

СИСТЕМНЫЙ АНАЛИЗ ДАННЫХ И МОДЕЛИРОВАНИЕ ПОВЕДЕНИЯ ПОТРЕБЛЕНИЯ ЭЛЕКТРОЭНЕРГИИ НА КАЖДЫЙ ЧАС СУТОК С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

**Дзгоев А. Э.¹, Лагунова А. Д.²,
(РТУ МИРЭА, Москва)**

Карацев С. Т.³
(Финансовый университет при Правительстве РФ
(Владикавказский филиал), Владикавказ)

Конюшок И. А.⁴
(ООО «Рестадвайзер», Санкт-Петербург)

Комаров И. А.⁵
(ООО «Фриссон», Санкт-Петербург)

Хузмиев И. М.⁶
(ПАО «Россети Северный Кавказ», Пятигорск)

Гладышев О. Я.⁷
(Филиал АО "СО ЕЭС" ОДУ Северо-Запада,
Санкт-Петербург)

Потребление электроэнергии является ключевым фактором устойчивого развития в энергетической отрасли, и точное прогнозирование его изменений имеет важное значение для эффективного управления большими электро-энергетическими системами и ресурсами. Целью данного исследования явля-

¹ Алан Эдуардович Дзгоев, к.т.н., доцент, доцент кафедры цифровой трансформации (dzgoviev@mirea.ru).

² Анна Дмитриевна Лагунова, к.э.н., доцент, зав. кафедрой цифровой трансформации (lagunova@mirea.ru).

³ Станислав Таймуразович Карацев, ст. преподаватель базовой кафедры «Корпоративные инфокоммуникационные системы» (stkaratsev@fa.ru).

⁴ Илья Андреевич Конюшок, начальник отдела развития искусственного интеллекта ООО «Рестадвайзер», разработчик ИИ (konushok@mail.ru).

⁵ Иван Александрович Комаров, генеральный директор ООО «Фриссон» (info@frisson.ai).

⁶ Игорь Маратович Хузмиев, к.т.н., заместитель директора Департамента корпоративных и технологических АСУ (huzmiev-im@rossetisk.ru).

⁷ Олег Ярославович Гладышев, главный эксперт, Служба перспективного развития (gladushev-oya@odusz.so-ups.ru).

ется разработка математической (регрессионной) модели для прогнозирования поведения электропотребления на каждый час следующих суток для энергосбытовых компаний современными методами машинного обучения и искусственного интеллекта. Рассматриваются различные методы искусственного интеллекта, применяемые для моделирования и прогнозирования потребления электроэнергии. К этим методам относятся: линейная модель, случайный лес и две реализации градиентного бустинга над решающими деревьями. Научный подход, основанный на технологии искусственного интеллекта Boosting, позволяет максимально снизить ошибку прогнозирования электропотребления в крупных энергетических компаниях. Авторами разработана новая, полезная и качественная регрессионная модель, адекватно описывающая экспериментальные данные по потреблению электроэнергии за каждый час суток. Выполнено тестирование разработанной регрессионной модели на реальных производственных данных энергетической компании. Проведенное исследование и полученные результаты позволяют авторам сделать вывод о том, что разработанная математическая модель методом машинного обучения LightGBM может быть использована энергосбытовыми компаниями для почасового планирования электропотребления при подаче заявок на оптовый рынок электроэнергии и мощности (ОРЭМ) на несколько дней вперед. Исследование было выполнено на языке программирования Python.

Ключевые слова: моделирование электропотребления, обработка данных, оценка качества моделей, математическая статистика, регрессионная модель, искусственный интеллект.

1. Введение

Актуальность выбранной темы для науки и практики обусловлена растущей потребностью в эффективном управлении энергоресурсами с учётом продолжающейся цифровой трансформации электроэнергетической отрасли Российской Федерации. Финансовые операции энергетических и производственных компаний зависят от их способности удовлетворять свои потребности в электроэнергии [19].

В современном мире, где энергетические ресурсы ограничены, планирование потребления электроэнергии играет важную роль в оптимизации работы энергетических систем и обеспечении устойчивого развития производственных предприятий. Поэтому корректное краткосрочное прогнозирование нагрузки имеет решающее значение для эффективного планирования и управления электроэнергетическими системами и ресурсами [22].

Краткосрочное прогнозирование электропотребления рассчитывается от нескольких часов до нескольких дней. Обычно компании прогнозируют ежедневное, почасовое потребление электроэнергии определенного региона или предприятия. Краткосрочные прогнозные оценки помогают электрогенерирующим компаниям страны в планировании необходимых объёмов генерации электроэнергии и диспетчеризации генерирующих установок [32].

Для решения вышеуказанной задачи краткосрочного прогнозирования электропотребления исследователи из Пакистана Momina Qureshi др. [25], а также из Южной Кореи Jaewon Chung и Beakcheol Jang [14] используют различные методы прогнозирования и анализа временных рядов: регрессионный анализ, искусственные нейронные сети (ИНС), алгоритмы машинного обучения, декомпозицию временных рядов, а также некоторые гибридные подходы, которые объединяют два или более методов для получения лучшей адекватности математических моделей, в том числе алгоритм LSTM. Среди традиционных методов множественная линейная регрессия широко использовалась исследователями для краткосрочного прогнозирования потребления электроэнергии путем выявления влияния экономических, демографических, климатических факторов для различных стран [26, 31].

Исследователи из Бразилии Mesquita Lopes Cabreira и др. в своей статье [24] описывают этапы методологии, теоретическое описание моделей временных рядов и иерархические подходы к прогнозированию энергетических процессов. Они предлагают использовать двухуровневую иерархию в промышленном секторе Бразилии, учитывая потребление электроэнергии по географическим регионам и штатам федерации, включая Федеральный округ. К обработанным данным подбираются классические математические модели временных рядов экспоненциального сглаживания и Бокса – Дженкинса, которые рассчитывают прогнозные значения. Исследователи из Южной Кореи Son N. и Shin Y. в своем исследовании [34] предлагают использовать алгоритм краткосрочного и среднесрочного прогнозирования потребления электроэнергии путем объединения модели

рекуррентной искусственной нейронной сети (управляемые рекуррентные блоки GRU), подходящей для долгосрочного прогнозирования, и аддитивной модели Prophet, подходящей для учёта сезонности и обработки событий.

В статье [21] авторы показывают и описывают результаты способности нескольких методов: ИНС, адаптивной нейронечёткой системы вывода (ANFIS), метода опорных векторов наименьших квадратов (LSSVM) и нечёткой модели временных рядов (FTS) прогнозировать потребление электроэнергии в разных странах с разным уровнем развития. Переменные, рассматриваемые в этом исследовании, включали ежемесячное потребление электроэнергии в семи странах: Норвегия, Швейцария, Малайзия, Египет, Алжир, Болгария и Кения.

Исследователи из Узбекистана Рахмонов И.У. и др. в статье [30] описывают математическую модель прогнозирования электропотребления, используя метод ARIMA, которая успешно справилась с задачей, показав относительную ошибку 6%.

Ученые из Турции в своём научном труде [36] предлагают использовать для решения задачи краткосрочного прогнозирования на производственных предприятиях один из самых мощных методов, который известен своей способностью изучать основные характеристики данных для моделей временных рядов – метод Prophet. В статье доказано, что без использования температуры критерии качества модели, такие как средняя абсолютная ошибка (MAE) и квадратный корень из среднеквадратической ошибки (RMSE), показали лучшие результаты, чем с температурой.

В статье [35] рассматривается гибридная модель для краткосрочного прогнозирования энергетической нагрузки на основе трансферного обучения, используя метод машинного обучения Light Gradient Boosting Machine (LightGBM) для интеллектуальных сетей в интеллектуальных энергетических системах, которая позволяет получить ошибку прогнозирования MAPE менее 3%.

В статье [17] предлагается использовать новый метод для среднесрочного прогнозирования потребления электроэнергии на основе Transformer-LightGBM. Метод предварительно обра-

батывает исходные данные рынка электроэнергетики, включая анализ пропущенных значений и выбросов, затем проводится корреляционный анализ, чтобы извлечь признаки с сильной связью со среднесрочными прогнозами потребления электроэнергии. Впоследствии нейронная сеть трансформатора изучает сложные закономерности в динамических временных рядах. Наконец, LightGBM используется для объединения характеристик электропотребления, на основании которых алгоритм рассчитывает прогнозы.

