ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ПРОФЕССИОНАЛЬНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

МОСКОВСКОЙ ОБЛАСТИ

«ФИЗИКО-ТЕХНИЧЕСКИЙ КОЛЛЕДЖ»

**КУРСОВАЯ РАБОТА ПО ПРЕДМЕТУ:**

**«**[**МДК Технологии машинного обучения и анализа данных**](https://schools.school.mosreg.ru/v2/subject?school=2000002775503&class=2390346913078445148&subject=2394909087469797876) **»**

«Прогнозирование суточной нагрузки в электросети на основе анализа временных рядов и сезонных факторов»

Выполнил: Студент 4го курса, группы, ИСП-23В,

Фёдоров Андрей

Долгопрудный 2025

## **СОДЕРЖАНИЕ**

**ВВЕДЕНИЕ**

**1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ**

1.1 Методы прогнозирования временных рядов в энергетике

1.2 Алгоритмы машинного обучения для задач регрессии

1.3 Метрики оценки качества прогнозов

**2 ПРАКТИЧЕСКАЯ ЧАСТЬ**

2.1 Анализ и предобработка исходного датасета

2.1.1 Описание датасета

2.1.2 Обработка пропусков и аномалий

2.1.3 Создание datetime индекса и проверка временной целостности

2.2 Создание признаков (Feature Engineering)

2.2.1 Временные признаки

2.2.2 Циклические кодирования

2.2.3 Пиковые периоды

2.2.4 Лаговые переменные

2.2.5 Скользящие статистики

2.2.6 Признаки из суб-счетчиков

2.3 Разведочный анализ данных (EDA)

2.3.1 Анализ распределения целевой переменной

2.3.2 Суточная сезонность

2.3.3 Недельная сезонность

2.3.4 Месячная сезонность

2.3.5 Анализ автокорреляции

2.4 Построение и обучение моделей машинного обучения

2.4.1 Подготовка данных: разделение на train/test без утечек

2.4.2 Обучение RandomForest, XGBoost, LightGBM

2.4.3 Сравнение метрик: MAE, RMSE, R²

2.4.4 Анализ важности признаков

2.5 Разработка Telegram бота для тестирования модели

2.5.1 Архитектура бота: конфигурация, генератор реалистичных данных

2.5.2 Реализация интерактивных прогнозов

2.5.3 Визуализация результатов и сравнение с реальными данными

2.6 Анализ результатов и систематических ошибок

2.6.1 Выявление проблем: ночное потребление завышено в 3 раза

2.6.2 Анализ причин: доминирование lag\_2h\_ago и циклические ошибки

2.6.3 Предложения по улучшению модели

**ЗАКЛЮЧЕНИЕ**

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

**ПРИЛОЖЕНИЯ**

## ****ВВЕДЕНИЕ****

### ****Актуальность темы****

В современном мире электроэнергия стала таким же важным ресурсом, как вода или воздух. Каждый день мы используем электричество для всего: от зарядки телефона до отопления домов. Но проблема в том, что потребление электроэнергии постоянно меняется - днем мы используем больше, ночью меньше, зимой включаем обогреватели, летом - кондиционеры.

Если энергетические компании не могут точно предсказать, сколько электричества понадобится завтра или через неделю, возникают проблемы. Либо производится слишком много энергии (и она пропадает зря), либо слишком мало (и возникают веерные отключения). Именно поэтому прогнозирование потребления электроэнергии стало такой важной задачей.

С развитием технологий машинного обучения появилась возможность строить точные прогнозы на основе исторических данных. Это как предсказывать погоду, только для энергопотребления.

### ****Цель и задачи исследования****

****Цель работы**** - разработать систему прогнозирования суточного потребления электроэнергии в домашних хозяйствах с использованием методов машинного обучения.

Чтобы достичь этой цели, мне нужно было решить несколько конкретных задач:

1. ****Изучить и подготовить данные**** - взять реальные данные о потреблении электроэнергии и привести их в порядок (убрать пропуски, исправить ошибки)
2. ****Провести анализ данных**** - понять, как меняется потребление в течение суток, недели, месяцев
3. ****Создать признаки для модели**** - преобразовать дату и время в понятные для компьютера параметры
4. ****Обучить несколько моделей машинного обучения**** - попробовать разные алгоритмы и выбрать лучший
5. ****Разработать Telegram-бота**** - сделать простой и понятный интерфейс для тестирования модели
6. ****Проанализировать результаты**** - понять, где модель ошибается и почему

### ****Объект и предмет исследования****

****Объект исследования**** - процесс потребления электроэнергии в домашних хозяйствах.

****Предмет исследования**** - временные ряды потребления электроэнергии и их сезонные закономерности.

### ****Практическая значимость****

Результаты моей работы могут быть полезны:

* ****Энергетическим компаниям**** - для более точного планирования генерации электроэнергии
* ****Потребителям**** - для понимания своих паттернов потребления и экономии
* ****Студентам**** - как пример реального применения машинного обучения
* ****Разработчикам**** - готовое решение для создания подобных систем

### ****Структура работы****

Моя курсовая работа состоит из двух основных частей:

****Теоретическая часть**** - где я изучил, какие вообще бывают методы прогнозирования и как оценивать их качество.

****Практическая часть**** - где я:

* Взял реальные данные за 6 месяцев (более 260 тысяч записей)
* Преобразовал их в удобный формат
* Нашел интересные закономерности (оказывается, в выходные потребляют на 36% больше)
* Обучил несколько моделей машинного обучения
* Сделал Telegram-бота для демонстрации работы модели
* Обнаружил и проанализировал проблемы в работе модели

В конце я подвожу итоги и делаю выводы о том, что получилось, а что можно улучшить.

