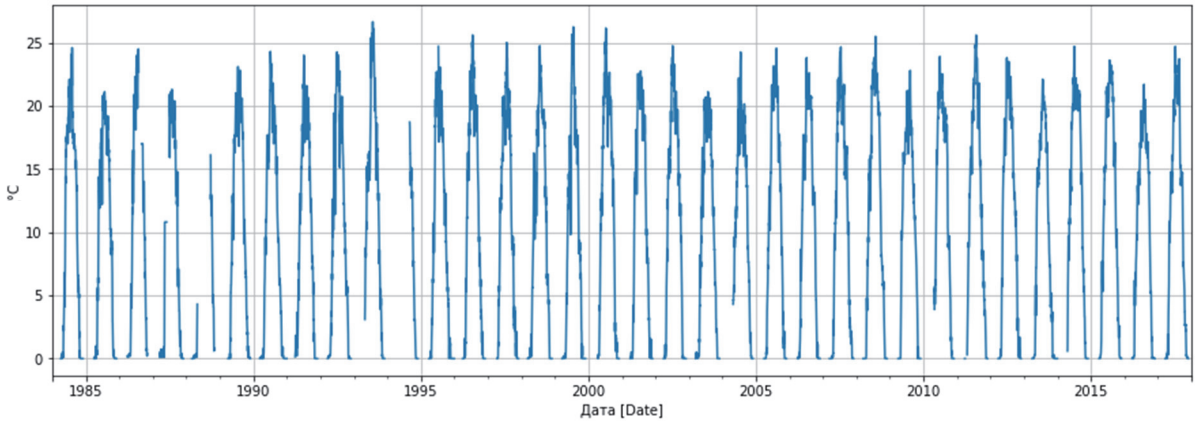

Рис. 3. Дневные значения среднедневной температуры
Figure 3. Daily mean temperatures

1. Методы решения задачи

Исследовано применение трех алгоритмов машинного обучения для предсказания паводков на р. Амур: линейная регрессия, нейронные сети и градиентный бустинг в реализации CatBoost. Данные алгоритмы выбраны по следующим практическим соображениям: во-первых, они обладают крайне развитой экосистемой вспомогательных решений; во-вторых, подавляющее большинство специалистов по машинному обучению умеет с ними работать. Это позволят использовать полученные результаты для быстрого построения промышленной системы.

Линейная регрессия – это одна из наиболее изученных и распространенных статистических моделей, описывающая зависимость целевой переменной y от другой или нескольких других целевых переменных x через линейную зависимость.

Регрессионная модель описывается следующим уравнением:

$$y = f(x, b) + \varepsilon, E[\varepsilon],$$

где b – параметры модели; ε – случайная ошибка модели, а $f(x, b)$ имеет вид

$$f(x, b) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k,$$

где b_j – параметры регрессии; x_j – регрессоры; k – количество факторов модели [16].

Параметры подбираются через минимизацию квадратичной ошибки на обучающей выборке:

$$\min_b \sum_{i=1}^{N} (y_i - f(x_i, b)),$$

где i – номер объекта из обучающей выборке; N – размер обучающей выборке.

В данной работе использована реализация линейной регрессии из библиотеки для языка программирования Python scikit-learn[4].

Нейронная сеть – математическая модель и ее программная реализация, созданная на основе принципов организации и функционирования биологических нейронных сетей[5]. «Нейрон» в сети получает сигнал в виде вектора действительных чисел, обрабатывает его и отдает одно действительное число, называемое сигналом. Сигнал является результатом вычислений нелинейной функции над взвешенной суммой входных значений. Благодаря этому нейронные сети в отличие от линейной регрессии способны распознавать нелинейные закономерности в данных. Нейроны между слоями связаны друг с другом, и каждая связь в каждом нейроне имеет вес. Вес в нейронах изменяется во время обучения нейронной сети и отвечает за усиление или ослабление сигнала в соединении.

