Математические структуры и моделирование
2023. № 4 (68). С. 59–70

УДК 004.896
DOI 10.24147/2222-8772.2023.4.59-70

ОБЗОР И АПРОБИРОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КРАТКОСРОЧНОГО ПРОГНОЗИРОВАНИЯ

Д.И. Васина
магистрант, e-mail: bacina94@mail.ru

Омский государственный технический университет, Омск, Россия

Аннотация. Определяется растущее значение солнечной энергии в устойчивом производстве энергии. Традиционные модели прогнозирования с трудом справлялись со сложными нелинейными закономерностями, но появление методов машинного обучения значительно повысило точность и надёжность в этой области. Рассматриваются различные методики машинного обучения, используемые для прогнозирования солнечной энергии: линейная регрессия, CatBoost, XGBoost, искусственные нейронные сети. В работе объясняются математические принципы, лежащие в основе представленных методов, такие как функции потерь, ансамблевые модели, этапы градиентного спуска, термины регуляризации и конкретные функциональные формы этих методологий машинного обучения. Подчеркивается роль методов машинного обучения в значительном повышении точности и надежности прогнозов солнечной энергетики.

Ключевые слова: прогнозирование выработки, машинное обучение, деревья решений, линейная регрессия, искусственные нейронные сети.

1. Введение

В последние годы использование солнечной энергии стало ключевым решением в стремлении к устойчивому производству энергии. Поскольку применение солнечных технологий продолжает развиваться, надёжность и эффективность производства становятся более значимыми [1, 2]. Одной из ключевых задач, с которыми сталкиваются в этой области, является точное прогнозирование и оптимизация производства, поскольку оно зависит от динамичных и непредсказуемых факторов окружающей среды [3, 4].

Прогнозирование солнечной энергии основывалось на традиционных моделях, которые с трудом приспосабливаются к сложным, нелинейным закономерностям. Однако появление методов машинного обучения вызвало волну преобразований в этой области, предложив сложные инструменты для повышения точности и надёжности прогнозов солнечной энергии [5, 6].

В статье рассматривается применение различных методологий машинного обучения – линейной регрессии, CatBoost, XGBoost и искусственных нейронных сетей (ИНС), раскрывается их роль в повышении точности и надёжности прогнозов солнечной энергии.

1.1. Линейная регрессия

Линейная регрессия является фундаментальным статистическим методом, используемым в прогнозировании для установления взаимосвязи между двумя переменными.

В работе В.П. Корнеева, К.С. Хрисанфовой «Применение полиномиальных признаков в задачах линейной регрессии» линейная регрессия определяется как алгоритм машинного обучения для прогнозирования на основе линейной функции зависимости. Авторы проводят исследование сравнения эффективности линейной и полиномиальной регрессионных моделей [7].

В статье Т.А. Бойко «Разработка алгоритма построения модели множественной линейной регрессии» описывается построение модели множественной линейной регрессии, в том числе выявляется необходимость выполнения отбора независимых переменных, математическая часть построения модели и проверка точности прогнозирования. При построении модели множественной линейной регрессии автор акцентирует внимание на том, что важно учитывать p-value, R-квадрат и t-value для проверки качества модели [8].

В области прогнозирования линейная регрессия устанавливает зависимости между зависимой переменной (переменной, подлежащей прогнозированию) и одной или несколькими независимыми переменными (предикторами).

Целью модели является выделение наиболее подходящей линии, которая минимизирует разницу между прогнозируемыми значениями и фактическими. Минимизация выполняется с помощью метода наименьших квадратов, который вычисляет линию, минимизирующую сумму квадратов разностей между наблюдаемыми и прогнозируемыми значениями. Линейная регрессия может быть дополнительно расширена до множественной, где имеется более одной независимой переменной. Уравнение для множественной линейной регрессии имеет вид:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n, \tag{1}$$

где y – зависимая переменная; $x_1, x_2, \ldots, x_n$ – независимые переменные; b_0 – перехват; $b_1, b_2, \ldots, b_n$ – коэффициенты, представляющие влияние каждой независимой переменной на зависимую переменную. Как только модель линейной регрессии обучена на ретроспективных данных, ее можно использовать для прогнозирования будущих значений зависимой переменной на основе значений независимых переменных.