## ****1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ****

### ****1.1 Методы прогнозирования временных рядов в энергетике****

****Что такое временные ряды?****  
Временной ряд — это последовательность данных, собранных в последовательные моменты времени, где порядок наблюдений имеет важное значение. В отличие от обычных данных, где порядок не важен, во временных рядах каждая следующая точка зависит от предыдущих.

****Ключевые особенности временных рядов:****

* ****Зависимость от времени**** - текущее значение зависит от предыдущих
* ****Порядок важен**** - нельзя перемешивать данные
* ****Наличие трендов и сезонности**** - данные имеют закономерности во времени

****В моей работе**** я работаю с временным рядом потребления электроэнергии, где каждое измерение сделано с интервалом в 1 минуту, и потребление в текущую минуту зависит от потребления в предыдущие минуты.

****Особенности энергетических данных****  
Данные по энергопотреблению имеют несколько интересных особенностей, которые важно учитывать:

1. ****Сезонность**** - потребление повторяется по определенным циклам:
   * ****Суточная****: ночью мало, утром и вечером много
   * ****Недельная****: в выходные потребление отличается от рабочих дней
   * ****Годовая****: зимой и летом потребление разное
2. ****Тренд**** - общее направление изменения over времени. Например, потребление может постепенно расти из-за появления новых электроприборов.
3. ****Случайные колебания**** - непредсказуемые изменения из-за разных факторов (праздники, аварии, погода)

****Основные методы прогнозирования****

Классические методы:

* ****ARIMA**** - сложный статистический метод, который хорошо работает с сезонностью
* ****Экспоненциальное сглаживание**** - простой метод, где новые данные имеют больший вес

Современные методы:

* ****Машинное обучение**** - алгоритмы, которые сами учатся на данных находить закономерности
* ****Глубокое обучение**** - сложные нейронные сети для очень точных прогнозов

В своей работе я выбрал машинное обучение, потому что эти методы хорошо справляются с сложными закономерностями в данных и их проще реализовать.

****Почему это важно для энергетики?****  
Точные прогнозы помогают:

* Избежать перепроизводства или нехватки электроэнергии
* Снизить costs на обслуживание сетей
* Предотвратить аварии и веерные отключения
* Оптимизировать работу электростанций

### ****1.2 Алгоритмы машинного обучения для задач регрессии****

****Что такое задача регрессии?****  
В машинном обучении есть два основных типа задач:

* ****Классификация**** - когда мы пытаемся определить категорию (например, "кошка" или "собака")
* ****Регрессия**** - когда мы пытаемся предсказать число (например, "сколько электроэнергии будет потребляться завтра в 18:00")

Моя задача - это регрессия, потому что я предсказываю конкретное число (киловатты).

****Алгоритмы, которые я использовал:****

****1. Random Forest****  
Представьте, что вам нужно угадать вес человека. Вы можете спросить нескольких людей и взять среднее значение их предположений. Random Forest работает похожим образом - он создает много "деревьев решений" (простых моделей), и затем усредняет их прогнозы.

Плюсы:

* Работает хорошо "из коробки", не требует тонкой настройки
* Не переобучается так сильно, как отдельные деревья
* Может работать с разными типами данных

Минусы:

* Медленнее, чем некоторые другие алгоритмы
* Сложнее интерпретировать результаты

****2. XGBoost****  
Это как очень умная версия Random Forest. Если Random Forest создает деревья независимо, то XGBoost каждое следующее дерево учится исправлять ошибки предыдущего.

Плюсы:

* Очень высокая точность
* Быстрая работа
* Хорошо справляется с разными типами данных

Минусы:

* Требует больше настроек
* Может переобучиться на маленьких наборах данных

****3. LightGBM****  
Это еще более продвинутая версия, оптимизированная для скорости и эффективности. Она особенно хороша для больших наборов данных.

Плюсы:

* Очень быстрая
* Экономит память
* Хорошая точность

Минусы:

* Может быть слишком чувствительной к настройкам

****Почему я выбрал именно эти алгоритмы?****  
Я решил попробовать все три, потому что:

1. Они хорошо зарекомендовали себя в задачах прогнозирования
2. Они по-разному работают с данными
3. Я мог сравнить их и выбрать лучший
4. Они не требуют суперкомпьютера для обучения

### ****1.3 Метрики оценки качества прогнозов****

****Зачем нужны метрики?****  
Когда мы создаем модель, нам нужно понять, насколько она хороша. Мы не можем просто посмотреть на прогнозы и сказать "вроде нормально". Нужны конкретные числа, которые покажут качество работы модели.

****Основные метрики, которые я использовал:****

****1. MAE (Mean Absolute Error) - Средняя абсолютная ошибка****  
Это самая простая для понимания метрика. Она показывает, "в среднем на сколько киловатт ошибается наша модель".

Пример: Если реальное потребление 2.0 кВт, а модель предсказала 1.8 кВт, то ошибка = |2.0 - 1.8| = 0.2 кВт

Плюсы: Легко интерпретировать  
Минусы: Не учитывает большие ошибки

****2. RMSE (Root Mean Square Error) - Среднеквадратичная ошибка****  
Эта метрика сильнее штрафует за большие ошибки. Если модель иногда сильно ошибается, RMSE будет высоким.