В настоящее время на большинстве производственных предприятий Российской Федерации применяются методы экспертных оценок, когда краткосрочное прогнозирование осуществляется сотрудником компании на основе использования простейших арифметических операций, что не может обеспечить высокую точность. Однако современные подходы к экономическому и техническому управлению и развитие информационных технологий предъявляют всё более жесткие требования к точности решения задач прогнозирования электропотребления [10]. Эффективным решением данной задачи является создание новой, полезной и качественной математической модели прогнозирования, адекватно описывающей экспериментальные данные о фактическом электропотреблении за каждый час суток.

В статье [8] описаны результаты исследования применения методов регрессионного множественного моделирования электропотребления на основе использования данных с переменной структурой. Разработанная авторами модель позволяет выявлять особенности технологического процесса, оценивать вклад различных факторов на показатели потребления электроэнергии и на этой основе формулировать более точные рекомендации по повышению энергоэффективности промышленного производства.

Статистических методов, использующихся в процессе разработки математической модели, известно большое количество, все они имеют свои достоинства и недостатки. Выбор статистического метода осуществляется под требования конкретного предприятия путём исследования эффективности его приме-

ния, используя данные изучаемого процесса в организации, в том числе в условиях неполноты информации [11]. Таким образом, правильный выбор статистического метода – важный этап в системном анализе данных и разработке математической (регрессионной) модели, адекватно описывающей экспериментальные данные. Непрерывное развитие информационных систем и средств вычислительной техники даёт возможность расширять круг используемых методов прогнозирования и совершенствовать их [3], например, метод скользящей матрицы [1], можно эффективно использовать для краткосрочного и оперативного прогнозирования электропотребления на основе малой выборки данных [6].

Экономически актуальность прогнозирования электропотребления для производственных и энергетических предприятий обосновывается следующим образом:

- деятельность энергосбытовых компаний, являющихся участниками оптового рынка электроэнергии, включает в себя покупку объёмов электроэнергии на этом рынке и реализацию ее розничным потребителям. При этом объёмы электроэнергии, закупаемые на оптовом рынке, должны быть равны составленным днём ранее (или более) прогнозам. В случае отклонения фактически потребленной электроэнергии от заявленной энергосбытовое предприятие вынуждено докупать объёмы электроэнергии или продавать излишне закупленные объёмы электроэнергии на балансирующем рынке по невыгодной для предприятия цене. При больших отклонениях на энергосбытовое предприятие могут быть наложены меры оперативного воздействия [9]. В итоге энергосбытовое предприятие, являющееся гарантирующим поставщиком, несёт финансовые потери и снижение экономических показателей;

- особенность деятельности крупных производственных предприятий горно-металлургического комплекса страны состоит в том, что ежегодно уменьшается количество богатых месторождений полезных ископаемых, следовательно, производственные работы проводятся на больших объёмах руды с увеличенным расходом электрической энергии в несколько раз. Поэтому затраты за электропотребление на производственных

предприятиях составляют до 30% от себестоимости выпускаемой продукции. Если в месяц крупные производственные предприятия оплачивают расход электроэнергии 350–500 млн руб. (в том числе затраты из-за некорректных прогнозных оценок при подаче заявок на ОРЭМ), то актуальность точного прогнозирования и экономии денежных средств достаточно высока.

В связи с вышеописанным энергосбытовым компаниям и крупным производственным предприятиям крайне важно рассчитывать корректные краткосрочные и оперативные прогнозныe оценки и работать на ОРЭМ, не переплачивая колоссальные суммы денежных средств при оплате расхода электроэнергии. Таким образом, возникает задача моделирования и прогнозирования электропотребления с наименьшей абсолютной и относительной ошибкой, результаты решения которой описываются авторами в данной статье.

Ученые проводят исследования, посвященные разработке математических моделей и методов для прогнозирования электропотребления в регионе с помощью различных методов: статистических, машинного обучения и их комбинаций. В качестве данных для решения задачи прогнозирования часто используются временные ряды.

Следует отметить, что при всем теоретическом многообразии и большом практическом опыте решения задач прогнозирования электропотребления универсальных прогнозных методов и моделей не существует, поскольку процессы потребления электроэнергии зависят от множества факторов и специфики производства каждого в отдельности предприятия [7].

2. Системный анализ данных электропотребления

2.1. СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ

В качестве зависимой переменной были выбраны значения регионального потребления электроэнергии (Y) (МВт·ч.) за каждый час суток в период с 01.01.2019 г. до 29.06.2023 г. энергосбытовой компании ПАО «Россети Северный Кавказ» АО «Северный Кавказэнерго» находящейся в г. Владикавказе (Республика Се-

верная Осетия – Алания). Исходная выборка значений зависимой переменной Y состояла из 39 384 строк.

Структура экспериментальных данных электропотребления представлена на рис. 1. Данные загружаются в `Pandas.DataFrame`. `Pandas.DataFrame` – двумерная динамическая структура данных для потенциально гетерогенной табличной информации. Библиотека `Pandas` используется в широком спектре областей, таких как академические исследования, розничная торговля, финансы, экономика, статистика, аналитика, а также в электроэнергетике [29].

Электропотребление	
Дата и время	
2019-01-01 00:00:00	215.191
2019-01-01 01:00:00	202.239
2019-01-01 02:00:00	188.462
2019-01-01 03:00:00	179.276
2019-01-01 04:00:00	173.035
...	...
2023-06-29 19:00:00	176.542
2023-06-29 20:00:00	188.429
2023-06-29 21:00:00	183.351
2023-06-29 22:00:00	170.048
2023-06-29 23:00:00	152.483

39384 rows × 1 columns

Рис. 1. Загруженные данные из Excel

Рассчитаны основные описательные статистики для анализа значений зависимой переменной:

- количество наблюдений (N) = 39384 значения электропотребления за каждый час суток, кВт·ч.;
- выборочное среднее:

$$(1) \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n} = 184,464; \text{ кВт} \cdot \text{ч};$$

— стандартное отклонение:

$$(2) \quad sd_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} = 43,164; \text{ кВт} \cdot \text{ч};$$

— минимальное значение электропотребления за час, кВт·ч.: $Y_{min} = 93,956 \text{ кВт} \cdot \text{ч};$

— максимальное значение электропотребления за час, кВт·ч.: $Y_{max} = 321,055 \text{ кВт} \cdot \text{ч};$

— первый квартиль (25-й перцентиль) = 156,352 кВт·ч.;

— второй квартиль (50-й перцентиль, или медиана) =
= 179,882 кВт·ч.;

— третий квартиль (75-й перцентиль) = 213,912 кВт·ч.

Продолжая проводить разведочный анализ данных, построим гистограмму (рис. 2).

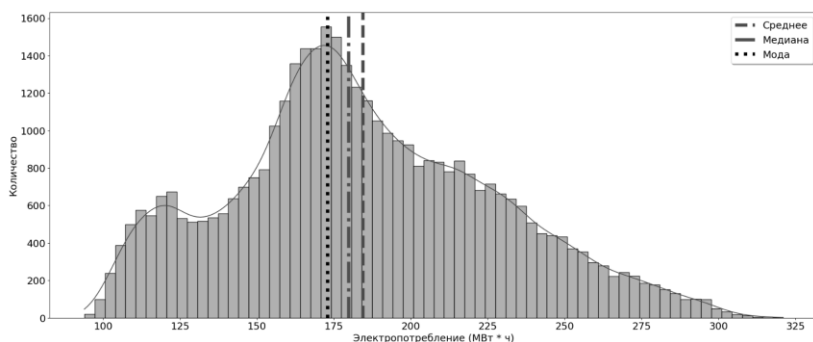


Рис. 2. Вероятностное распределение экспериментальных данных электропотребления за каждый час суток (весь период наблюдения)

По ней видно, что распределение отлично от нормального (нет симметрии относительно среднего; медиана не равна среднему значению и моде). Гипотеза о нормальности распределения на больших выборках (больше 5000 наблюдений) почти всегда отклоняется.

В том, что распределение отлично от нормального можно также убедиться, построив Q-Q plot (график Квантиль-Квантиль) (рис. 3).