1.2. Метод ансамбля градиентного бустинга деревьев решений CatBoost

Метод CatBoost как алгоритм повышения градиента работает путём минимизации заданной функции потерь при помощи добавления деревьев решений в ансамбль. Используется аддитивная модель, в которой каждое новое дерево подгоняется к остаточным ошибкам существующего ансамбля. Окончательное предсказание – это сумма предсказаний, сделанных всеми деревьями в ансамбле.

Метод минимизирует заданную дифференцируемую функцию потерь (например, среднеквадратичную ошибку для регрессии или логарифмическую потерю для клас-

сификации) с использованием градиентного спуска. CatBoost стремится найти оптимальные параметры (древовидные структуры) путем итеративного перемещения в направлении, противоположном градиенту функции потерь.

А.Д. Моргоева, Р.В. Клюев в своей научной работе «Прогнозирование потребления электрической энергии промышленным предприятием с помощью методов машинного обучения» утверждают, что развитие и применение методов интеллектуального анализа данных способствует уменьшению и рационализации использования ресурсов. В статье описано исследование, которое проводилось на промышленном предприятии с энергоёмким и устаревшим оборудованием. По его результатам авторы делают вывод, что модель машинного обучения, основанная на алгоритме градиентного бустинга библиотеки CatBoost, даёт достоверный прогноз потребления электроэнергии, что является актуальным с точки зрения экономических выгод [9].

Математическая основа и применение CatBoost позволяют ему эффективно обрабатывать категориальные данные при построении точных прогностических моделей в различных областях.

Фундаментальные компоненты и математические принципы CatBoost включают комбинацию повышения градиента и обработки категориальных признаков:

1. $L(y, F(x))$ – функция потерь, используемая в CatBoost. Эта функция представляет ошибку между истинной целью y и прогнозом модели $F(x)$.
2. $F(x) = \Sigma_{m=1}^{M} f_m(x)$ – предсказание с помощью ансамблевой модели, где M – это количество деревьев b и $f_m(x)$ представляет выходные данные каждого отдельного дерева.
3. Шаг градиентного спуска – на каждой итерации новое дерево подгоняется к отрицательному градиенту функции потерь для каждой выборки i в отношении текущего модельного прогноза:

$$r_{im} = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}. \tag{2}$$

1.3. Метод ансамбля градиентного бустинга деревьев решений XGBoost

XGBoost (Extreme Gradient Boosting) является улучшенным алгоритмом машинного обучения, который относится к семейству gradient boosting. XGBoost использует специально разработанную структуру и алгоритм повышения производительности, что определяет его эффективным с точки зрения вычислений. Метод имеет встроенный механизм для обработки пропущенных значений во время обучения. Алгоритм включает в себя методы регуляризации L1 и L2, контролирующие сложность модели для предотвращения переобучения. XGBoost позволяет определять свои собственные целевые функции и критерии оценки.

T. Chen в работе «XGBoost: A Scalable Tree Boosting System» проводит практическое исследование и приходит к заключению о том, что XGBoost – один из методов, который используется во многих сферах деятельности. Бустинг даёт самые современные результаты по многим стандартным критериям классификации. Инновации метода включают в себя: новый алгоритм древовидного обучения, предназначенный для обработки разреженных данных; теоретически обоснованную процедуру

создания эскиза взвешенных квантилей, которая позволяет обрабатывать веса экземпляров в приближенном древовидном обучении [10].