Пример: Ошибка 2 кВт "весит" в 4 раза больше, чем ошибка 1 кВт

Плюсы: Учитывает серьезность ошибок  
Минусы: Сложнее интерпретировать

****3. R² (R-квадрат) - Коэффициент детерминации****  
Эта метрика показывает, насколько хорошо модель предсказывает данные по сравнению с простым средним значением.

Как интерпретировать:

* R² = 1.0 - идеальный прогноз
* R² = 0.0 - модель не лучше, чем просто предсказывать среднее значение
* R² < 0 - модель хуже, чем предсказание среднего

Пример: Если R² = 0.9, это значит, что модель объясняет 90% изменчивости данных

Плюсы: Показывает общее качество модели  
Минусы: Не показывает величину ошибок в абсолютных значениях

****Как я использовал эти метрики:****  
В моей работе я сравнивал все три метрики для каждой модели:

* MAE показывал, на сколько киловатт в среднем ошибается модель
* RMSE показывал, есть ли у модели серьезные выбросы ошибок
* R² показывал, насколько модель лучше простого среднего

Это позволило мне комплексно оценить качество каждой модели и выбрать лучшую.

## ****2 ПРАКТИЧЕСКАЯ ЧАСТЬ****

### ****2.1 Анализ и предобработка исходного датасета****

#### ****2.1.1 Описание датасета****

В качестве исходных данных для исследования был использован датасет "Потребление электроэнергии в домашних условиях" (Household Power Consumption), содержащий реальные измерения энергопотребления частного домохозяйства.

****Основные характеристики датасета:****

* ****Период измерений****: 6 месяцев (январь - июнь 2007 года)
* ****Количество записей****: 260 640 измерений
* ****Интервал измерений****: 1 минута
* ****Размер файла****: ~120 МБ

****Структура данных:****  
Датасет содержит следующие столбцы:

* Date - дата измерения (формат ДД/ММ/ГГ)
* Time - время измерения (формат ЧЧ:ММ:СС)
* Global\_active\_power - общая активная мощность (кВт) - ****целевая переменная****
* Global\_reactive\_power - общая реактивная мощность (кВт)
* Voltage - напряжение в сети (В)
* Global\_intensity - общая сила тока (А)
* Sub\_metering\_1 - потребление кухни (кВт)
* Sub\_metering\_2 - потребление прачечной (кВт)
* Sub\_metering\_3 - потребление систем отопления/охлаждения (кВт)

****Особенности данных:****  
Данные собирались с помощью интеллектуальной системы учета, что обеспечило высокую точность и детализацию измерений. Наличие суб-счетчиков позволяет анализировать потребление по отдельным зонам дома.

#### ****2.1.2 Обработка пропусков и аномалий****

****Анализ пропущенных значений:****  
При первичном анализе было выявлено 26 397 пропущенных значений (1.01% от общего объема данных). Распределение пропусков по столбцам:

\*Таблица 2.1 - Статистика пропусков по признакам\*

| **Признак** | **Количество пропусков** | **Процент пропусков** |
| --- | --- | --- |
| Global\_active\_power | 3 771 | 1.45% |
| Global\_reactive\_power | 3 771 | 1.45% |
| Voltage | 3 771 | 1.45% |
| Global\_intensity | 3 771 | 1.45% |
| Sub\_metering\_1 | 3 771 | 1.45% |
| Sub\_metering\_2 | 3 771 | 1.45% |
| Sub\_metering\_3 | 3 771 | 1.45% |

****Методы обработки пропусков:****  
Для сохранения целостности временного ряда была применена стратегия заполнения медианными значениями по каждому столбцу. Такой подход был выбран потому что:

* Медиана устойчива к выбросам
* Сохраняет общее распределение данных
* Проста в реализации и интерпретации

****Обнаружение аномалий:****  
С помощью метода межквартильного размаха (IQR) были выявлены выбросы:

* Global\_active\_power: 14 349 выбросов (5.51%)
* Sub\_metering\_1: 23 199 выбросов (8.90%)
* Global\_intensity: 14 666 выбросов (5.63%)

Выбросы не удалялись, так как они представляют реальные пики потребления и важны для моделирования экстремальных ситуаций.

#### ****2.1.3 Создание datetime индекса и проверка временной целостности****

****Формирование временного индекса:****  
Для корректной работы с временным рядом был создан единый столбец datetime объединением полей Date и Time.

****Проверка временной целостности:****

* ****Период данных****: с 2007-01-01 00:00:00 по 2007-06-30 23:59:00
* ****Длительность****: 181 день
* ****Интервал между измерениями****: 1 минута (постоянный)
* ****Отсутствуют разрывы**** во временном ряде

****Валидация данных:****

* Проверена уникальность временных меток
* Подтверждена последовательность измерений
* Убедились в корректности распознавания дат (учтен параметр dayfirst=True для европейского формата)

****Результат предобработки:****  
После всех этапов обработки получен чистый датасет с 260 640 записями, готовый для дальнейшего анализа и построения признаков. Все пропуски заполнены, временной индекс корректно создан, данные проверены на целостность.

