МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ  
Федеральное государственное автономное образовательное учреждение высшего образования  
«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»  
ШКОЛА КОМПЬЮТЕРНЫХ НАУК

ОТЧЕТ  
О РАЗРАБОТКЕ WEB-ПРИЛОЖЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ЧИСЛА ПАЦИЕНТОВ ПЕРВИЧНОГО ПРИЕМА В МЕДУЧРЕЖДЕНИЯХ ТЮМЕНИ С УЧЕТОМ КЛИМАТИЧЕСКИХ ДАННЫХ

ФГАОУ ВО «ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»  
(наименование организации)

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ НАУК  
(наименование структурного подразделения)

| Выполнил  обучающийся 4 курса,  2021 года поступления | (подпись) | Ершов Александр  Андреевич |
| --- | --- | --- |
| Выполнил обучающийся 4 курса,  2021 года поступления | (подпись) | Земсков Никита Александрович |
| Научный руководитель проекта,  (должность) | (подпись) | Ступников Андрей Анатольевич |
|  |  |  |

ОГЛАВЛЕНИЕ

[ВВЕДЕНИЕ 3](#_heading=h.5pi442w2jtnh)

[ГЛАВА 1. МЕТОДЫ И МОДЕЛИ ПРОГНОЗИРОВАНИЯ 5](#_heading=h.30j0zll)

[1.1 Предметная область 5](#_heading=h.s2kz8ok678sc)

[1.2 Модель градиентного бустинга 9](#_heading=h.kpnmvn90rakr)

[1.3 Модель XGBoost 12](#_heading=h.yn758s3rvggc)

[1.4 Модель LightGBM 13](#_heading=h.fi82oua6r06y)

[1.5 Модель CatBoost 14](#_heading=h.vtdijdz5fxoy)

[1.7 Модель ARIMA 16](#_heading=h.mhfa2td4rmr9)

[1.8 Нейронная сеть LSTM 16](#_heading=h.61iuldjamcrh)

[1.9 Линейная регрессия 19](#_heading=h.f0kc15jpajc6)

[ГЛАВА 2. РЕАЛИЗАЦИЯ 23](#_heading=h.3dy6vkm)

[2.1 Постановка задачи 23](#_heading=h.z337ya)

[2.2 Подбор стека технологий 23](#_heading=h.k2icv52s8sib)

[2.3 Реализация моделей 27](#_heading=h.bhlprb85un8r)

[2.4 Разработка серверной части приложения 30](#_heading=h.er1y7rgroutm)

[2.5 Интерфейс пользователя 31](#_heading=h.16ai35jv5st6)

[2.6 Функционал системы 33](#_heading=h.dkneyaa070tv)

[2.7 Архитектура веб-приложения 33](#_heading=h.8umi4hlxx8dr)

[ГЛАВА 3. СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ И ОЦЕНКА ИХ КАЧЕСТВА 35](#_heading=h.2jxsxqh)

[3.1 Модель Gradient Boosting 35](#_heading=h.rvdch5vn7gzn)

[3.2 Модель LightGBM 40](#_heading=h.adr9m6ca5tv3)

[3.3 Модель CatBoost 45](#_heading=h.4666ds83py3o)

[3.4 Модель XGBoost 49](#_heading=h.e2pchc2sycvx)

[3.5 Модель LSTM 54](#_heading=h.21m5d26whkkb)

[3.6 Модель ARIMA 58](#_heading=h.agk4pg3y69gs)

[ГЛАВА 4. ТЕСТИРОВАНИЕ 67](#_heading=h.1y810tw)

[4.1 Тестирование модели 67](#_heading=h.p3iwxet5ab9k)

[4.2 Тестирования пользовательского интерфейса 69](#_heading=h.9quk2gnp9yh7)

[4.3 Процедура тестирования серверной части 69](#_heading=h.9quk2gnp9yh7)

[4.4 Результаты тестирования 70](#_heading=h.nkfmbx7kvwi7)

[ЗАКЛЮЧЕНИЕ 71](#_heading=h.4i7ojhp)

[СПИСОК ЛИТЕРАТУРЫ 72](#_heading=h.2xcytpi)

#### 

#### 

#### ВВЕДЕНИЕ

За последние годы в России выявлена тенденция увеличения нагрузки на медучреждения, в частности в учреждения первичной медико-санитарной помощи. Повышенная нагрузка связана как со вспышкой вируса COVID-19, так и с повышением общего числа заболеваний населения. По данным министерства здравоохранения начиная с 2020 года количество визитов в амбулаторно-поликлинические учреждения составляет в пределах 1,4 млрд, в зависимости от года [1]. Значительное увеличение нагрузки на медучреждения требует быстрее оказывать медицинскую помощь, что потенциально, ведет к снижению качества оказываемых медицинских услуг.