Нейронные сети обучаются с помощью алгоритма обратного распространения ошибки [17]. Его суть в расчете градиента функции потерь относительно веса нейронной сети для одного обучающего экземпляра, а затем изменения веса значений веса в направлении антиградиента функции потерь. Эффективность данного метода позволяет использовать его для обучения многослойных нейронных сетей.

Будем проверять нейронную сеть с одним полно связным слоем размером 100, функцией активации ReLU [18] и оптимизатором Adam. Реализация этой модели будет взята из Python scikit-learn[6].

Градиентный бустинг (CatBoost). В работе использована разновидность градиентного бустинга, называющаяся градиентный бустинг над деревьями.

Градиентный бустинг над деревьями представляет собой ансамбль деревьев решений. В основе его алгоритма лежит итеративное обучение деревьев решений с целью минимизации функции потерь. Благодаря особенностям деревьев решений градиентный бустинг способен работать с категориальными признаками и справляться с нелинейными закономерностями в данных.

В работе будет исследована реализация этого алгоритма из библиотеки CatBoost, одним из преимуществ которой является умение работать с пропусками в данных.

2. Эксперименты

Подготовка данных и их разбиение. Для выявления оптимальной модели использовались только исторические данные об уровнях воды и толщине льда без сложных объясняющих признаков, поскольку они избыточны для задачи определения наилучшего алгоритма машинного обучения, а построение наиболее точной модели не являлось целью исследования. Объясняющие признаки описывают сезонность временных рядов и распределение изменений их значений за различные промежутки времени.

Как и в любых реальных данных до расчета фич и обучения моделей была выполнена очистка данных. В датасете обнаружены явные ошибки и выбросы, выглядящие, например, как на рис. 4.

Сначала были удалены явные ошибки в данных. Это наблюдения, в которых минимальный уровень воды за день превышал максимальный, средний уровень превышал максимальный, средний уровень был меньше минимального, потому что с точки зрения математики такое невозможно. В таких наблюдениях показателям, отвечающим за описание уровня воды, были присвоены значения NaN. Всего таких наблюдений – 0,04 %.

Значения меньше 0 не будут удаляться, потому что они являются результатом неправильного выбора нуля графика гидрологического поста. Обычно на практике за нуль графика принимается значение на 0,5 м ниже наблюдавшегося уровня воды, и, возможно, из-за обмеления реки нуль мог уменьшиться. Следовательно, несмотря на наличие отрицательных значений динамика изменения уровня воды должна оставаться корректной. Наглядно это можно увидеть на рис. 2, где периодически наблюдаются значения меньше 0, но визуально динамика уровня воды выглядит корректно. Всего таких наблюдений около 5 %.

[4] Sklearn.linear_model. Linear regression. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (дата обращения: 02.02.2022).

[5] Нейронная сеть // Большая российская энциклопедия: в 35 т. Т. 35. / гл. ред. Ю.С. Осипов. М.: Большая российская энциклопедия, 2017.

[6] Scikit-learn. 1.17. Neural network models (supervised). URL: https://scikit-learn.org/stable/modules/neural_networks_supervised.html (дата обращения: 02.02.2022).

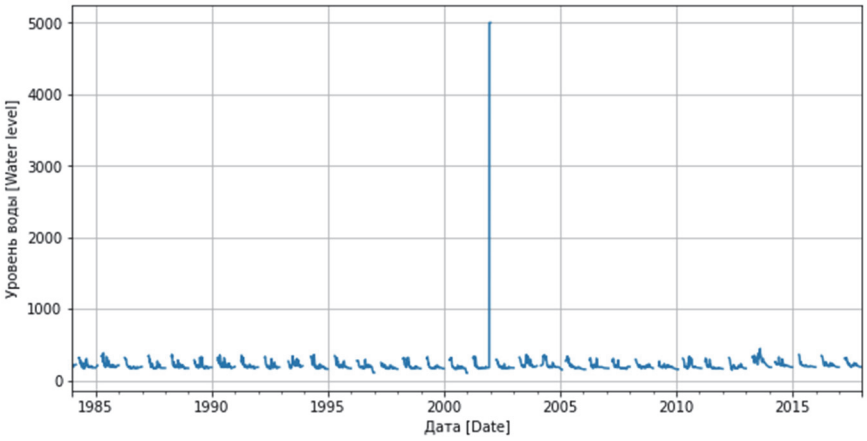

Рис. 4. Ежедневные значения минимального уровня воды для датчика 6535
Figure 4. Daily minimum water levels for 6535 sensor