## ****2.2 Создание признаков (Feature Engineering)****

### ****2.2.1 Временные признаки****

Для учета временных закономерностей были созданы следующие признаки:

****Базовые временные характеристики:****

* hour - час дня (0-23)
* day\_of\_week - день недели (0-понедельник, 6-воскресенье)
* month - месяц года (1-12)
* is\_weekend - индикатор выходного дня (0-рабочий, 1-выходной)

****Обоснование выбора:****  
Эти признаки позволяют модели учитывать основные временные циклы, выявленные в ходе EDA анализа:

* Суточные колебания потребления
* Различия между рабочими и выходными днями
* Сезонные изменения по месяцам

### ****2.2.2 Циклические кодирования****

Для корректной работы с циклическими nature временных данных применено преобразование в синусы и косинусы:

****Созданные признаки:****

* hour\_sin, hour\_cos - циклическое кодирование часа
* day\_of\_week\_sin, day\_of\_week\_cos - циклическое кодирование дня недели
* month\_sin, month\_cos - циклическое кодирование месяца

****Преимущество подхода:****

* Устраняет разрыв между 23:00 и 00:00
* Сохраняет временную близость (23:00 близко к 00:00)
* Позволяет модели лучше понимать циклические паттерны

### ****2.2.3 Пиковые периоды****

На основе анализа суточной сезонности выделены ключевые периоды потребления:

****Определенные периоды:****

* is\_morning\_peak - утренний пик (7:00-9:00)
* is\_evening\_peak - вечерний пик (18:00-22:00)
* is\_night - ночное время (0:00-5:00)
* is\_midday - дневное время (10:00-16:00)

****Дополнительные уточняющие признаки:****

* is\_early\_morning - раннее утро (4:00-6:00)
* is\_late\_evening - поздний вечер (21:00-23:00)
* is\_deep\_night - глубокая ночь (1:00-4:00)

### ****2.2.4 Лаговые переменные****

Для учета временной зависимости созданы лаговые признаки на основе анализа автокорреляции:

****Стратегические лаги:****

* lag\_same\_day\_24h - потребление вчера в это же время
* lag\_week\_ago\_168h - потребление неделю назад в это же время
* lag\_48h\_ago - потребление 2 дня назад
* lag\_72h\_ago - потребление 3 дня назад
* lag\_96h\_ago - потребление 4 дня назад

****Взаимодействия лагов с временными периодами:****

* lag\_24h\_morning, lag\_24h\_evening, lag\_24h\_night - лаги с учетом времени суток
* lag\_24h\_weekend - лаги с учетом выходных дней
* lag\_48h\_morning, lag\_48h\_evening - двухдневные лаги по периодам
* lag\_week\_morning, lag\_week\_evening - недельные лаги по периодам

****Обоснование выбора:****  
Лаги выбраны на основе анализа автокорреляции, который показал значимую зависимость до 4 дней и выраженную недельную сезонность.

### ****2.2.5 Скользящие статистики****

Для учета динамики потребления рассчитаны скользящие статистики:

****Долгосрочные статистики:****

* rolling\_mean\_24h - среднее потребление за последние 24 часа
* rolling\_mean\_168h - среднее потребление за последние 7 дней (168 часов)
* rolling\_mean\_7d\_past - среднее потребление за последние 7 дней

****Краткосрочные статистики:****

* rolling\_mean\_3h\_past - среднее потребление за последние 3 часа

****Важное замечание:****  
Все скользящие статистики рассчитаны только на основе предыдущих данных с двойным сдвигом для исключения утечки информации в будущее.

### ****2.2.6 Признаки из суб-счетчиков****

На основе данных суб-счетчиков созданы дополнительные признаки:

****Относительные показатели:****

* kitchen\_ratio - доля потребления кухни от общего
* laundry\_ratio - доля потребления прачечной от общего
* ac\_heating\_ratio - доля потребления систем климат-контроля

****Бинарные индикаторы активности:****

* kitchen\_active - индикатор активности кухни
* laundry\_active - индикатор активности прачечной
* ac\_heating\_active - индикатор активности систем отопления/охлаждения

****Итоговый набор признаков:****  
Всего создано 57 признаков, которые комплексно описывают различные аспекты энергопотребления и позволяют модели выявлять сложные временные зависимости.

## ****2.3 Разведочный анализ данных (EDA)****

### ****2.3.1 Анализ распределения целевой переменной****

****Глобальная активная мощность (Global\_active\_power)**** - основной показатель потребления электроэнергии в домохозяйстве.

****Статистические характеристики:****

* ****Среднее значение****: 1.156 кВт
* ****Медиана****: 0.564 кВт
* ****Стандартное отклонение****: 1.175 кВт
* ****Минимальное значение****: 0.082 кВт
* ****Максимальное значение****: 10.670 кВт

****Особенности распределения:****

* Распределение сильно право-скошено (медиана значительно меньше среднего)
* Большинство значений сосредоточено в диапазоне 0.1-2.0 кВт
* Наличие редких пиков до 10.67 кВт - экстремальные нагрузки
* Коэффициент вариации 101.7% свидетельствует о высокой изменчивости данных

****Интерпретация:****  
Такое распределение типично для данных энергопотребления - длительные периоды низкого потребления чередуются с кратковременными пиковыми нагрузками.

### ****2.3.2 Суточная сезонность****

Анализ выявил четкие суточные паттерны потребления:

****Ключевые точки суточного цикла:****

* ****Минимум****: 0.489 кВт в 4:00 (глубокая ночь)
* ****Максимум****: 2.182 кВт в 21:00 (вечерний пик)
* ****Размах****: 1.692 кВт (в 4.5 раза!)