Согласно данным Федеральной службы государственной статистики [2], в 2022 году количество обращений за медицинской помощью в Тюменской области возросло на 15% по сравнению с предыдущим годом. Если рассматривать общую статистику зарегистрированных заболеваний у пациентов с диагнозом, установленным впервые в жизни, можно заметить насколько сильно возросло число заболеваний [2].

За 2020 год всего 3139 тысяч человек, 2021 год – 3526 тысяч человек, 2022 год – 3792 тысяч человек, 2023 год – 3448 тысяч человек [2]. Несмотря на увеличение числа обращений, медицинские учреждения сталкиваются с нехваткой человеческих ресурсов.

Это приводит к проблеме снижения качества оказываемых медицинских услуг, из-за повышенной нагрузки на медучреждения. Также не в последнюю очередь это может быть связано с изменением климатических условий в Тюменской области и по самой Тюмени. Таким образом, существует необходимость прогнозирования числа поступающих пациентов на ближайшие дни, что позволит медучреждениям заранее подготовиться к повышенным нагрузкам.

Технические решения, направленные на прогнозирование числа пациентов, могут существенно улучшить управление ресурсами в медицинских учреждениях. Использование математических моделей и алгоритмов машинного обучения для анализа данных о пациентах и климатических условиях позволит предсказывать количество обращений в ближайшие дни. Это, в свою очередь, даст возможность медицинским учреждениям заранее планировать распределение персонала и ресурсов.

В Тюмени, где климат характеризуется холодными зимами и теплым летом, резкие колебания температуры могут способствовать обострению хронических заболеваний, такие как сердечно-сосудистые патологии. Вероятно, существует зависимость между изменениями температуры и количеством обращений в поликлиники. Это отмечает необходимость учета климатических факторов при прогнозировании числа пациентов.

Целью данного проекта является разработка web-приложения, которое будет являться инструментом для прогнозирования числа пациентов первичного приема на основе климатических данных. В свою очередь, актуальность проекта очевидна, поскольку прогнозирование числа поступающих пациентов позволит медицинским учреждениям заранее подготовиться к повышенным нагрузкам, тем самым сохранить качество обслуживания пациентов.

#### ГЛАВА 1. МЕТОДЫ И МОДЕЛИ ПРОГНОЗИРОВАНИЯ

##### 1.1 Предметная область

Предметная область, которая была отражена во введении проекта - исследование влияния климатических условий на изменение состояния здоровья людей.

Для более подробного понимания того, что такое климатические условия, следует дать описание термину:

Климатические условия — это режим погоды, характерный для данной местности в зависимости от ее географического положения. Погода включает в себя такие параметры, как температура, влажность, осадки, ветер и другие метеорологические факторы. Метеорологические факторы — это элементы атмосферы, которые влияют на погоду и климат. Основные метеорологические факторы включают:

1. Температура — измеряет теплоту воздуха и влияет на состояние окружающей среды.
2. Влажность — содержание водяного пара в воздухе, которое влияет на ощущение тепла и формирование облаков.
3. Атмосферное давление — сила, с которой воздух давит на поверхность Земли, играет важную роль в формировании погодных систем.
4. Осадки — любые формы воды, выпадающие из атмосферы (дождь, снег, град), которые влияют на уровень влаги в почве и экосистемы.
5. Ветер — движение воздуха, которое влияет на распределение тепла и влаги, а также на формирование облаков и осадков.
6. Облачность — количество и тип облаков в атмосфере, которые влияют на солнечное излучение и температуру.

Далее приведены исследования, посвященные влиянию климатических условий на здоровье людей и увеличением числа посещений медучреждений.

Исследования показывают, что климатические условия оказывают значительное влияние на здоровье населения. В научной статье [3] исследование посвящено влиянию экстремальных температур (как холода, так и жары) на посещаемость отделений неотложной помощи по респираторным заболеваниям в Пекине, Китай. В ней исследуются эффекты экстремальных температур на здоровье, с акцентом на респираторные заболевания, в условиях изменения климата. Включены данные о ежедневных посещениях больниц за 2009-2012 годы в Пекине, а также информация о температуре, влажности, загрязнении воздуха (, ). Анализируются данные об обращениях в больницы по поводу респираторных заболеваний, включая возраст, пол пациентов, и погодные условия. Использована модель распределенного лагового нелинейного анализа (DLNM) для оценки задержанного эффекта температуры. Модели показывают нелинейную связь между температурой и визитами в отделения. По итогу данного исследования, было выявлено, что экстремально низкие температуры более длительно влияют на здоровье, особенно на детей, а пожилые люди более уязвимы к тепловым волнам. Женщины реагируют на холод более остро, чем мужчины.