При возможности выполнялось восполнение пропусков в показателях максимального уровня воды следующим образом: в наблюдениях с пропусками по максимальному уровню воды, но с заполненными значениями по среднему и минимальному уровням и при их равенстве, максимальному уровню присваивалось то же значение. Потому что минимум может быть равен среднему только при условии, что среднее считалось по одинаковым значения. Всего таких наблюдений около 0,83 %.

Далее была посчитана целевая переменная, равная максимальному уровню воды через 10 дней. Построена одна модель для всех датчиков по следующим причинам: во-первых, так модель получает больше данных для обучения; во-вторых, данные по многим датчикам выполняют некоторую регуляризацию модели, так как ей придется выучивать паттерны, которые работают в большинстве мест реки.

Все наблюдения, для которых отсутствовала целевая переменная удалены. Далее посчитаны дополнительные объясняющие признаки:

1) изменение максимального уровня, температуры и потребления воды за 1, 5, 10, 15, 20, 30, 50, 60, 180, 365 дней, чтобы учесть влияние сезонности и индивидуальные магнитуды изменений;

2) среднее однодневных изменений уровня воды за 7, 30, 90, 365 дней;

3) среднее однодневных изменений температуры за 7, 30, 90 дней;

4) среднее однодневных изменений потребления воды за 7 дней;

5) стандартное отклонение однодневных изменений уровня воды за 7, 30, 90, 365 дней;

6) стандартное отклонение однодневных изменений температуры за 7, 30, 90 дней;

7) стандартное отклонение однодневных изменений потребления воды за 7 и 30 дней.

Все дополнительные объясняющие признаки выше и температура воды вошли в финальный список фич, на которых обучаются модели и делается предсказание.

В конце данные были разбиты на обучающую и тестовую выборки. В обучающую вошли первые 80 % дат, все остальные даты вошли в тестовую. Разбиение было сделано по датам, чтобы добиться максимальной корректности эксперимента за счет того, что в обучающей выборке не окажется дат из тестовой.

Метрики. Для оценки качества модели использованы следующие метрики:

1. *Коэффициент эффективности модели Нэша – Сатклифа (NSE)* [19]:

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T}(Q_o^t - Q_m^t)^2}{\sum_{t=1}^{T}(Q_0^t - \overline{Q_0})^2}.$$

Это классическая метрика для оценки предиктивной силы гидрологической модели. Она принимает значение 1, если были получены идеальные предсказания, 0 – если предсказания были так же хороши, как средние и отрицательные значения для предсказаний, работающих хуже, чем среднее.

2. *Коэффициент детерминации*

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(D_o^t - D_m^t)^2}{\sum_{t=1}^{T}(D_o^t - \overline{D_o})^2}.$$

Эта метрика аналогична коэффициенту эффективности модели Нэша – Сатклифа, но вместо

абсолютных значений уровней использует относительные изменения уровня воды.

3. *Симметричная средняя абсолютная ошибка в процентах (SMAPE):*

$$\text{SMAPE} = \frac{100\,\%}{T}\sum_{t=1}^{T}\frac{|Q_m^t - Q_o^t|}{(|Q_m^t| + |Q_o^t|)/2}.$$

4. *Средняя абсолютная ошибка (MAE):*

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|D_o^t - D_m^t|,$$

где T – количество наблюдений в выборке; t – индекс наблюдения; Q_o^t – наблюдаемое значение уровня воды в наблюдение t; Q_m^t – предсказанное значение уровня воды в наблюдение t; $\overline{Q_o}$ – среднее наблюдаемое значение; D_o^t – целевое изменение уровня воды для объекта t; D_m^t – предсказанное изменение уровня воды в момент t; $\overline{D_o}$ – среднее целевое изменение уровня воды.