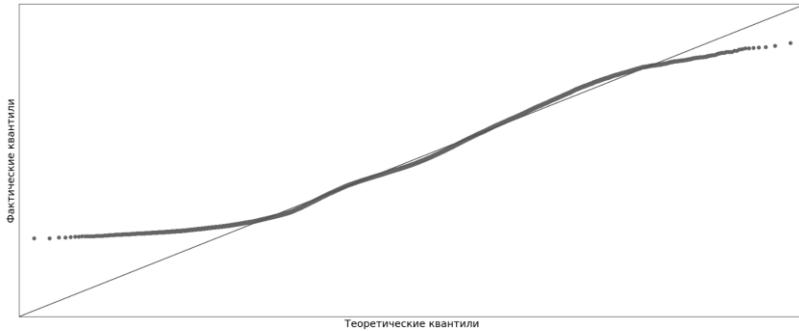


Рис. 3. График Квантиль-Квантиль

Q-Q plot — это вероятностный график для сравнения двух распределений вероятностей путем сопоставления их квантилей друг с другом [37]. Если два сравниваемых распределения подобны, точки на графике будут примерно лежать на тождественной линии $y = x$. В данном случае теоретические квантили — это значения из нормального распределения, а фактические — из распределения «Электропотребление». По графику на рис. 3 видно, что присутствуют сильные отклонения «на хвостах».

Был проведен анализ данных на аномальные значения с помощью графика Box Plot (рис. 4) [5]. По графику на рис. 4 видно, что некоторые наблюдения превышают верхнюю границу статистически значимой выборки (upper fence):

$$(3) \text{ upper_fence} = Q3 + 1,5 \cdot IQR,$$

где $Q3$ — третий квартиль; IQR — межквартильный размах ($Q3 - Q1$); $Q1$ — первый квартиль.

В результате проведённого системного анализа данных можно сделать вывод о том, что 0,218 % значений зависимой переменной являются аномальными. Эти аномальные значения приходятся на зимний период. Было решено их проигнорировать.

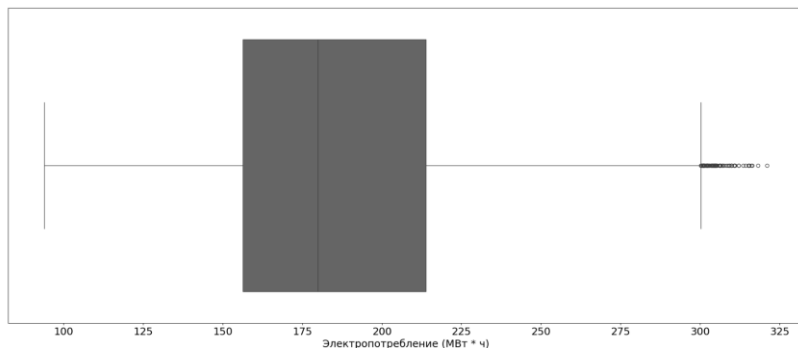


Рис. 4. График «Box plot»

Для того чтобы обнаружить тренд во временном ряде, строят скользящее среднее исследуемого процесса. Авторами было построено скользящее среднее электропотребления с окном в одну неделю (168 точек). Судя по графику на рис. 5, в ряде не выявлен ярко выраженный тренд.

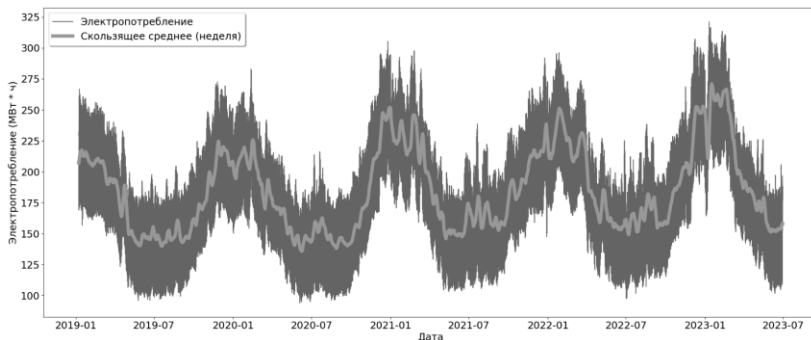


Рис. 5. Временной ряд со скользящим средним

Рис. 2 показывает, что исходное значение вероятности отличается от Гауссовского распределения. Поэтому для решения задачи прогнозирования электропотребления на каждый час следующих суток при разработке математической модели решено было использовать ансамблевый метод машинного обучения LightGBM, так как он позволяет моделировать нелинейные зависимости.

Также необходимо отметить, что с помощью алгоритма LightGBM можно моделировать и прогнозировать временные ряды, у которых отсутствует ярко выраженный тренд, что доказано выполненными расчётами и полученными результатами в ходе проведения данного исследования апробирования разработанной математической модели в энергетической компании ПАО «РОССЕТИ Северный Кавказ» в течение одного года (в период с июля 2022 по июнь 2023 гг.).

3. Процесс моделирования электропотребления

3.1. ИССЛЕДОВАНИЕ ВЛИЯНИЯ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ (ФАКТОРОВ) НА ПОТРЕБЛЕНИЕ ЭЛЕКТРОЭНЕРГИИ ЗА КАЖДЫЙ ЧАС СУТОК

Проведено исследование влияния независимых переменных (X) на зависимую переменную (Y – потребление электроэнергии). В качестве факторов использовались значения, связанные со временем: «Период времени суток», «Час», «День недели», «Выходной», «День в году», «Месяц». Также, использовались лаги целевой переменной (предыдущие значения): «Электропотребление лаг 1 день», «Электропотребление лаг 7 дней». Итоговая матрица признаков X (независимых переменных X) будет принимать следующий вид (рис. 6.).

Проанализировать связь между количественными значениями независимых переменных X и целевой (зависимой) переменной Y (потребление электроэнергии) поможет матрица корреляции (рис. 7). Коэффициент линейной корреляции Пирсона рассчитан по формуле

$$(4) \quad r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot (y_i - \bar{y})^2}}.$$

Между лагами зависимой переменной и самой зависимой переменной ярко выражена сильная положительная корреляционная связь.

Дата и время	Электропотребление	час	Период времени суток	День недели	Выходной	Месяц	День в году	Электропотребление лаг 1 день	Электропотребление лаг 7 дней
2019-01-08 00:00:00	195.697	0	0	1	0	1	8	191.298	215.191
2019-01-08 01:00:00	184.357	1	0	1	0	1	8	179.960	202.239
2019-01-08 02:00:00	174.130	2	0	1	0	1	8	172.706	198.462
2019-01-08 03:00:00	170.228	3	0	1	0	1	8	169.160	179.276
2019-01-08 04:00:00	169.023	4	0	1	0	1	8	167.794	173.035
—	—	—	—	—	—	—	—	—	—
2023-06-29 19:00:00	176.542	19	3	3	0	6	180	171.040	167.809
2023-06-29 20:00:00	188.429	20	3	3	0	6	180	183.232	177.173
2023-06-29 21:00:00	183.351	21	3	3	0	6	180	177.985	172.780
2023-06-29 22:00:00	170.048	22	3	3	0	6	180	164.705	158.533
2023-06-29 23:00:00	152.483	23	3	3	0	6	180	148.960	144.961

39216 rows × 9 columns

Рис 6. Матрица входных параметров

На рис. 7 показано, что коэффициент корреляции [4] между зависимой переменной, её дневным и недельным лагом равен 0,979 и 0,935 соответственно. Таким образом, доказана целесообразность использования лаговых переменных зависимой переменной Y_t -24 и Y_t -24·7, где t – время (час суток).

	Электропотребление	Электропотребление лаг 1 день	Электропотребление лаг 7 дней
Электропотребление	1.000	0.979	0.935
Электропотребление лаг 1 день	0.979	1.000	0.934
Электропотребление лаг 7 дней	0.935	0.934	1.000

Рис. 7. Матрица корреляции

Для анализа категориальных переменных были построены диаграммы Box plot (рис. 8–10), на которых графически отображена зависимость электропотребления от часа суток, от периода времени суток и от месяцев в году.

По графику на рис. 8 можно заметить, что есть определенная зависимость потребления электроэнергии от часа суток. Расход электроэнергии уменьшается в ночные часы и начинает увеличиваться с 5:00 утра. С 11:00 потребление электроэнергии немного уменьшается до 15:00. С 16:00 мы видим увеличение расхода электроэнергии до 18–19:00. Это поведение потребления электроэнергии доказывает необходимость включения независимой переменной «Час» в математическую модель.