****Пиковые периоды:****

* ****Утренний пик**** (7:00-9:00): 1.450 кВт
* ****Вечерний пик**** (18:00-22:00): 1.897 кВт
* ****Ночное время**** (0:00-5:00): 0.581 кВт

****Критические переходы:****

* ****Резкий рост****: +0.579 кВт с 6:00 до 7:00 (пробуждение)
* ****Вечерний спад****: -0.586 кВт с 22:00 до 23:00 (подготовка ко сну)

### ****2.3.3 Недельная сезонность****

Потребление существенно различается по дням недели:

****Статистика по дням:****

* ****Минимум****: 0.938 кВт (пятница)
* ****Максимум****: 1.580 кВт (воскресенье)
* ****Разница****: 0.642 кВт (68% увеличения)

****Сравнение типов дней:****

* ****Рабочие дни**** (пн-пт): 1.048 кВт в среднем
* ****Выходные дни**** (сб-вс): 1.432 кВт в среднем
* ****Разница****: +36.6% в выходные

****Стабильность потребления:****

* ****Наиболее стабильный****: воскресенье (коэффициент вариации 90.2%)
* ****Наиболее изменчивый****: вторник (коэффициент вариации 107.2%)

### ****2.3.4 Месячная сезонность****

Выявлены выраженные сезонные колебания:

****Потребление по месяцам:****

* ****Январь****: 1.546 кВт (максимум)
* ****Февраль****: 1.401 кВт
* ****Март****: 1.319 кВт
* ****Апрель****: 0.863 кВт
* ****Май****: 0.986 кВт
* ****Июнь****: 0.827 кВт (минимум)

****Сезонные изменения:****

* ****Зимние месяцы****: 1.474 кВт в среднем
* ****Весенние месяцы****: 1.056 кВт в среднем
* ****Летние месяцы****: 0.827 кВт в среднем
* ****Снижение зима→лето****: -43.9%

### ****2.3.5 Анализ автокорреляции****

****Автокорреляционная функция (ACF):****

* Сильная корреляция с лагом 1 день (+0.4915)
* Выраженная недельная сезонность (лаг 7 дней: +0.4008)
* Значимые корреляции до 14 дней вперед

****Частичная автокорреляционная функция (PACF):****

* Сильная непосредственная зависимость (лаг 1: +0.4942)
* Значимая зависимость на лаге 3 дня (+0.3606)
* Быстрое затухание корреляций

## **2.4 Построение и обучение моделей машинного обучения**

### **2.4.1 Подготовка данных: разделение на train/test без утечек**

Для корректной оценки качества моделей была применена стратегия временного разделения данных, которая исключает утечку информации из будущего в прошлое.

****Метод разделения:****

* ****Обучающая выборка****: первые 80% данных (208 512 записей)
* ****Тестовая выборка****: последние 20% данных (52 128 записей)
* ****Период обучения****: 1 января - 25 мая 2007 года
* ****Период тестирования****: 25 мая - 30 июня 2007 года

****Проверка на утечки:****  
Была проведена тщательная проверка, которая подтвердила:

* Отсутствие перекрытия дат между train и test выборками
* Временной зазор в 1 минуту между окончанием train и началом test
* Нет общих записей между выборками

****Особенность распределения:****  
При анализе обнаружена разница в среднем потреблении между выборками:

* Train выборка: 1.228 кВт (среднее)
* Test выборка: 0.870 кВт (среднее)
* Соотношение: 0.709 (снижение на 29.1%)

Эта разница объясняется сезонным снижением потребления с зимы на лето, что является естественным для энергетических данных.

### **2.4.2 Обучение RandomForest, XGBoost, LightGBM**

Были обучены три различные модели машинного обучения, каждая из которых имеет свои особенности работы с временными рядами.

****Random Forest Regressor:****

* Количество деревьев: 100
* Максимальная глубина: 20
* Использование всех доступных процессоров (n\_jobs=-1)

****XGBoost Regressor:****

* Количество estimators: 100
* Максимальная глубина: 10
* Скорость обучения: 0.1
* Режим ускоренного обучения

****LightGBM Regressor:****

* Количество estimators: 100
* Максимальная глубина: 10
* Скорость обучения: 0.1
* Оптимизирован для работы с большими данными

****Процесс обучения:****  
Все модели обучались на одинаковом наборе данных с 57 признаками. Время обучения варьировалось от 2 минут для LightGBM до 15 минут для RandomForest, что демонстрирует преимущество градиентного бустинга при работе с большими объемами данных.

### **2.4.3 Сравнение метрик: MAE, RMSE, R²**

Для объективной оценки моделей использовались три ключевые метрики качества, которые позволяют комплексно оценить точность прогнозов.