Надёжность XGBoost можно обосновать его эффективностью и способностью обрабатывать большие наборы данных, что применимо в различных сферах машинного обучения, включая задачи регрессии, классификации и ранжирования:

$$f(x) = \sum_{i=1}^{n} L(y_i, \widehat{y_i}) + \sum_{K=1}^{K} \Omega(f_k), \tag{3}$$

где $f(x)$ – целевая функция, которая сводится к минимуму; $L(y_i, \widehat{y_i})$ – функция потерь, измеряющая погрешность между фактическими y_i и прогнозируемыми значениями $\widehat{y_i}$; $\Omega(f_k)$ – член регуляризации, который определяет сложность отдельных деревьев, добавляя штраф за регуляризацию к весам дерева.

Модель прогнозирования методом XGBoost имеет следующий вид:

$$f(x) = \Sigma_{k=1}^{k} f_k(x_i). \tag{4}$$

На каждой итерации новые деревья подгоняются к отрицательному градиенту функции потерь:

$$r_{ik} = -\frac{\partial L(y_i, \widehat{y_i})}{\partial \widehat{y_i}}, \tag{5}$$

где r_{ik} – отрицательный градиент функции потерь по отношению к текущему прогнозу.

1.4. Метод прогнозирования на основе искусственной нейронной сети

Метод прогнозирования с использованием искусственной нейронной сети, особенно в последовательной модели, предполагает использование архитектуры глубокого обучения, предназначенной для обработки последовательных данных, таких как временные ряды, текстовые данные или любые данные с определённым порядком или последовательностью.

В работе авторов А.Н. Попова, А.Д. Венгерского «Разработка модели машинного обучения для прогнозирования генерации электроэнергии солнечными панелями на основе алгоритма градиентного бустинга» акцентируется внимание на том, что для предсказания солнечной радиации можно построить модель машинного обучения, но при этом подавляющее число используемых алгоритмов – нейронные сети [11].

В исследовании приводится архитектура ИНС, которая состоит из трёх слоев. Первый слой входной, он имеет 32 нейрона, на вход принимает 6 входных векторов; второй скрытый слой имеет 64 нейрона с функцией активации ReLU; третий – один выходной нейрон с линейной функцией активации. Оптимизатор для модели Adam, итераций обучения – 100, функцией потерь для задачи регрессии является среднеквадратичная ошибка (MSE) [12].

Процесс обучения включает в себя подачу последовательных данных в модель и обновление весов модели, чтобы минимизировать определённую функцию потерь. Нейронная сеть прямого действия (представленная в исследовании) выражается следующим образом:

1. Для одного нейрона (персептрона):
   Ввод: $x_1, x_2, \ldots, x_n$
   Вес: $\omega_1, \omega_2, \ldots, \omega_n$
   Смещение: b
   Функция активации $\sigma(z) = \begin{cases} 0, \text{если } z < 0 \\ z, \text{если } z \geqslant 0 \end{cases}$, таким образом $\sigma(z) = \max(0, z)$, т. е. если значение z больше или равно нулю, функция вернёт значение z, иначе она вернёт ноль. Это делает функцию ReLU простой и эффективной, она широко используется в нейронных сетях, поскольку обеспечивает быструю обучаемость и хорошую способность преодолевать проблему затухания градиента.

$$z = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n + b; \tag{6}$$

2. Распространение через скрытый слой:
   Для скрытого слоя с несколькими нейронами входные данные взвешиваются, суммируются и передаются через функцию активации:

$$z_j = \Sigma_{i=1}^{n} \omega_{ij} x_i + b_j \tag{7}$$

   для каждого нейрона в слое, $y_j = \sigma(z_j)$ для каждого нейрона, где z_j представляет собой взвешенную сумму плюс смещение для j-го нейрона, и $\sigma(z_j)$ является выходным сигналом после прохождения функции активации для этого нейрона.