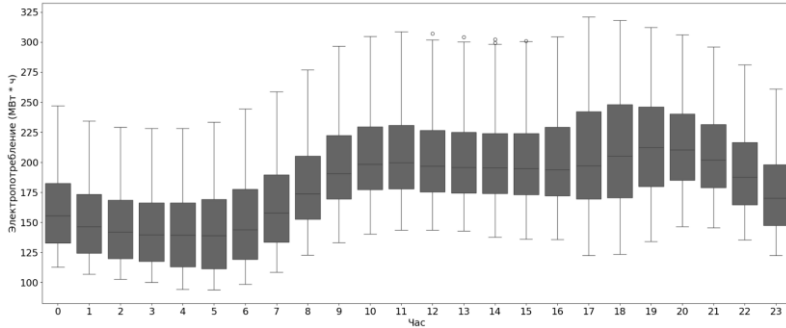


Рис. 8. Диаграмма Box plot, описывающая зависимость электропотребления от часа суток

Из рис. 9 заметно, что потребление электроэнергии увеличивается ближе к вечеру и уменьшается в ночной период суток.

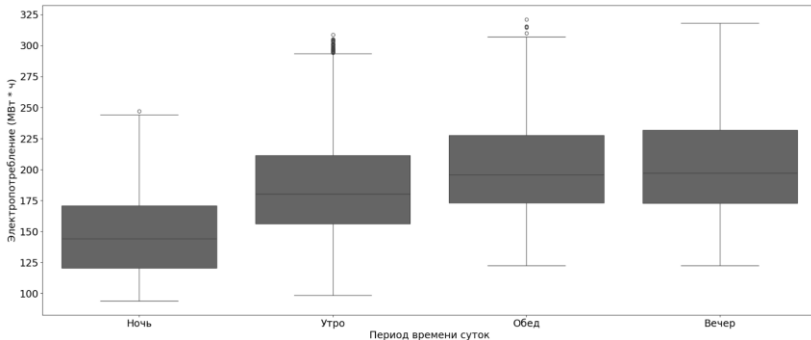


Рис. 9. Диаграмма Box plot описывающая зависимость электропотребления от периода времени суток

Рис. 10 показывает, что потребление электроэнергии уменьшается летом и увеличивается к зиме, что обосновывает включение независимой переменной «Месяц» в математическую модель для прогнозирования.

Графики на рис. 8–10 доказывают необходимость учёта значений вышеуказанных факторов при прогнозировании поведения потребления электроэнергии методами машинного обучения.

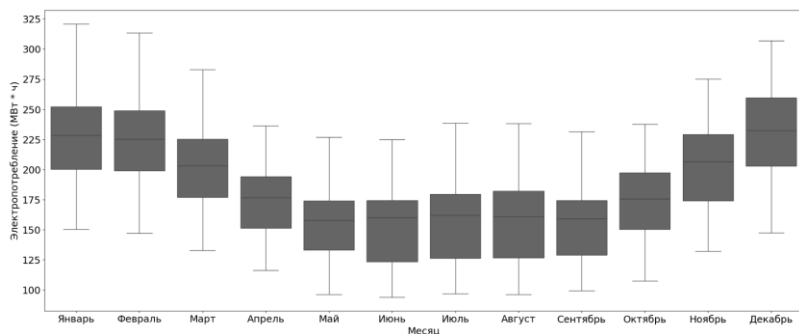


Рис. 10. Диаграмма Box plot описывающая зависимость электропотребления от месяца

3.2. МАТЕМАТИЧЕСКОЕ ОПИСАНИЕ МЕТОДОЛОГИИ МОДЕЛИРОВАНИЯ И ПРОГНОЗИРОВАНИЯ ЭЛЕКТРОПОТРЕБЛЕНИЯ

Алгоритм LightGBM представляет собой разработанную Microsoft библиотеку для обучения модели градиентного бустинга. LightGBM обладает многими преимуществами, включая разреженную оптимизацию, параллельное обучение, регуляризацию. LightGBM не наращивает дерево по уровням – строка за строкой – так, как это делают большинство других реализаций [33]. Вместо этого он добавляет листья деревьям. Алгоритм выбирает лист, который, по его мнению, приведет к наибольшему снижению значения функции потерь [18]. Кроме того, LightGBM не использует популярный алгоритм обучения деревьев решений на основе сортировки, который ищет наилучшую точку разделения на отсортированных значениях признаков, как это делает, например, XGBoost. Вместо этого LightGBM реализует оптимизированный алгоритм обучения дерева решений на основе гистограмм, который дает большие преимущества как в эффективности, так и в потреблении памяти.

Математическая модель градиентного бустинга для прогнозирования потребления электроэнергии на каждый час суток (5) была построена путём последовательного добавления деревьев решений, каждое из которых корректирует ошибки предыдущих деревьев.

Пусть $F(x)$ – целевая функция процесса потребления электроэнергии на каждый час следующих суток, представляющая прогноз разработанной математической модели. На t -м шаге обучения математическая модель прибавляет новое дерево $h_t(x)$, которое минимизирует ошибку моделирования электропотребления на текущем этапе. Формула математической модели для прогнозирования электропотребления на каждый час следующих суток имеет вид

$$(5) \quad F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x),$$

где η – коэффициент обучения, регулирующий скорость изменения модели.

Алгоритм метода машинного обучения LightGBM использует функцию потерь $L(y, F(x))$, например, среднеквадратическую ошибку для задач регрессии, которая представлена в формуле

$$(6) \quad L(y, F(x)) = \frac{1}{2} (y - F(x))^2.$$

Цель данного этапа исследования заключалась в поиске таких значений модели (5), которые бы минимизировали функции потерь на каждом этапе:

$$(7) \quad \arg \min_h \sum_{i=1}^N L(y_i, F_{t-1}(x_i) + h(x_i))^2.$$

Однако вместо прямого минимизирования функции потерь при прогнозировании электропотребления на каждый час суток метод машинного обучения LightGBM использует градиентный спуск. На t -м шаге алгоритм строит дерево, обученное на остатках (невязках), которые являются антиградиентом функции потерь:

$$(8) \quad g_i = \frac{\partial L(y_i, F_{t-1}(x_i))}{\partial F(x_i)},$$

где g_i – значение градиента для i -го объекта (значения независимых и зависимой переменных).

На этапе моделирования электропотребления мы использовали две техники метода LightGBM для повышения его эффективности:

– Gradient-Based One-Side Sampling (GOSS): отбор объектов (значений переменных модели) на основе значений их градиентов. Эта техника выбирает данные независимых и зависимых переменных с большими градиентами, что позволяет уменьшить размер исходной выборки (39 384 строк) и ускорить обучение без потери точности моделирования электропотребления;

– Exclusive Feature Bundling (EFB): объединение характеристик, которые редко принимают ненулевые значения одновременно. Это уменьшает размерность данных независимых и зависимой переменной и улучшает эффективность алгоритма.

Эти методы и другие особенности реализации метода, такие как разреженное представление данных и метод обучения на основе гистограмм, делают LightGBM оптимальным для задач, требующих работы с большими объёмами данных, такими как прогнозирование потребления электроэнергии на каждый час суток, а также нелинейными зависимостями процесса потребления электроэнергии с найденными факторами, описанными выше в данном исследовании [18].

Эти методы позволяют алгоритму снижать вычислительную сложность и потребление памяти, при этом сохраняя высокую точность. Далее приведен математический аппарат, описывающий каждый из методов.

GOSS фокусируется на уменьшении выборки исходных данных о потреблении электроэнергии за каждый час суток за предыдущие (3,5 года), не теряя при этом важной смысловой информации. Идея GOSS состоит в том, что объекты (значения переменных модели) с высокими значениями градиентов оказывают большее влияние на ошибку модели. Таким образом, GOSS сохраняет все объекты с высокими значениями градиентов и применяет случайную выборку к объектам низкими значениями градиентов. Это позволяет уменьшить общий размер выборки, ускоряя обучение модели.

Пусть g_i – градиент для i -го объекта; a – доля объектов с высокими значениями градиентов, которые сохраняются; b – доля объектов с низкими значениями градиентов, которые выбираются случайно. Тогда алгоритм функционирует следующим образом:

Шаг 1. Сортировка градиентов. Сначала объекты сортируются по абсолютному значению градиента $|g_i|$.

Шаг 2. Отбор экземпляров (определённых выборок):

- сохраняются верхние a 100% объектов по абсолютным значениям $|g_i|$.

- случайным образом выбирается $b \cdot 100\%$ из оставшихся объектов с низкими значениями градиентов.

Шаг 3. Компенсация пропорций. Для компенсации уменьшения количества данных объекты с малыми значениями градиентов масштабируются множителем $1 - a/b$ при вычислении функции потерь, чтобы уравновесить вклад различных групп данных.

Таким образом, целевая функция метода GOSS принимает вид

$$(9) \quad L = \sum_{i \in \text{high_grad}} g_i^2 + \frac{1-a}{b} \cdot \sum_{j \in \text{low_grad}} g_j^2.$$

где `high_grad` и `low_grad` обозначают соответственно подмножества данных с высокими и низкими значениями градиентов.