\*Таблица 2.2 - Сравнение метрик качества моделей\*

| **Модель** | **MAE (кВт)** | **RMSE (кВт)** | **R²** | **Интерпретация** |
| --- | --- | --- | --- | --- |
| RandomForest | 0.1473 | 0.3296 | 0.8872 | Хорошее качество |
| XGBoost | 0.1318 | 0.3009 | 0.9060 | Высокое качество |
| ****LightGBM**** | ****0.1259**** | ****0.2916**** | ****0.9117**** | ****Наилучшее качество**** |

****Анализ результатов:****

****LightGBM показала наилучшие результаты:****

* Средняя ошибка прогноза: 125.9 Вт
* Объясненная дисперсия: 91.17%
* Максимальная ошибка: 4.44 кВт

****Сравнительная характеристика:****

* LightGBM превосходит RandomForest на 14.5% по MAE
* LightGBM превосходит XGBoost на 4.5% по MAE
* Все модели демонстрируют высокое качество (R² > 0.88)

****Практическая интерпретация:****  
Для энергетической компании ошибка в 126 Вт на одно домохозяйство является приемлемой, особенно учитывая высокую изменчивость бытового потребления. При масштабировании на 1000 домов совокупная ошибка составит около 126 кВт, что соответствует потреблению небольшого офисного здания.

### **2.4.4 Анализ важности признаков**

Для понимания работы модели и выявления наиболее значимых факторов потребления был проведен анализ важности признаков для лучшей модели - LightGBM.

\*Таблица 2.3 - Топ-10 наиболее важных признаков\*

| **Место** | **Признак** | **Важность (%)** | **Интерпретация** |
| --- | --- | --- | --- |
| 1 | rolling\_mean\_3h\_past | 18.77% | Краткосрочная история |
| 2 | kitchen\_ratio | 12.83% | Активность кухни |
| 3 | ac\_heating\_ratio | 11.37% | Системы климат-контроля |
| 4 | laundry\_ratio | 7.77% | Активность прачечной |
| 5 | rolling\_mean\_24h | 7.53% | Суточное среднее |
| 6 | rolling\_mean\_168h | 4.40% | Недельное среднее |
| 7 | lag\_same\_day\_24h | 3.20% | Вчера в это же время |
| 8 | hour\_sin | 3.07% | Циклический час |
| 9 | hour\_cos | 3.03% | Циклический час |
| 10 | month\_cos | 2.93% | Циклический месяц |

****Ключевые выводы из анализа важности:****

****Доминирование временных паттернов:****

* Признаки, связанные с временными закономерностями, составляют 45% общей важности
* Скользящие статистики и лаги - наиболее информативные признаки
* Циклические кодирования времени доказали свою эффективность

****Роль суб-счетчиков:****

* Данные суб-счетчиков collectively составляют 32% важности
* Активность кухни - второй по значимости фактор
* Системы отопления/охлаждения - третий по значимости

****Сбалансированность модели:****  
Распределение важности признаков является сбалансированным - ни один признак не доминирует чрезмерно (максимальная важность 18.77%), что свидетельствует о корректной работе модели и отсутствии переобучения на отдельные факторы.

****Практическая значимость:****  
Анализ важности признаков подтверждает правильность выбранного подхода к feature engineering - созданные временные и циклические признаки действительно несут важную информацию для прогнозирования потребления.

## **2.5 Разработка Telegram бота для тестирования модели**

### **2.5.1 Архитектура бота: конфигурация, генератор реалистичных данных**

Для практического тестирования работы модели был разработан Telegram бот, который предоставляет интерактивный интерфейс для получения прогнозов энергопотребления.

****Архитектура бота:****

* ****Фреймворк****: python-telegram-bot
* ****Модель****: LightGBM (наилучшая модель)
* ****Признаки****: 57 предварительно созданных признаков
* ****Конфигурация****: отдельный config.py файл с настройками

****Генератор реалистичных данных:****  
Для обеспечения работы бота в условиях отсутствия реальных текущих данных был разработан генератор реалистичных признаков, основанный на статистике из EDA анализа:

****Основные компоненты генератора:****

* ****Реальные суточные профили****: использование фактических средних значений потребления по часам из обучающей выборки
* ****Сезонные коэффициенты****: учет месячных колебаний (зима +27.5%, лето -28.5%)
* ****Недельные паттерны****: различие рабочих и выходных дней (+36.6%)
* ****Реалистичные лаги****: генерация исторических данных на основе суточных профилей

****Реалистичные суточные профили:****  
Генератор использует фактические средние значения потребления по часам:

* Ночной минимум: 0.489 кВт в 4:00
* Утренний пик: 1.518 кВт в 7:00
* Вечерний максимум: 2.182 кВт в 21:00

****Преимущества подхода:****

* Сохранение реальных временных паттернов
* Учет сезонных и недельных колебаний
* Возможность тестирования на различных датах
* Реалистичное взаимодействие с моделью

### **2.5.2 Реализация интерактивных прогнозов**

Бот предоставляет пользователю три основных сценария взаимодействия через инлайн-кнопки:

****Сценарии использования:****

1. ****Прогноз на завтра**** - демонстрация работы модели на 1 день вперед
2. ****Прогноз на послезавтра**** - тестирование на 2 дня вперед
3. ****Сравнение обоих прогнозов**** - анализ различий между последовательными днями

****Техническая реализация:****

* ****Асинхронная обработка**** запросов для обеспечения отзывчивости
* ****Визуализация результатов**** в виде графиков
* ****Детальная статистика**** по каждому прогнозу
* ****Честная оценка**** качества с указанием систематических ошибок

****Пример работы прогноза:****  
При запросе прогноза на конкретную дату бот:

1. Генерирует реалистичные признаки для каждого часа суток
2. Передает их в обученную модель LightGBM
3. Получает прогнозы потребления для 24 часов
4. Визуализирует результаты в виде графика
5. Предоставляет текстовую аналитику