3. Распространение по последовательной модели:
   В последовательной нейронной сети информация передаётся с помощью временных шагов. На временном шаге t входные данные x_t объединяются с предыдущим скрытым состоянием h_{t-1} и проходят через нейрон(ы).

$$h_{t-1} = \sigma(w_{hx} x_t + w_{hh} h_{t-1} + b_h), \tag{8}$$

   где w_{hx} и w_{hh} – веса из входных данных и предыдущего скрытого состояния; b_h – это смещение; h_t – вывод или скрытое состояние в момент времени t, которое передаётся на следующий временной шаг.

2. Проведение исследования

Набор данных представлен в виде таблицы и содержит следующие значения с наименованиями:

Date, time – информация о временном промежутке, начиная с 15.05.2020 и заканчивая 18.06.2020, в виде 2020.05.15 00:00 (дд.мм.гггг чч:мм). Интервал времени – каждые 15 минут;

P_{DC} – значение выработки постоянного тока, выраженное в кВт*ч ;

P_{AC} – значение выработки переменного тока, выраженное в кВт*ч;

T_A – температурное значение на солнечной станции, измеряемое в C^o;

T_M – значение показания модуля температуры, измеряемое в C^o;

Ir – электромагнитное и корпускулярное излучение Солнца, его количество за 15-минутный интервал времени, выраженный в кВт·ч/м^2;

H – время, час;

D – день;

D_W – день недели;

M_{15} – 15-минутный отрезок времени в месяце.

Набор данных о выработке электроэнергии собирается на уровне инвертора – к каждому инвертору подключено несколько линий солнечных панелей. Данные датчиков собираются на уровне предприятия – единый массив датчиков оптимально размещён на предприятии (табл. 1) [13].

Таблица 1. Данные о выработке электроэнергии

Date, time	P_{DC}	P_{AC}	T_A	T_M	Ir	H	D	D_W	M_{15}
15.05.2020 00:00	***0***	***0***	***25,2***	***22,9***	0	0	1	4	1
15.05.2020 00:15	0	0	25,1	22,8	0	0	1	4	2
15.05.2020 00:30	0	0	24,9	22,6	0	0	1	4	3
…	…	…	…	…	…	…	…	…	…
29.05.2020 09:00	6317.0	618,7	24,4	40,1	0,5	9	15	4	37
29.05.2020 09:15	8053.6	788,2	25,4	47,3	0,7	9	15	4	38
…	…	…	…	…	…	…	…	…	…
16.06.2020 23:15	0	0	22,9	21,3	0	23	33	1	94
16.06.2020 23:30	0	0	22,9	21,2	0	23	33	1	95
16.06.2020 23:45	0	0	22,892	21,2	0	23	33	1	96

На рис. 1 представлена корреляционная матрица зависимостей параметров набора данных, представленных в таблице.