Поскольку модели будут предсказывать изменение уровня воды, то Q_m^t будет получаться из суммы текущего значения уровня воды и предсказанного.

Метрика NSE выбрана потому, что является классической в задаче моделирования будущего уровня вод, R^2 – одна из наиболее распространенных метрик для задачи регрессии, а SMAPE и MAE выбраны как вспомогательные метрики, так как могут с некоторой вероятностью определить аномальное поведение модели.

Моделирование. Поскольку линейная регрессия и нейронная сеть не умеют обрабатывать пропуски в данных, то для экспериментов их необходимо предварительно обработать.

Обработка пропусков сделана в два этапа:

1) в первую очередь там, где возможно, пропуски были заполнены последним известным значением;

2) все оставшиеся пропуски заполнены нулем. Идея заполнения нулем следующая: поскольку и в линейной регрессии, и в узлах нейронной сети происходит линейная комбинация, то зануление фичи уберет ее из линейной комбинации.

Для градиентного бустинга не было необходимости выполнять заполнение пропусков, потому что его реализация в библиотеки CatBoost умеет их обрабатывать. Это свойство является одним из преимуществ данного алгоритма, так как, во-первых, снижает затраты человеческого времени и вычислительных ресурсов на предварительную обработку данных, а во-вторых, поскольку пропуски обрабатываются отдельно, то наличие пропуска в определенном месте само по себе может являться информацией, способной повысить качество предсказания. В экспериментах использовались настройки бустинга, при которых пропущенные значения обрабатывались как самые маленькие во всей выборке. Это метод обработки пропущенных значений по умолчанию.

После обработки пропусков каждая модель была обучена на одном и том же наборе обучающих данных и затем протестирована на одной и той же тестовой выборке. В результате получены результаты, представленные в табл. 1 и 2.

Таблица 1

Качество работы моделей на обучающей выборке

Название модели	NSE	R^2	SMAPE	MAE
Линейная регрессия	0,917	0,131	19,92	30,54
Нейронная сеть	0,917	0,278	18,74	28,08
Градиентный бустинг	0,937	0,359	16,57	26,21

Table 1

Model performance on train set

Model name	NSE	R^2	SMAPE	MAE
Linear regression	0.917	0.131	19.92	30.54
Neural network	0.917	0.278	18.74	28.08
Gradient boosting	0.937	0.359	16.57	26.21

Таблица 2

Качество работы моделей на тестовой выборке

Название модели	NSE	R^2	SMAPE	MAE
Линейная регрессия	0,916	0,084	19,37	32,29
Нейронная сеть	0,913	0,175	18,13	29,95
Градиентный бустинг	0,929	0,224	16,28	28,19

Table 2

Model performance on test set

Model name	NSE	R^2	SMAPE	MAE
Linear regression	0.916	0.084	19.37	32.29
Neural network	0.913	0.175	18.13	29.95
Gradient boosting	0.929	0.224	16.28	28.19

По результатам моделей на тестовой выборке можно безоговорочно утверждать, что наилучшей моделью машинного обучения для предсказания паводков на р. Амур является градиентный бустинг в реализации CatBoost. Его значения NSE на 1 % превысили аналогичные показатели нейронной сети и линейной регрессии, а по показателям R^2 прирост качества составил в 1,28 и 2,6 раз соответственно. Второе место с большим отрывом от третьего получает нейронная сеть. Отчасти превосходство градиентного бустинга в данной задаче можно объяснить обилием данных и наивысшей сложностью модели среди представленных – при таком количестве данных она способно найти в них больше паттернов, чем остальные. Но необходимо заметить, что линейная регрессия переобучилась меньше остальных моделей, поэтому она, скорее всего, могла бы стать лучшей на меньшем количестве наблюдений.