### **2.5.3 Визуализация результатов и сравнение с реальными данными**

Для наглядного представления результатов были разработаны специализированные типы визуализаций:

****Типы графиков:****

1. ****Суточный профиль прогноза**** - потребление по часам для одной даты
2. ****Сравнительный анализ**** - два прогноза на разные даты на одном графике
3. ****Сопоставление с реальными данными**** - прогноз против исторических средних

****Элементы визуализации:****

* ****Цветовые зоны****: выделение пиковых периодов (утро, вечер, ночь)
* ****Сравнительные линии****: реальные средние значения из EDA анализа
* ****Аннотации****: пояснения к критическим точкам графика
* ****Легенда****: объяснение элементов визуализации

****Пример сравнительного анализа:****  
На графике сравнения отображаются:

* Синяя линия: прогноз на завтра
* Красная линия: прогноз на послезавтра
* Зеленая пунктирная: реальные средние значения
* Цветовые зоны: ночное время, утренний и вечерний пики

****Аналитическая информация:****  
К каждому графику прилагается текстовая аналитика:

* Среднее потребление за сутки
* Пиковая нагрузка и время ее достижения
* Сравнение с реальными историческими данными
* Выявление систематических ошибок модели

****Объективность представления:****  
Важной особенностью бота является честное представление результатов - пользователю явно указываются на проблемы модели и расхождения с реальными данными, что делает тестирование максимально объективным.

## **2.6 Анализ результатов и систематических ошибок**

### **2.6.1 Выявление проблем: ночное потребление завышено в 2-3 раза**

В ходе тестирования модели через Telegram бот были выявлены систематические ошибки в прогнозах:

****Ключевые проблемы:****

1. ****Завышение ночного потребления****
   * Реальное ночное потребление: 0.5-0.8 кВт
   * Прогнозируемое: 2.5-4.5 кВт
   * Превышение: в 3-5 раз
2. ****Недостаточное относительное увеличение утренних пиков**** 
   * Реальный утренний пик: 1.5-1.8 кВт
   * Прогнозируемый: 3.5-4.5 кВт
   * Хотя абсолютное значение выше, относительное увеличение меньше ожидаемого
3. ****Общий сдвиг масштаба****
   * Все прогнозы систематически завышены на 2-3 кВт
   * Сохранение правильной формы суточного профиля

****Пример конкретных расхождений:****

* 4:00: реально 0.49 кВт → прогноз 2.33 кВт (+375%)
* 7:00: реально 1.52 кВт → прогноз 3.77 кВт (+148%)
* 21:00: реально 2.18 кВт → прогноз 5.70 кВт (+161%)

### **2.6.2 Анализ причин: доминирование rolling\_mean\_3h\_past и циклические ошибки**

Анализ важности признаков и работы модели позволил выявить основные причины систематических ошибок:

****Основные причины проблем:****

1. ****Доминирование краткосрочных статистик****
   * Признак rolling\_mean\_3h\_past имеет 18.77% важности
   * Модель слишком сильно ориентируется на ближайшую историю
   * Недостаточный учет долгосрочных сезонных паттернов
2. ****Проблемы с генерацией реалистичных лагов****
   * Лаговые переменные генерируются на основе усредненных данных
   * Не учитываются индивидуальные особенности конкретных дней
   * Накопление ошибок при генерации цепочки лагов
3. ****Смещение данных в обучающей выборке****
   * Train выборка содержит больше зимних данных (высокое потребление)
   * Test выборка и прогнозы - летние данные (низкое потребление)
   * Модель не полностью адаптируется к сезонным изменениям
4. ****Ограничения генератора признаков****
   * Использование усредненных значений вместо реальных данных
   * Недостаточный учет случайных колебаний потребления
   * Упрощенная модель взаимодействия признаков

****Технические аспекты:****

* Модель сохраняет правильную форму суточного профиля
* Все пики и провалы находятся в правильных временных интервалах
* Относительные соотношения между часами сохраняются
* Проблема в абсолютном масштабе, а не в паттернах

### **2.6.3 Предложения по улучшению модели**

На основе анализа выявленных проблем были сформулированы конкретные предложения по улучшению точности прогнозов:

****Краткосрочные улучшения:****

1. ****Коррекция масштаба прогнозов****
   * Внедрение пост-обработки прогнозов
   * Применение сезонных поправочных коэффициентов
   * Использование скользящей нормализации
2. ****Улучшение генератора признаков****
   * Более точная генерация лаговых переменных
   * Учет реальной волатильности потребления
   * Добавление случайных колебаний в допустимых пределах
3. ****Балансировка важности признаков****
   * Ограничение влияния доминирующих признаков
   * Увеличение веса циклических и сезонных признаков
   * Применение регуляризации для сбалансированного обучения

****Долгосрочные улучшения:****

1. ****Сбор актуальных данных****
   * Интеграция с реальными системами учета
   * Использование текущих данных вместо генерации
   * Постоянное обновление модели на новых данных
2. ****Усовершенствование архитектуры модели****
   * Эксперименты с нейронными сетями (LSTM)
   * Использование ансамблевых методов
   * Применение специализированных подходов для временных рядов
3. ****Расширение набора признаков****
   * Добавление погодных данных
   * Учет календарных особенностей (праздники)
   * Интеграция с внешними факторами влияния

****Практическая реализуемость:****  
Большинство предложенных улучшений могут быть реализованы в рамках существующей архитектуры без кардинальной переработки системы, что делает их практичными для внедрения в краткосрочной перспективе.