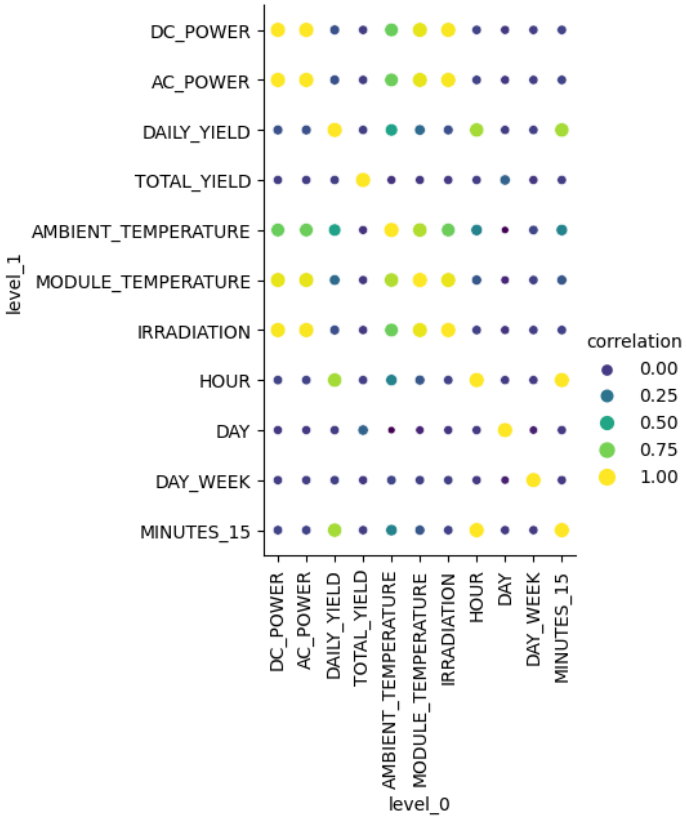

Рис. 1. Корреляционная матрица параметров набора данных

Корреляционная матрица представлена в виде графика, в котором коэффициент корреляции между всеми парами переменных отображается 5 уровнями коэффициента корреляции. Показано, насколько сильно и в каком направлении (положительном или отрицательном) связаны между собой переменные. Значения коэффициента корреляции варьируются от 1 до 0 (взятые по модулю), где 1 означает идеальную положительную, а 0 – отсутствие корреляции. На графике корреляционной матрицы видно, какие параметры имеют сильную корреляцию между собой (близкие к 1), какие – слабую (близкие к 0). Например, температура модуля и температура окружающей среды, уровень солнечного излучения имеют положительную корреляцию.

Проведено краткосрочное прогнозирование выработки мощности переменного (рис. 2) и постоянного тока (рис. 3) на сутки вперёд с использованием методов CatBoost, XGBoost, ИНС и линейной регрессии.

На рис. 2(а–г) отображены результаты прогнозирования мощности переменного тока солнечными панелями. Прогнозы выработки электроэнергии переменным током солнечными панелями с использованием различных методов были оценены на основе предоставленных параметров. Результаты показали различную прогностическую эффективность различных методов, CatBoost (рис. 2а) и XGBoost (рис. 2б) доказали относительно высокие прогностические способности, продемонстрировав многообещающую точность прогнозирования выходной мощности переменного тока. С другой стороны, метод линейной регрессии (рис. 2г) продемонстрировал наи-

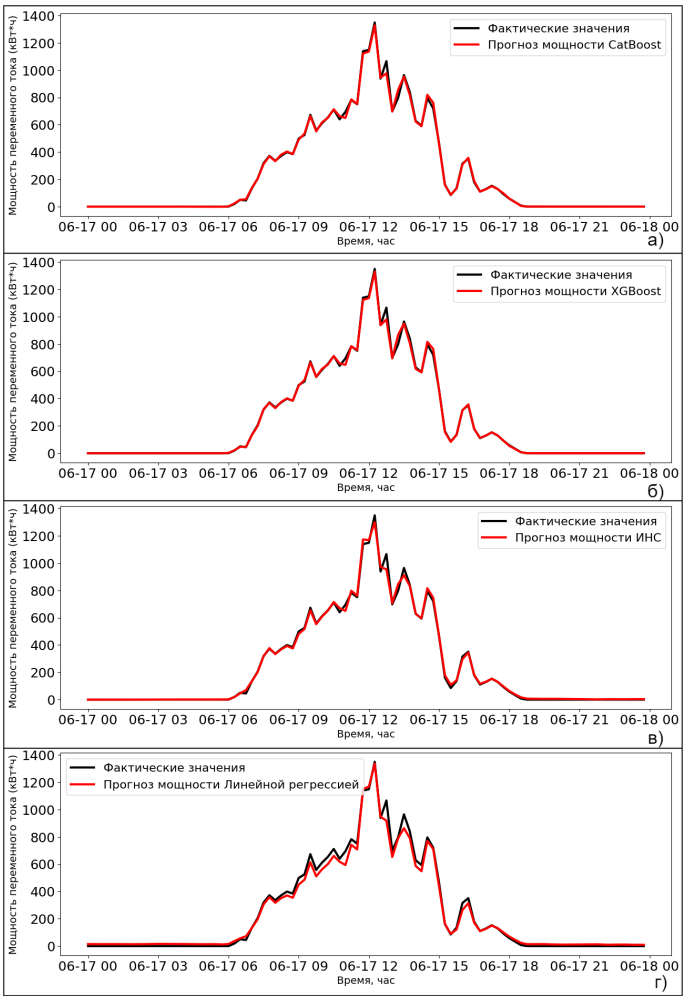

Рис. 2. Прогнозирование выработки мощности переменного тока методами: CatBoost (а); XGBoost (б); ИНС (в); линейной регрессии (г)