Считаем необходимым отметить, что GOSS — это метод, использующий тот факт, что в Gradient Boosted Decision Tree (GBDT) нет собственного веса экземпляра данных. Поскольку экземпляры данных с гетерогенными градиентами играют разную роль в вычислении *information gain*, или «приросте информации» (в теоретико-информационном подходе является важным критерием в машинном обучении моделей, используется для количественной оценки релевантности и зависимости независимых переменных от зависимой переменной, синоним «расхождения Куллбэка – Лейблера»), экземпляры с большими градиентами будут вносить более значимый вклад в значение этого критерия. Таким образом, чтобы сохранить точность информации, GOSS накапливает экземпляры с большими градиентами и случайным образом удаляет экземпляры с малыми градиентами. Чтобы компенсировать влияние на распределение данных, при вычислении *information gain* GOSS вводит постоянный множитель для экземпляров данных с малыми градиентами. В частности, GOSS сначала сортирует эти экземпляры в соот-

ветствии с абсолютным значением их градиентов и выбирает верхние экземпляры $a \cdot 100\%$. Затем случайным образом выбирает $b \cdot 100\%$ экземпляров из остальных данных. После этого GOSS увеличивает выборочные данные с малыми градиентами на константу $(1 - a)/b$ при расчете значений information gain. Поступая так, уделяется больше внимания недостаточно обученным экземплярам без существенного изменения исходного распределения данных.

В результате проведенных расчетов можно сделать вывод о том, что метод GOSS позволил нам ускорить обучение за счет сокращения экспериментальных данных о потреблении электроэнергии и при этом сохранил точность оценки градиентов.

Второй метод – EFB – решает проблему высокой размерности данных о потреблении электроэнергии, объединяя взаимно-исключающие признаки в одну группу, что позволяет сократить общее количество признаков (факторов, независимых переменных), не теряя информации. Далее представлен алгоритм работы.

Шаг 1. Определение взаимной исключительности.

Два признака считаются взаимно-исключающими, если они редко одновременно принимают ненулевые значения в выборке. Например, если x_i и x_j – признаки и вероятность того, что $x_i \neq 0$ и $x_j \neq 0$ одновременно, очень мала, они могут быть объединены в один признак.

Шаг 2. Построение набора эксклюзивных признаков:

- все признаки разбиваются на несколько «пакетов», или «наборов», каждый из которых содержит только взаимно-исключающие признаки;

- если m – исходное число признаков, то после применения EFB количество новых признаков m' будет существенно меньше m , что снижает размерность.

Шаг 3. Объединение признаков.

Пусть B_k – объединённый набор признаков, который состоит из признаков, удовлетворяющих условию взаимной исключительности. Тогда новый признак f_{B_k} вычисляется для каждого набора B_k как

$$(10) f_{B_k} = \sum_{i \in B_k} x_i,$$

где x_i – исходные признаки (факторы).

EFB – это практически безубыточный метод сокращения количества значимых факторов. В разреженном признаковом пространстве многие признаки почти исключительные. Это означает, что они редко одновременно принимают ненулевые значения. Факторы, которые закодированы с помощью метода One-hot encoding, – наглядный пример исключительных признаков. EFB объединяет эти признаки, уменьшая размерность для того чтобы повысить эффективность алгоритма, сохраняя при этом высокий уровень точности. Набор таких независимых переменных, объединённых в один признак, называется *exclusive feature bundle*, или набор эксклюзивных факторов.

В результате проведенного этапа данного исследования можно сделать вывод о том, что метод EFB сократил размерность признакового пространства, что снизило вычислительную нагрузку и потребление памяти.

3.3. ПОДБОР ГИПЕРПАРАМЕТРОВ МОДЕЛИ

Алгоритм LightGBM объединяет два вышеописанных метода при анализе значимости независимых переменных и повышают эффективность математической модели, обеспечивая множество преимуществ по сравнению с другими моделями градиентного бустинга [16].

Однако решающее значение в применении модели LightGBM имеет проведенное исследование по выявлению факторов, влияющих на электропотребление и установление (настройка) оптимальных значений гиперпараметров, чтобы математическая модель не переобучилась, а также не показала слабую адекватность имеющимся экспериментальным данным по потреблению электроэнергии на каждый час следующих суток. Авторы Khanh-Toan Nguyen, Thanh-Ngoc Tran, Huy-Tuan Nguyen в своей статье [20] привели достаточно доказательств в пользу того, что использование гиперпараметров по умолчанию приводит к повышению ошибок прогнозирования.

В данном исследовании при построении математических моделей (5) для временных рядов при решении задачи моделирования и прогнозирования электропотребления на каждый час суток для энергосбытовых компаний мы учитывали временную зависимость данных и для этого применили метод кросс-валидации на скользящем окне (TimeSeriesSplit) в сочетании с настройкой гиперпараметров с помощью GridSearchCV, что обеспечило нам выбор оптимальных параметров математической модели (1) при учёте временных взаимосвязей [13].

В данном исследовании с помощью настройки GridSearchCV мы выполнили систематический перебор комбинаций гиперпараметров для модели и нашли такие значения, которые минимизировали ошибку моделирования электропотребления.

Пусть у нас есть модель с параметрами

$$(11) \theta = (\theta_1, \theta_2, \dots, \theta_k),$$

каждый из которых может принимать значения из множества V_i . Пространство поиска для этих параметров составляет декартово произведение

$$(12) V = V_1 \times V_2 \times \dots \times V_k.$$

С помощью метода GridSearchCV мы последовательно оценили математическую модель на каждом элементе множества V , что позволило выбрать оптимальные параметры θ^* , минимизирующие метрику ошибки на валидационных данных:

$$(13) \theta^* = \arg \min_{\theta \in V} \text{Ошибка}(\theta).$$

С помощью метода TimeSeriesSplit мы адаптировали кросс-валидацию для временных рядов значений электропотребления, где данные делятся на обучающую и валидационную выборки, соблюдая порядок времени и избегая утраты взаимосвязей. На каждом разбиении обучающая выборка расширялась, добавляя более ранние временные периоды, и проверка производилась на следующих временных периодах. Для n разбиений с помощью метода TimeSeriesSplit подвыборки были организованы следующим образом:

$$(14) \text{Train_1} = [1, 2, \dots, k], \quad \text{Test_1} = [k + 1, \dots, k + m],$$

$$\text{Train_2} = [1, 2, \dots, k + m], \quad \text{Test_2} = [k + m + 1, \dots, k + 2m]$$

и т.д., где k – размер обучающей выборки, m – размер тестовой выборки [13].

Таким образом, на каждом шаге сохраняется структура временного ряда со значениями потребления электроэнергии за каждый час суток.

Предположим, что пространство поиска состоит только из двух гиперпараметров, один из которых значим, а другой – нет. Мы хотим настроить их так, чтобы повысить точность модели. Если каждый из них имеет 3 разных значения, то всё пространство поиска имеет 9 возможных вариантов. Можно попробовать каждый из них, чтобы найти оптимальное значение для обоих гиперпараметров. Этот подход и называется Grid Search.

Проблема с кросс-валидацией на временных рядах состоит в том, что он имеет временную структуру и случайно перемешивать значения всего ряда без сохранения этой структуры нельзя, иначе в процессе потеряются все взаимосвязи наблюдений друг с другом. Для этого необходимо использовать метод под названием TimeSeriesSplit. Суть метода заключается в следующем: начиная с некоторой подвыборки обучающих данных выдаются прогнозные значения на следующий временной период (test_size) с определенным сдвигом (gap) и оценивается их качество. Затем тот период, на который был выдан прогноз (т.е. значения независимых и зависимой переменной), добавляется в исходную подвыборку данных. Далее выдается прогноз на следующий определенный временной период (test_size) и так повторяется 5 раз (n_splits) (рис. 11).

Авторами минимизировалось значение MAE, а для метода TimeSeriesSplit были установлены следующие аргументы: test_size (размер тестовой выборки) = 2160, n_splits (количество разбиений) = 5, gap (сдвиг по времени для оценки) = 120. Были подобраны $n_estimators$ (количество решающих деревьев) и max_depth (их глубина). Оптимальными параметрами оказались $n_estimators = 70$, $\text{max_depth} = 6$.