В научной статье [4] отображена тема исследования влияния загрязнение воздуха на частоту посещений отделений неотложной помощи по сердечно-сосудистым заболеваниям в городах Онтарио, Канада. В проблеме темы поднимаются вопросы загрязнение воздуха, которые остаются актуальными для системы здравоохранения, особенно с увеличением роста сердечно-сосудистых и респираторных заболеваний. В исследовании использовались данные о здоровье из Национальной системы отчетности по амбулаторной помощи, а также данные о загрязнении воздуха и метеорологические данные, собранные с апреля 2004 года по декабрь 2011 года. Для более подробного понимания стоит раскрыть термин амбулаторной помощи. Амбулаторная помощь - это внебольничная медицинская помощь, оказываемая людям, которые приходят на прием к врачу в поликлинику. В исследовании, был применен метод временной регрессии с учетом временных задержек, а также различие между полами людей, для выявления зависимости между медицинскими обращениями и загрязнением воздуха. Излагая суть результатов исследования, было определено, что краткосрочное воздействие загрязнение воздуха связано с увеличением числа обращений в отделения неотложной помощи по респираторным заболеваниям. Были также обнаружены положительные зависимости для признаков и , с временными задержками как для мужчин, так и для женщин. Стоит пояснить данные формулы, - это диоксид азота, он является одним из загрязняющих веществ, образующихся в результате сжигания ископаемого топлива, который вызывает раздражение дыхательных путей в организме человека. Далее, - это мелкие частицы с диаметров 2.5 микрометра, они представляют собой смесь различных твердых частиц, а также капель жидкости, которые находятся в воздухе. Они образуются в результате выбросов промышленных предприятий, эти частицы могут проникать в легкие, тем самым вызывать проблемы со здоровьем, такие как сердечно-сосудистые заболевания. Последним является - диоксид серы, это газ без цвета и с резким запахом, который образуется в результате сжигания ископаемого топлива, содержащего серу, таких как уголь и нефть. Он может вызывать проблемы в организме человека, связанные с дыхательной системой, усугубить заболевание легких. Результаты исследования говорят о том, что требуется дальнейшее изучение влияния загрязнение воздуха на здоровье человека и определение иных зависимостей.

В научной статье [5] также исследовалась тема изменения климата и его влияние на здоровье человека, включая инфекционные заболевания. Основная проблема, которая была отмечена - это недостаточная доказательная база о влиянии изменения климата на заболевания, в частности экстремальных погодных условий, таких как тепловые волны, поскольку возникают сложности в интерпретации данных о здоровье, поскольку они могут быть обусловлены множеством факторов, таких как экономические факторы и социальные. Для анализа использовались исторические события, такие как тепловая волна, возникшая в Европе в 2003 году. Применялись модели прогнозирования здоровья, а также проводилась оценка рисков на основе климатических изменений и их воздействия на здоровье. В результате исследования было показано, что изменения климата приводит к серьезным последствиям для здоровье людей. Более подробно было отмечено, что в результате тепловой волны в Европе в 2003 году, было зафиксировано более 2000 избыточных смертей в Англии. Во Франции количество смертей составило 14800, что равно увеличению смертности на 60% в течении трех недель августа того года. Только в Париже число смертей увеличилось на 140% в этот период. Все эти результаты говорят о том, насколько серьезное влияние оказывают экстремальные температуры на здоровье людей.

Аналогично, в статье [6] рассматривается влияние климатических условий на здоровье население, однако исследование было по большей степени посвящено оценке уязвимости человеческого здоровья и проблеме участия здравоохранения в изучении климатических процессах, а также способах адаптации населения к изменяющимся климатическим условиям. В исследовании проводится анализ существующих барьеров в проведении оценок, таких как недостаток данных о здоровье и климате, а также отмечается необходимость в более детальном анализе зависимости здоровья человека от климатических условий. В результатах исследования было отмечено, что регулярное обновление оценок критически важно для повышения устойчивости к климатическим изменениям и защиты здоровья населения.

Таким образом, в каждой статье отмечается важность понимания взаимосвязи между климатическими условиями и здоровьем человека, а также отмечается необходимость дальнейших исследований в данной области.