Заключение

Стояла задача определить наилучшую модель машинного обучения для предсказания паводков на р. Амур. Выбор происходил среди трех наиболее распространенных в индустрии моделей: линейная регрессия, нейронная сеть и градиентный бустинг. Использовались наблюдения Росгидромета с 1984 по 2018 г. Проведена тщательная подготовка данных. В результате экспериментов наилучшие результаты продемонстрировал градиентный бустинг. В дальнейшем полученные результаты можно использовать для построения модели наибольшей точности для р. Амур, а именно создать разнообразные и сложные объясняющие признаки и подобрать оптимальные параметры для применения градиентного бустинга к этой задаче. Предположительно результаты исследования переносимы на другие реки, по которым количество наблюдений сравнимо с Амуром.

Список литературы / References

1. Yoganath A, Junichi Y. *Global trends in water related disasters: an insight for policymakers*. Tsukuba: International Centre for Water Hazard and Risk Management (UNESCO); 2009.

2. Arduino G, Reggiani P, Todini E. Recent advances in flood forecasting and flood risk assessment. *Hydrology and Earth Sciences*. 2005;9(4):280–284.

3. Moore R, Bell V, Jones D. Forecasting for flood warning. *Comptes Rendus Geosciences.* 2005;337(1–2): 203–217.

4. Tullos D. Assessing environmental impact assessments: a review and analysis of documenting environmental impacts of large dams. *Journal of Environmental Management*. 2008;90:208–223.

5. DiFrancesco K, Tullos D. Flexibility in water resources management: review of concepts and development of assessment measures for flood management systems. *Journal of the American Water Resources Association*. 2014;50(6):1527–1539.

6. Makhinov AN, Kim VI, Voronov BA. Flooding in the Amur basin in 2013: causes and consequences. *Vestnik of the Far East Branch of the Russian Academy of Sciences*. 2014;(2(174)):5–14. (In Russ.)

Махинов А.Н., Ким В.И., Воронов Б.А. Наводнение в бассейне Амура 2013 года: причины и последствия // Вестник ДВО РАН. 2014. № 2 (174). С. 5–14.

7. Ramírez J. Prediction and modeling of flood hydrology and hydraulics. In: Wohl EE. (ed.) *Inland Flood Hazards: Human, Riparian and Aquatic Communities*. Cambridge: Cambridge University Press; 2010. p. 498.

8. Sahraei S, Asadzadeh M, Unduche F. Signature-based multi-modelling and multi-objective calibration of hydrologic models: application in flood forecasting for Canadian Prairies. *Journal of Hydrology.* 2020;588:125095. https://doi.org/10.1016/j.jhydrol.2020.125095

9. Aqil M, Kita I, Yano A. Analysis and prediction of flow from local source in a river basin using a Neuro-fuzzy modeling tool. *Journal of Environmental Management*. 2007;85(1):215–223.

10. Chang FJ, Hsu K, Chang LC, Yu Y. *Flood forecasting using machine learning methods.* MDPI AG; 2019.

11. Dipanjan S, Raghav B, Tushar S. The Python machine learning ecosystem. In: *Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems.* New York: Apress; 2002. p. 67–118.

12. Carvalho D, Pereira E, Cardoso J. Machine learning interpretability: a survey on methods and metrics. *Electronics.* 2019;8(8):832. https://doi.org/10.3390/electronics8080832

13. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*. Montréal; 2018. p. 6638–6648. https://doi.org/10.48550/arXiv.1706.09516

14. Novorotsky PV. The Amur's flow rate fluctuations for the last 110 years. *Geography and Natural Resources.* 2007;(4):86–90. (In Russ.)

Новороцкий П.В. Колебания стока Амура за последние 110 лет // География и природные ресурсы. 2007. № 4. С. 86–90.

15. Makhinov AN. Amur terrigene and chemical discharge formation. *Proceedings of the International Kyoto Symposium*. Kyoto; 2005. p. 61–65.

16. Demidenko E. *Linear and non linear regression*. *Finance and statistics*. Moscow; 1981. p. 302. (In Russ.)

Демиденко Е. Линейная и нелинейная регрессия. Финансы и статистика. М., 1981. С. 302.

17. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature.* 1986;323:533–536.

18. Hinton GE, Nair V. Rectified linear units improve restricted Boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning*. Haifa; 2010. p. 807–814.