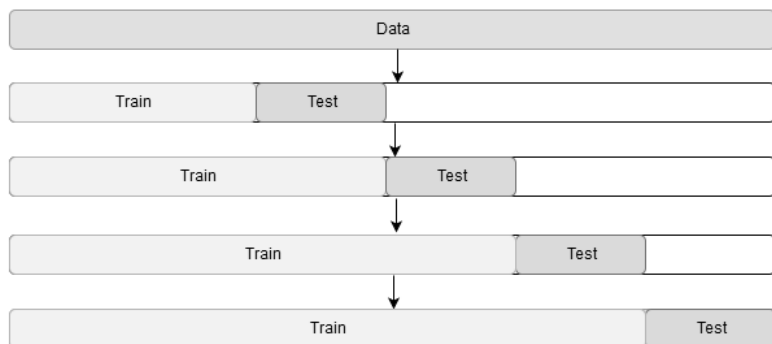


Рис. 11. Принцип метода *TimeSeriesSplit*

Необходимо отметить, что в качестве преимуществ методов *TimeSeriesSplit* и *GridSearchCV* при решении задачи моделирования электропотребления было выявлено, что сочетание алгоритмов *TimeSeriesSplit* и *GridSearchCV* минимизировало одну из важнейших проблем машинного обучения – «проблему утечки» (ситуация в моделировании, когда на этапе построения модели используются значения независимых переменных, недоступные при расчете прогнозной оценки) и обеспечило корректную валидацию временных рядов электропотребления, где текущие данные не должны использовать будущие данные для обучения математической модели (1). Сделанный нами вывод подтверждается проведенным экспериментом, тестированием на реальных данных в компании ПАО «Россети Северный Кавказ» в период с 2022 по 2023 гг. и полученными научно-обоснованными результатами моделирования электропотребления, адекватно описывающими производственные данные.

3.3. ОЦЕНКА ЗНАЧИМОСТИ ФАКТОРОВ

После того как была разработана математическая модель, необходимо оценить значимость факторов. Для этого можно построить столбчатую диаграмму критерия прироста информации (Information Gain) для каждого фактора (рис. 12). В машинном обучении прирост информации – это количество информа-

ции, полученной о случайной величине в результате наблюдения другой случайной величины [27]:

$$(15) \quad IG(S, a) = H(S) + H(S | a),$$

где $IG(S, a)$ – прирост информации для набора данных S ; $H(S)$ – информационная энтропия набора данных S до разбиения; $H(S|a)$ – это условная информационная энтропия набора данных S с разбиением a :

$$(16) \quad H(S | a) = \sum_v \frac{S_a(v)}{S} \cdot H(S_a(v)),$$

где $S_a(v)/s$ – отношение количества примеров в наборе данных с переменной a , имеющей значение v , и $H(S_a(v))$ – это информационная энтропия группы образцов, где переменная a имеет значение v . Информационным критерием качества для задачи регрессии является MSE .



Рис. 12. Столбчатая диаграмма Information gain

Как видно из графика на рис. 12, по имеющимся данным, оказалось, что большей значимостью из всех факторов обладают лаговые переменные.

3.3. АНАЛИЗ РЕЗУЛЬТАТОВ РАЗРАБОТАННЫХ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ ПО РАЗЛИЧНЫМ МЕТОДАМ МАШИННОГО ОБУЧЕНИЯ

Необходимо отметить, что в ходе исследования был проведён анализ различных методов машинного обучения: CatBoost [23], Huber Regressor [28], Random Forest [12], Linear Regression [15], LightGBM. Результаты оценки качества моделей по крите-

риям $RMSE$ (корень из среднеквадратичной ошибки), MAE (средняя абсолютная ошибка), $MAPE$ (средняя абсолютная процентная ошибка), $R^2_{adjusted}$ (скорректированный коэффициент детерминации) и представлены в таблицах 1–5.

Таблица 1. Результаты оценки качества модели CatBoost

	Train	Validation	Test
$RMSE$	4,594	8,607	8,821
MAE	3,287	6,737	6,811
$MAPE$ (%)	1,805	3,853	3,510
$R^2_{adjusted}$	0,988	0,925	0,965

Таблица 2. Результаты оценки качества модели Huber Regression

	Train	Validation	Test
$RMSE$	8,126	9,689	8,993
MAE	5,797	7,253	6,420
$MAPE$ (%)	3,189	3,960	3,230
$R^2_{adjusted}$	0,961	0,908	0,964

Таблица 3. Результаты оценки качества модели Random Forest

	Train	Validation	Test
$RMSE$	2,742	7,656	8,439
MAE	1,938	5,497	5,897
$MAPE$ (%)	1,078	2,989	2,903
$R^2_{adjusted}$	0,996	0,941	0,968

Таблица 4. Результаты оценки качества Linear Regression

	Train	Validation	Test
$RMSE$	8,099	8,569	9,011
MAE	5,923	6,227	6,525
$MAPE$ (%)	3,288	3,405	3,290
$R^2_{adjusted}$	0,962	0,927	0,964

Таблица 5. Результаты оценки качества модели LightGBM

	Train	Validation	Test
$RMSE$	6,180	7,490	8,172
MAE	4,493	5,328	5,777
$MAPE$ (%)	2,490	2,893	2,847
$R^2_{adjusted}$	0,978	0,944	0,970

Для наглядного представления информации о точности модели были построены графики. Самый простой график – линейный. На нём необходимо отразить фактические и прогнозные значения. Проблема состоит в отображении больших объёмов данных: они будут иметь нерепрезентативный вид. На рис. 13 отображены первые 720 фактических и прогнозных значений на тестовой выборке.

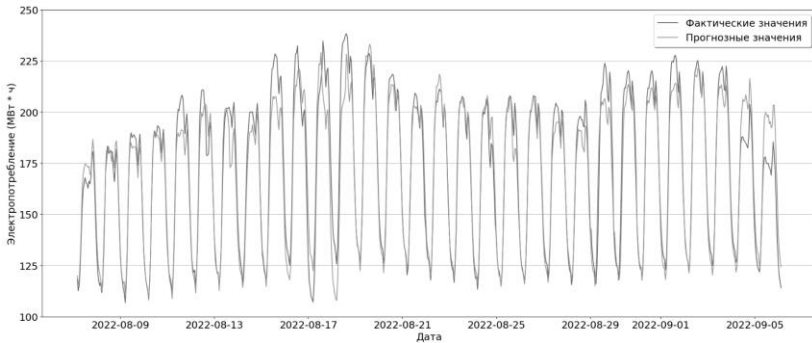


Рис. 13. График фактических и прогнозных значений

Для отображения всех точек проще воспользоваться диаграммой рассеяния. Диаграмма рассеяния изображает значения двух переменных в виде точек на декартовой плоскости. На диаграмме рассеяния каждому наблюдению соответствует точка, координаты которой равны значениям каких-либо двух параметров этого наблюдения. Диаграммы рассеяния используются для демонстрации наличия или отсутствия корреляции между двумя переменными [2].

На рис. 14 представлена диаграмма рассеяния фактических и прогнозных значений электропотребления. В заголовке графика также отражён коэффициент корреляции между ними, он равен 0,986.

Данный график демонстрирует отличную точность прогнозирования с помощью алгоритма LightGBM.

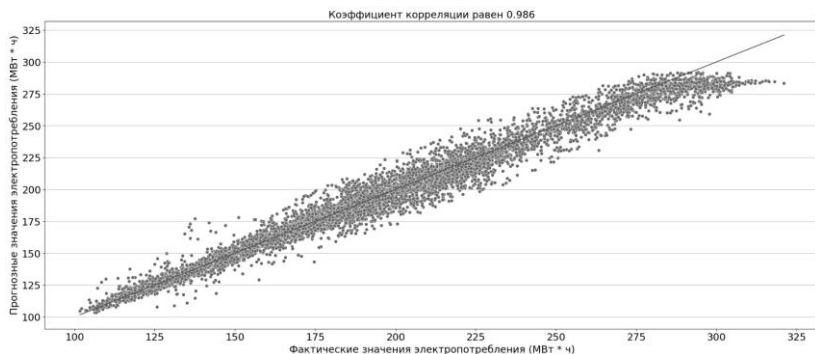


Рис. 14. Диаграмма рассеяния фактических и прогнозных значений электропотребления

Проведена практическая апробация разработанной новой, полезной и качественной математической модели на данных одной из ведущих энергетических компаний Российской Федерации ПАО «Россети Северный Кавказ». При моделировании и прогнозировании потребления электроэнергии на каждый час суток в течение 24 дней получены следующие результаты:

- средняя абсолютная ошибка прогнозирования электропотребления на каждый час следующих (первых) суток составила 3,120%;
- средняя абсолютная ошибка прогнозирования электропотребления на каждый час вторых суток составила 3,70%;
- средняя абсолютная ошибка прогнозирования электропотребления на каждый час третьих суток составила 4%.
- средняя абсолютная ошибка прогнозирования электропотребления за три дня составила 3,606 %.