## **ЗАКЛЮЧЕНИЕ**

### **Обобщение результатов исследования**

В ходе выполнения курсовой работы была успешно решена поставленная задача по разработке системы прогнозирования суточного потребления электроэнергии в домашних хозяйствах. Проведенное исследование позволило достичь следующих ключевых результатов:

****Основные достижения:****

1. ****Создан полный ML-пайплайн**** от сбора и предобработки данных до развертывания работающего прототипа в виде Telegram бота
2. ****Проведен комплексный EDA анализ****, который выявил четкие временные закономерности:
   * Суточные колебания потребления с размахом 1.692 кВт
   * Недельная сезонность (+36.6% в выходные дни)
   * Месячные изменения (-43.9% от зимы к лету)
   * Выраженная автокорреляция до 7 дней
3. ****Разработана эффективная система признаков**** из 57 параметров, включающая:
   * Циклическое кодирование времени
   * Лаговые переменные на основе ACF анализа
   * Скользящие статистики без утечек данных
   * Признаки из суб-счетчиков потребления
4. ****Обучены и сравнены три модели**** машинного обучения, где LightGBM показала наилучшие результаты:
   * MAE: 0.126 кВт (126 Вт средней ошибки)
   * R²: 0.9117 (91.17% объясненной дисперсии)
   * Устойчивое превосходство над другими алгоритмами
5. ****Разработан интерактивный Telegram бот**** для тестирования модели, который:
   * Предоставляет интуитивный интерфейс для прогнозов
   * Визуализирует результаты с сравнением реальных данных
   * Честно демонстрирует limitations модели

### **Практическая значимость и перспективы развития**

****Реальный вклад работы:****

Несмотря на выявленные систематические ошибки, разработанная система имеет практическую ценность:

1. ****Для образовательных целей**** - работа представляет полноценный пример применения ML для прогнозирования временных рядов
2. ****Для энергетических компаний**** - продемонстрирован подход к созданию систем прогнозирования бытового потребления
3. ****Для дальнейших исследований**** - выявленные проблемы и предложенные решения создают основу для совершенствования модели

****Перспективы улучшения:****

На основе анализа ошибок определены четкие направления для развития:

1. ****Непосредственные улучшения****:
   * Внедрение постобработки для коррекции масштаба
   * Балансировка важности признаков
   * Улучшение генератора реалистичных данных
2. ****Стратегические направления****:
   * Интеграция с реальными системами учета
   * Добавление внешних факторов (погода, праздники)
   * Эксперименты с архитектурами LSTM
3. ****Масштабирование системы****:
   * Адаптация для многоквартирных домов
   * Прогнозирование для различных типов потребителей
   * Создание системы раннего предупреждения пиковых нагрузок

### **Критическая оценка limitations**

Важно объективно оценить limitations текущего решения:

****Технические ограничения:****

* Систематическое завышение прогнозов в 2-3 раза
* Зависимость от качества генерации признаков
* Ограниченность обучающих данных (только 6 месяцев)

****Методологические ограничения:****

* Использование усредненных данных вместо реальных измерений
* Отсутствие учета внешних факторов влияния
* Упрощенная модель потребительского поведения

****Несмотря на эти ограничения****, работа демонстрирует работоспособный подход к решению сложной задачи прогнозирования энергопотребления и создает прочную основу для дальнейших улучшений.

### **Итоговые выводы**

1. ****Гипотеза исследования подтверждена**** - методы машинного обучения эффективны для прогнозирования энергопотребления с учетом временных закономерностей
2. ****Достигнута основная цель**** - разработана система прогнозирования с точностью 91.2% по коэффициенту детерминации
3. ****Практическая реализуемость доказана**** - создан работающий прототип в виде Telegram бота
4. ****Выявлены направления совершенствования**** - систематические ошибки модели имеют четкие причины и пути решения

Работа демонстрирует, что даже в условиях ограниченных данных и ресурсов можно создать эффективную систему прогнозирования, которая служит надежной основой для дальнейшего развития и совершенствования.

## **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. ****Hastie, T., Tibshirani, R., & Friedman, J.**** (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – 745 с. – Основы машинного обучения и статистического анализа
2. ****Hyndman, R. J., & Athanasopoulos, G.**** (2021). Forecasting: Principles and Practice. – 438 с. – Методы прогнозирования временных рядов
3. ****Géron, A.**** (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. – 851 с. – Практическое применение машинного обучения
4. ****Brownlee, J.**** (2018). Deep Learning for Time Series Forecasting. – 562 с. – Специализированные методы для временных рядов
5. ****Документация scikit-learn**** (2023). Official documentation for scikit-learn machine learning library. – Реализация алгоритмов ML
6. ****Документация LightGBM**** (2023). Microsoft LightGBM official documentation. – Градиентный бустинг для больших данных
7. ****UCI Machine Learning Repository**** (2023). Individual household electric power consumption dataset. – Исходные данные для исследования
8. ****Telegram Bot API Documentation**** (2023). Official documentation for Telegram Bot API. – Разработка интерфейса бота
9. ****Matplotlib Documentation**** (2023). Python plotting library documentation. – Визуализация результатов
10. ****Pandas Documentation**** (2023). Data analysis and manipulation tool for Python. – Обработка и анализ данных

## **ПРИЛОЖЕНИЯ**