меньшую прогностическую способность среди представленных методов. Несмотря на свою простоту, он показал меньшую точность при отображении сложных взаимосвязей между входными параметрами и выходной мощностью переменного тока

На рис. 3(а–г) изображены графики прогноза выработки постоянного тока солнечными панелями. Так же, как и при прогнозировании выработки переменного тока, хуже себя проявил метод линейной регрессии (рис. 3г), а метод искусственных нейронных сетей (рис. 3в) занизил показатели во все пиковые временные периоды.

Проанализировав графики, можно утверждать, что методы CatBoost (рис. 3а) и XGBoost (рис. 3б) качественно прогнозируют оба показателя выработки мощностей и выдают приблизительно одинаково точные результаты.

Для оценивания качества эффективности представленных методов краткосрочного прогнозирования применялась метрика MAPE, которая характеризует отклонение в процентах от фактических значений (табл. 2).

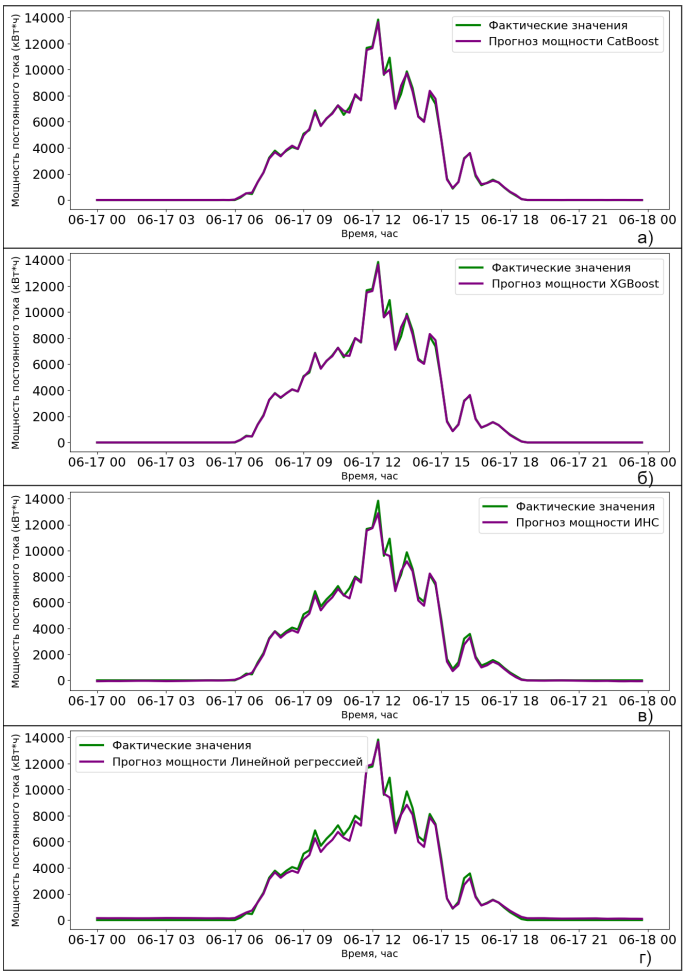

Рис. 3. Прогнозирование выработки мощности постоянного тока методами: CatBoost (а); XGBoost (б); ИНС (в); линейной регрессии (г)

В табл. 2 отображены средние значения абсолютной процентной ошибки (MAPE) для прогнозирования мощности переменного тока (P_{AC}) и мощности постоянного тока (P_{DC}).