19. Nash J, Sutcliffe JV. River flow forecasting through conceptual models. Part I. A discussion of principles. *Journal of Hydrology*. 1970;10(3):282–290.

Сведения об авторах

Александров Никита Эдуардович, аспирант, департамент инновационного менеджмента в отраслях промышленности, Инженерная академия, Российский университет дружбы народов, Российская Федерация, 117198, Москва, ул. Миклухо-Маклая, д. 6; ORCID: 0000-0001-8183-0257; 1042210208@rudn.ru

Ермаков Дмитрий Николаевич, доктор политических наук, доктор экономических наук, кандидат исторических наук, доцент департамента инновационного менеджмента в отраслях промышленности, Инженерная академия, Российский университет дружбы народов, Российская Федерация, 117198, Москва, ул. Миклухо-Маклая, д. 6; ORCID: 0000-0002-0811-0058, eLIBRARY SPIN-код: 6835-3155; ermakov-dn@rudn.ru

Бром Алла Ефимовна, доктор технических наук, профессор кафедры промышленной логистики, факультет инженерного бизнеса и менеджмента, Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет), Российская Федерация, 105005, Москва, ул. 2-я Бауманская, д. 5, стр. 1; ORCID: 0000-0003-3633-1197, eLIBRARY SPIN-код: 3110-1259; allabrom@bmstu.ru

Омельченко Ирина Николаевна, доктор технических наук, доктор экономических наук, декан факультета инженерного бизнеса и менеджмента, Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет), Российская Федерация, 105005, Москва, ул. 2-я Бауманская, д. 5, стр. 1; ORCID: 0000-0003-4707-1079, eLIBRARY SPIN-код: 7548-0546; logistic@ibm.bmsru.ru

Шкодинский Сергей Всеволодович, доктор экономических наук, профессор, заведующий кафедрой экономического и финансового образования, Московский государственный областной университет, Российская Федерация, 141014, Мытищи, ул. Веры Волошиной, д. 24; профессор департамента инновационного менеджмента в отраслях промышленности, Инженерная академия, Российский университет дружбы народов, Российская Федерация, 117198, Москва, ул. Миклухо-Маклая, д. 6; ORCID: 0000-0002-5853-3585, eLIBRARY SPIN-код: 5372-2519; sh-serg@bk.ru

About the authors

Nikita E. Aleksandrov, Ph.D student, Department of Innovation Management in Industries, Academy of Engineering, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation; ORCID: 0000-0001-8183-0257; 1042210208@rudn.ru

Dmitry N. Ermakov, Dr. of Political Sciences, Dr. of Economics, Ph.D of Historical Sciences, Professor, Department of Innovation Management in Industries, Academy of Engineering, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation; ORCID: 0000-0002-0811-0058, eLIBRARY SPIN-code: 6835-3155; ermakov-dn@rudn.ru

Alla E. Brom, Dr. of Economics, Professor of the Department of Industrial Logistics, Faculty of Engineering Business and Management, Bauman Moscow State Technical University, 5 2-ya Baumanskaya St, bldg 1, Moscow, 105005, Russian Federation; ORCID: 0000-0003-3633-1197, eLIBRARY SPIN-code: 3110-1259; allabrom@bmstu.ru

Irina N. Omelchenko, Dr. of Technical Sciences, Dr. of Economics, Dean of the Faculty of Engineering Business and Management, Bauman Moscow State Technical University, 5 2-ya Baumanskaya St, bldg 1, Moscow, 105005, Russian Federation; ORCID: 0000-0003-4707-1079, eLIBRARY SPIN-code: 7548-0546; logistic@ibm.bmsru.ru

Sergey V. Shkodinsky, Doctor of Economics, Professor, Head of the Department of Economic and Financial Education, Moscow State Regional University, 24 Very Voloshinoy St, Mytishi, 141014, Russian Federation; Professor of the Department of Innovation Management in Industries, Academy of Engineering, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation; ORCID: 0000-0002-5853-3585, eLIBRARY SPIN-code: 5372-2519; sh-serg@bk.ru