3.3. АНАЛИЗ ОСТАТКОВ РАЗРАБОТАННОЙ МОДЕЛИ

Важным этапом оценки разработанной регрессионной модели является анализ её остатков. Модель должна одинаково хорошо прогнозировать как большие значения целевой переменной, так и малые. Недопустимы также большие отклонения по абсолютной величине. Гистограм-

ма должна быть симметрична относительно нуля. Медиана, мода и среднее должно стремиться к нулю. На рис. 15 представлена гистограмма распределения остатков модели. Из рис. 16 видно, что отклонения не превышают границ $0 \pm 1\sigma$.

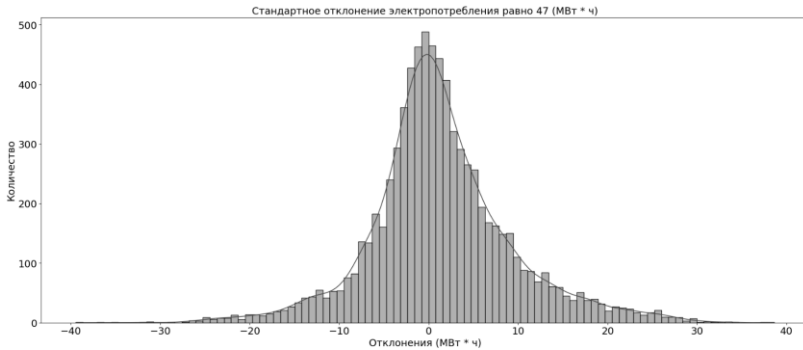


Рис. 15. Гистограмма распределения остатков

При анализе диаграммы отклонений Box Plot, представленной на рис. 16, было выявлено, что 6,719% остатков выходят за границы:

Lower fence:

$$Q1 - (1,5 \cdot IQR) = -2,799 - (1,5 \cdot 7,893) = -14,638 \text{ (МВт} \cdot \text{ч)};$$

Upper fence:

$$Q3 + (1,5 \cdot IQR) = 5,097 + (1,5 \cdot 7,893) = 16,936 \text{ (МВт} \cdot \text{ч)},$$

где Lower fence – нижняя граница статистически значимой выборки данных; Upper fence – верхняя граница статистически значимой выборки данных; $Q1$ – первый квартиль; $Q3$ – третий квартиль; IQR – межквартильный размах.

Также были рассчитаны максимальная положительная ошибка и максимальная отрицательная ошибка прогнозирования:

$$\text{Максимальная положительная ошибка} = 39,296 \text{ (МВт} \cdot \text{ч)};$$

$$\text{Максимальная отрицательная ошибка} = -39,733 \text{ (МВт} \cdot \text{ч)}.$$

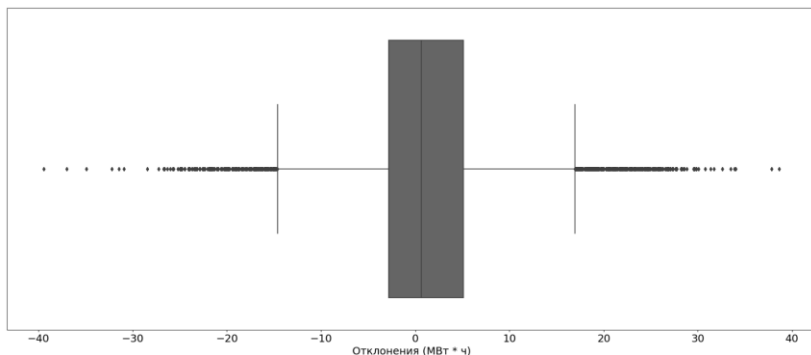


Рис. 16. Диаграмма отклонений BOX PLOT

4. Выводы и рекомендации

В результате проведенного экспериментального исследования была разработана новая, полезная, адекватная и качественная регрессионная модель для системы поддержки принятия решений при прогнозировании поведения электропотребления в энергетической компании. Исследование показало, что применение разработанной модели позволяет улучшить точность прогнозов, оптимизировать распределение ресурсов и повысить эффективность управления энергетическими процессами и ресурсами в большой электроэнергетической системе Российской Федерации.

Системный анализ результатов показал, что разработанная математическая модель способствует снижению издержек компании за счет оптимального использования электроэнергии и предотвращения возможных простоев или перерасхода ресурсов. Это обеспечивает более стабильную и надежную работу энергетических систем, снижает риски финансовых потерь и способствует экономической эффективности энергетических и производственных предприятий.

Полученные результаты исследовательской работы могут быть использованы для улучшения оперативного управления и планирования энергопотреблением, что снизит издержки компаний и повысит их конкурентоспособность на энергетическом

рынке. Для производственных предприятий корректное прогнозирование помогает точнее планировать производственные мощности и оптимально распределять ресурсы. Для энергетических компаний – обеспечивать стабильность, эффективность и устойчивость работы энергосистемы, а также экономически выгодное энергоснабжение в регионе.

С помощью технологии искусственного интеллекта Boosting разработана математическая модель LightGBM, которая позволяет определенной энергетической компании моделировать и прогнозировать поведение потребления электроэнергии со средней абсолютной ошибкой за три дня 3,606% (МВт·ч.).

Литература

1. АЛКАЦЕВ М.И., ДЗГОЕВ А.Э., БЕТРОЗОВ М.С. *Исследование и разработка метода прогнозирования потребления электроэнергии в системе управления электроснабжением региона* // Известия высших учебных заведений. Проблемы энергетики. – 2012. – № 5-6. – С. 30–37.
2. БЛАГОВЕЩЕНСКИЙ Ю. *Тайны корреляционных связей в статистике*. – М.: Научная Книга, Инфра-М, 2009. – 158 с.
3. ВАЛЬ П.В., ПОПОВ Ю.П. *Концепция разработки системы прогнозирования электропотребления промышленного предприятия в условиях оптового рынка* // Промышленная энергетика. 2011. – №10. – С. 31-35.
4. ГМУРМАН В. *Теория вероятностей и математическая статистика: Учебное пособие для вузов*. – М.: Высшая школа, 2004. – 479 с.
5. ГРЕШИЛОВ А.А., СТАКУН В.А., СТАКУН А.А. *Математические методы построения прогнозов*. – М.: Радио и связь, 1997. – 112 с.
6. ДЗГОЕВ А.Э. *Методы обработки и анализа данных для разработки предиктивных моделей : учебное пособие*. – М.: РТУ МИРЭА, 2024. – 147 с. // Лань : электронно-библиотечная система. – URL: <https://e.lanbook.com/book/420860> (дата обращения: 15.10.2024).

7. КАРПЕНКО С.М., КАРПЕНКО Н.В., БЕЗГИНОВ Г.Ю. *Прогнозирование электропотребления на горнопромышленных предприятиях с использованием статистических методов* // Горная промышленность. – 2022. – №1. – С. 82–88.
8. КАРПЕНКО С.М., КАРПЕНКО Н.В., САХАРОВ В.Е. *Моделирование электропотребления на основе данных с переменной структурой* // Энергобезопасность и энергосбережение. – 2021. – №1. – С. 13–17.
9. КУРБАНГАЛИЕВ У.К., ЛИСИЦИН Н.В. *Требования к коммерческому учёту электрической энергии и мощности в условиях оптового рынка* // Энергетик. – №1. – 1996. – С. 2–4.
10. МАТЮНИНА Ю.В., МАКАРЕНКО И.Г. *Работа потребителей на рынках электроэнергии: учебное пособие*. – М.: Издательский дом МЭИ, 2008. – 64 с.
11. МОРГОЕВ И.Д., ДЗГОЕВ А.Э., КЛЮЕВ Р.В. и др. *Прогнозирование потребления электроэнергии предприятиями народнохозяйственного комплекса в условиях неполноты информации* // Известия Кабардино-Балкарского научного центра РАН. – 2022. – №3(107). – С. 9–20.
12. BERGMEIR C., BENÍTEZ J. M. *On the use of cross-validation for time series predictor evaluation* // Information Sciences. – 2012. – Vol. 191. – P. 192–213. – DOI: 10.1016/j.ins.2011.12.028.
13. BREIMAN L. *Random Forests* // Machine Learning. – 2001. – Vol. 45(1). – P. 5–32.
14. CHUNG J., JANG B. *Accurate prediction of electricity consumption using a hybrid CNN-LSTM model based on multivariable data* // PLoS ONE. – 2022. – Vol. 17(11). – P. e0278071. – DOI: 10.1371/journal.pone.0278071.
15. DALAL S., LILHORE DR U.K., SETH B. et al. *A Hybrid Model for Short-Term Energy Load Prediction Based on Transfer Learning with LightGBM for Smart Grids in Smart Energy Systems* // Journal of Urban Technology. – 2024. – DOI: 10.1080/10630732.2024.2380639.
16. FILIPOVA-PETRAKIEVA S.K., DOCHEV V. *Short-Term Forecasting of Hourly Electricity Power Demand: Regresion and Cluster Methods for Short-Term Prognosis* // Engineering,