При прогнозировании мощности P_{AC} XGBoost демонстрирует низкие значения MAPE = 2,44 %, что указывает на высокий уровень точности прогнозирования мощности переменного тока. CatBoost показывает чуть более высокий показатель MAPE = 3,33 % (соответствует требуемой точности). Метод на основе ИНС демонстрирует более высокий показатель MAPE = 7,09 %, линейная регрессия имеет значение ошибки – 14,01 %, что демонстрируют графики прогнозов (см. рис. 2 и 3).

При прогнозировании мощности P_{DC} аналогичным образом, XGBoost обеспечивает низкий показатель MAPE = 2,22 % для прогнозирования мощности постоянного тока, что подчёркивает его сильную прогностическую способность. CatBoost имеет MAPE = 4,04 %, что также соответствует требуемой точности. Для метода на основе ИНС MAPE = 8,09 %, также наблюдалось при прогнозировании мощности переменного тока, линейная регрессия показывает самый высокий показатель MAPE = 13,46 %, что указывает на меньшую точность прогнозирования мощности

Таблица 2. Ошибка прогнозирования мощности переменного и постоянного тока

Значение метрики MAPE, %							
При прогнозировании P_{AC}				При прогнозировании P_{DC}			
XGBoost	*CatBoost*	ИНС	Линейная регрессия	*XGBoost*	*CatBoost*	ИНС	Линейная регрессия
2,44%	3,33%	7,09%	14,01%	2,22%	4,04%	8,09%	13,46%

постоянного тока, по сравнению с другими методами.

3. Выводы

Таким образом, рассмотрено растущее значение солнечной энергии в устойчивом производстве энергии и необходимость точного прогнозирования её выработки. В исследовании описаны различные методологии машинного обучения, включая линейную регрессию, CatBoost, XGBoost и искусственные нейронные сети, указывается их роль в повышении точности и надёжности прогнозов солнечной энергии.

По результатам установлено, что метод линейной регрессии неэффективен при прогнозировании солнечной энергии из-за сложностей установления взаимосвязей между входными данными и выходными.

Методы CatBoost и XGBoost, использующие деревья решений, продемонстрировали перспективные возможности прогнозирования. Они эффективно использовали сложность данных, что привело к высокой точности прогнозирования выходной мощности переменного тока, превосходящей традиционные методы. Искусственные нейронные сети, разработанные для обработки последовательных данных, продемонстрировали высокую точность в прогнозировании солнечной энергии. Однако расхождения в прогнозировании периодов пикового времени указывают на необходимость дальнейшего уточнения гиперпараметров модели.

Литература

1. Глебов В.В., Братышев С.Н., Верига А.В. Краткосрочное прогнозирование выработки электроэнергии сетевой солнечной электростанции методом рекуррентных нейронных сетей // Цифровые технологии и защита информации в современном обществе : сборник докладов Международной научно-практической конференции (Астрахань, 29–30 ноября 2021 г.). Астрахань : Астраханский государственный университет, 2021. С. 4–9.
2. Горшенин А.Ю. Формирование выборки исходных данных для машинного обучения модели краткосрочного прогнозирования электропотребления // Автоматизация в промышленности. 2023. № 10. С. 37–41. DOI: 10.25728/avtprom.2023.10.08.
3. Эвок Д.А., Дьячков Я.А., Микулицкий М.В. Прогнозирование температур поверхности солнечных панелей по стандартам NOCT и NMOT // Энерго- и ресурсосбережение в теплоэнергетике и социальной сфере : материалы Международной научно-технической конференции студентов, аспирантов, учёных. Челябинск : Изд-во ЮУрГУ, 2023. С. 76–79.