- Technology & Applied Science Research. – 2022. – Vol. 12, No. 2. – P. 8374–8381. – DOI: 10.48084/etasr.4787.
17. FREEDMAN D.A. *Statistical Models: Theory and Practice*. – Cambridge University Press, 2009. – P. 26.
18. GUANG YANG, SONGHUI DU, QINGLING DUAN et al. *A Novel Data-Driven Method for Medium-Term Power Consumption Forecasting Based on Transformer-LightGBM* // Mobile Information Systems. – 2022. – P. 1–16. – DOI: 10.1155/2022/5465322.
19. GUPTA P., BAGCH A. *Introduction to Pandas* // In: Essentials of Python for Artificial Intelligence and Machine Learning. – December 2023. – DOI: 10.1007/978-3-031-43725-0_5.
20. HUBER P.J., RONCHETTI E.M. *Robust Statistics*. – John Wiley & Sons, Inc., Hoboken. – 354 p. – DOI: 0.1002/9780470434697.
21. IFTIKHAR H., ZYWIOLEK J., LÓPEZ-GONZALES J.L. et al. *Electricity consumption forecasting using a novel homogeneous and heterogeneous ensemble learning* // Front. Energy Res. – 2024. – Vol. 12. – 1442502. – DOI: 10.3389/fenrg.2024.1442502.
22. KE GUOLIN, MENG QI, FINLEY T. et al. «*LightGBM: A Highly Efficient Gradient Boosting Decision Tree*» // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.
23. KHANH-TOAN NGUYEN, THANH-NGOC TRAN, HUY-TUAN NGUYEN *Research on the Influence of Hyperparameters on the LightGBM Model in Load Forecasting* // Engineering, Technology & Applied Science Research. – 2024. – Vol. 14, No. 5. – P. 17005–17010.
24. LEE M.H.L., SER Y.C., SELVACHANDRAN G. et al. *A Comparative Study of Forecasting Electricity Consumption Using Machine Learning Models* // Mathematics. – 2022. – Vol. 10. – P. 1329. – DOI: 10.3390/math10081329.
25. LIN Y., LUO H., WANG D. et al. *An ensemble model based on machine learning methods and data preprocessing for short-term electric load forecasting* // Energies. – 2017. – Vol. 10. – P. 1186.

26. MAHANTA N., TALUKDAR R. *Forecasting of Electricity Consumption by Seasonal Autoregressive Integrated Moving Average Model in Assam* // India Int. Journal of Energy Economics and Policy. – 2024. – Vol. 14(5). – P. 393–400.
27. MESQUITA LOPES CABREIRA M., LEITE COELHO DA SILVA F., DA SILVA CORDEIRO J. et al. *A Hybrid Approach for Hierarchical Forecasting of Industrial Electricity Consumption in Brazil* // Energies. – 2024. – Vol. 17. – P. 3200. – DOI: 10.3390/en17133200.
28. NOWOZIN S. *Improved information gain estimates for decision tree induction* // arXiv preprint. – arXiv:1206.4620. – 2012.
29. PROKHORENKOVA L., GUSEV G., VOROBEV A. et al. *Catboost: unbiased boosting with categorical features* // In: Advances in Neural Information Processing Systems. – 2018. – P. 6639–6649.
30. QURESHI M., ARBAB M.A., REHMAN S.U. *Deep learning-based forecasting of electricity consumption* // Scientific Reports. – 2024. – Vol. 14. – P. 6489. – DOI: 10.1038/s41598-024-56602-4.
31. RAKHMONOV I.U., NIYOZOV N.N., KURBONOV N.N. et al. *Forecasting of electricity consumption by industrial enterprises with a continuous nature of production* // E3S Web of Conferences. – 2023. – Vol. 384. – P. 01030. – DOI: 10.1051/e3sconf/202338401030.
32. SHAH I., ARSHAD I., SAJID A. *Statistical Techniques for Modelling and Forecasting Pakistan's Electricity Consumption* // Forecasting. – 2023. – Vol. 1. – P. 1–14.
33. SHI H. *Best-first decision tree learning*. Diss. The University of Waikato, 2007.
34. SON N., SHIN Y. *Short- and Medium-Term Electricity Consumption Forecasting Using Prophet and GRU* // Sustainability. – 2023. – Vol. 15. – P. 15860. – DOI: 10.3390/su152215860.
35. YILDIZ U., KORKUT S.O. *Electricity Consumption Forecasting using the Prophet Model in Industry: A Case Study* // Int. Conf. on Computing, Intelligence and Data Analytics. – 2024. – DOI: 10.1007/978-3-031-53717-2_10.

36. WILK M.B., GNANADESIKAN R. *Probability plotting methods for the analysis of data* // Biometrika. Biometrika Trust. – 1968. – Vol. 55(1). – P. 1–17.
37. ZHANG D., GONG Y. *The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure* // IEEE Access. – 2020. – Vol. 8. – P. 220990–221003. – DOI: 10.1109/ACCESS.2020.3042848.

SYSTEM DATA ANALYSIS AND MODELING OF ELECTRICITY CONSUMPTION BEHAVIOR FOR EVERY HOUR OF THE DAY USING MACHINE LEARNING METHODS

Alan Dzgoev, Cand. Sc., associate Professor of the Department of Digital Transformation, RTU MIREA, Moscow (dzgoev@mirea.ru).

Anna Lagunova, Cand. Sc., associate Professor, Head of the Department of Digital Transformation, RTU MIREA, Moscow (lagunova@mirea.ru).

Stanislav Karatsev, Senior lecturer of the basic department "Corporate infocommunication systems", Financial University under the Government of the Russian Federation (Vladikavkaz branch), Vladikavkaz (stkaratsev@fa.ru).

Ilya Konushok, Head of Artificial Intelligence Development Department, AI Developer, Restadviser LLC, St. Petersburg (konushok@mail.ru).

Ivan Komarov, General Director of Frisson LLC, St. Petersburg (info@frisson.ai).

Igor Huzmiev, Cand. Sc., Deputy Director of the Department of Corporate and Technological Automated Control Systems of PJSC "ROSSETI Severny Kavkaz", Pyatigorsk (huzmiev-im@rossetisk.ru).

Oleg Gladushev, Chief Expert of the Prospective Development Service, Branch of JSC "SO UES" of the North-West Regional Control Department, St. Petersburg (gladushev-oia@odusz.so-ups.ru).

Abstract: Electricity consumption is a key driver of sustainable development in the energy industry, and accurately predicting its changes is essential for the efficient management of large electric power systems and resources. The aim of this study is

to develop a mathematical (regression) model for predicting the behavior of electricity consumption for each hour of the next day for energy supply companies using modern methods of machine learning and artificial intelligence. This article discusses various artificial intelligence methods used to model and predict electricity consumption. These methods include a linear model, a random forest, and two implementations of gradient boosting over decision trees. A scientific approach based on Boosting artificial intelligence technology allows to minimize the error in forecasting electricity consumption in large energy companies. The authors have developed a new, useful and high-quality regression model that adequately describes experimental data on electricity consumption for each hour of the day. The developed regression model was tested on real production data of an energy company. The conducted research and the obtained results allow the authors to conclude that the mathematical model developed using the machine learning method – LightGBM, can be used by energy supply companies for hourly planning of electricity consumption when submitting applications to the wholesale electricity and capacity market (WECM) for several days in advance. The research was carried out in the Python programming language.

Keywords: energy consumption modeling, data processing, model quality assessment, mathematical statistics, regression model, artificial intelligence.

УДК 004.8+004.6+519.2+ 004.942+ 519.7+303.732.4

ББК 22.172+ 22.18

*Статья представлена к публикации
членом редакционной коллегии Я.И. Квинто.*

*Поступила в редакцию 25.10.2024.
Опубликована 31.03.2025.*