4. Горшенин А.Ю., Денисова Л.А. Прогнозирование выработки электроэнергии ветроэлектростанцией с применением рекуррентной нейронной сети // Известия Тульского государственного университета. Технические науки. 2023. № 4. С. 39–45. DOI: 10.24412/2071-6168-2023-4-39-45.
5. Тюньков Д.А., Сапилова А.А., Грицай А.С., Алексеенко Д.А., Хамитов Р.Н. Методы краткосрочного прогнозирования выработки электрической энергии солнечными электростанциями и их классификация // Электротехнические системы и комплексы. 2020. № 3 (48). С. 4–10. DOI: 10.18503/2311-8318-2020-3(48)-4-10.
6. Горшенин А.Ю., Васина Д.И. Сравнение используемых методов при прогнозировании выработки электроэнергии ветроэлектростанциями // Математические структуры и моделирование. 2023. № 3 (67). С. 36–42. DOI: 10.24147/2222-8772.2023.3.36-42.
7. Корнеев В.П., Хрисанфова К.С. Применение полиномиальных признаков в задачах линейной регрессии // Тенденции развития науки и образования. 2023. № 94–5. С. 154–156. DOI: 10.18411/trnio-02-2023-275.
8. Бойко Т.А. Разработка алгоритма построения модели множественной линейной регрессии // Экономика и управление: проблемы, решения. 2018. Т. 8, № 12. С. 113–119.
9. Моргоева А.Д., Моргоев И.Д., Клюев Р.В., Гаврина О.А. Прогнозирование потребления электрической энергии промышленным предприятием с помощью методов машинного обучения // Известия Томского политехнического университета. Инжиниринг георесурсов. 2022. Т. 333, № 7. С. 115–125. DOI: 10.18799/24131830/2022/7/3527.
10. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD'16. 2016. P. 785–794. DOI: 10.1145/2939672.2939785.
11. Попов А.Н., Венгерский А.Д. Разработка модели машинного обучения для прогнозирования генерации электроэнергии солнечными панелями на основе алгоритма градиентного бустинга // Интеллектуальная энергетика : сборник научных статей кафедры «Электроснабжение промышленных предприятий» АлтГТУ им. И.И. Ползунова / сост.: С.О. Хомутов, В.И. Сташко. Барнаул : Межрегиональный центр электронных образовательных ресурсов, 2021. С. 101–103.
12. Саетова Л.Г., Горохов М.М. Нейронная сеть и регрессия: описание линейной регрессии в нейронных сетях // Информационные технологии в науке, промышленности и образовании : сборник трудов научно-технической конференции в рамках Всероссийского молодёжного научного форума «Общение студентов и аспирантов в научной и профессиональной сферах» (Ижевск, 26 мая 2021 г.). Ижевск : Ижевский государственный технический университет им. М.Т. Калашникова, 2021. С. 15–21.
13. Васина Д.И. Описание программы для проведения анализа выработки мощности постоянного и переменного токов солнечной электростанции // Актуальные вопросы энергетики : материалы Всероссийской научно-практической конференции с международным участием (Омск, 25–26 мая 2023 г.) / редколлегия: П.А. Батраков (отв. ред.) [и др.]. Омск : Омский государственный технический университет, 2023. С. 133–137.

REVIEW AND APPROBATION OF MACHINE LEARNING METHODS FOR SHORT-TERM FORECASTING

D.I. Vasina
Master's Degree Student, e-mail: bacina94@mail.ru

Omsk State Technical University, Omsk, Russia

Abstract. The article defines the growing importance of solar energy in sustainable energy production. Traditional forecasting models had difficulty coping with complex nonlinear patterns, but the advent of machine learning methods has significantly increased accuracy and reliability in this area. The article discusses various machine learning techniques used to predict solar energy: linear regression, CatBoost. XGBoost, artificial neural networks. The paper explains the mathematical principles underlying the presented methods, such as loss functions, ensemble models, gradient descent stages, regularization terms and specific functional forms of these machine learning methodologies. The role of machine learning methods in significantly improving the accuracy and reliability of solar energy forecasts is emphasized.

Keywords: output forecasting, machine learning, decision trees, linear regression, artificial neural networks.

Дата поступления в редакцию: 14.05.2023