Исходя из проблемы, которая также была отражена во введении, идея проекта заключается в разработке web-приложения, которое будет являться инструментом для прогнозирования количества пациентов первичного приема в медучреждения Тюмени с учетом климатических данных

Основными пользователями проекта являются руководители медучреждений, так как они отвечают за управление и распределение рабочего персонала. Определяют какое количество сотрудников должно выйти на место работы в определенный день.

Задача прогнозирования количества пациентов на будущие дни, с учетом климатических данных, является задачей регрессии. В свою очередь, задача регрессии - это один из основных типов задач в машинном обучении, целью которого является предсказание числовой величины на основе входных данных. Задача регрессии предполагает, что предсказанное значение числовой величины может быть любым в определенном диапазоне. Таким образом, для предсказания количества пациентов, первичного приема было решено рассмотреть модели:

1. Модель градиентного бустинга и ее модификации
2. Модель прогнозирования временных рядов ARIMA
3. Рекуррентная нейронная сеть LSTM
4. Линейная регрессия.

##### 1.2 Модель градиентного бустинга

Градиентный бустинг - это метод ансамблевого обучения, который используется для решения задач регрессии и классификации. Он работает следующим образом:

1. Инициализация: Сначала модель делает начальное предположение. Обычно это просто среднее значение целевой переменной в случае регрессии или же логарифм, в случае классификации.
2. Построение модели на каждом m-шаге, с учетом корректировки ошибок предыдущей модели. Новая модель предсказывает ошибки текущей модели
3. Вычисление градиента функции потерь по предсказаниям текущей модели
4. Обновление модели. Она добавляется к ансамблю с определенной скоростью обучения, которая уменьшает влияние каждой новой модели, чтобы избежать переобучения
5. Повторение процесса: Этот процесс повторяется пока не будет достигнуто определенное количество моделей или же пока ошибка не перестанет значимо уменьшаться.

Градиентный бустинг получил такое название, из-за того, что он использует градиентный спуск для минимизации ошибки модели. Он работает с помощью минимизации функции потерь, используя градиент этой функции или же по другому направление наибольшего убывания. По сути, это позволяет модели пошагово улучшать свои прогнозы, делая фокус на ошибках предыдущих моделей в ансамбле.

Далее, стоит отметить основные преимущества модели градиентного бустинга.

Преимущества:

1. Алгоритм может работать с любыми функциями потерь.
2. Справляется с пропущенными данными
3. Справляется с нелинейными зависимостями и взаимодействиями между признаками.

Третье преимущество являлось ключевым при выборе конечной модели для реализации проекта.

Недостатки:

1. Риск переобучения: модель может переобучиться на обучающих данных, особенно на небольших выборках, таким образом, модель будет склонна к переобучению при слишком большом количестве деревьев.
2. Алгоритм крайне чувствителен к выбросам, они могут повлиять на результаты.
3. Модель склонна к переобучению при слишком большом количестве деревьев.
4. Необходимость тщательной настройки гиперпараметров для достижения наилучших результатов, что по сути, требует много времени и вычислительных ресурсов

Архитектура модели выглядит следующим образом. В качестве базовых моделей, используются деревья решений с глубиной от 3 до 5, которые обучаются на ошибках предыдущих деревьев. Таким образом, архитектура модели представляет собой древовидную структуру, ансамбль деревьев.

Общий вид модели представляется как сумма всех предсказаний деревьев:

Где:

* - начальное предсказание.
* - m-е дерево.
* - коэффициент, который определяет вклад m-го дерева.

Основные гиперпараметры в градиентном бустинге:

1. Количество деревьев: Оно определяет сколько будет деревьев в ансамбле. Если сделать большое количество деревьев, то оно улучшит качество модели, однако повысит риск переобучения.
2. Скорость обучения: Определяет, насколько сильно каждое новое дерево будет корректировать ошибки предыдущего. Чем выше скорость обучения, тем быстрее будет обучаться модель, однако также увеличиться риск переобучения при большом количестве эпох.
3. Количество эпох: Определяет, сколько итераций обучения будет проходить модель на всем обучающем наборе данных.
4. Максимальная глубина деревьев: Ограничение глубины каждого дерева, чем меньше глубина, тем больше шанс предотвращения переобучения, однако может снизиться эффективность модели.
5. Минимальное количество образцов в листе: Этот гиперпараметр определяет количество образцов, которое должно быть в листе, что способствует снижению риска переобучения модели.
6. Тип потерь: Определяет функцию потерь, которую модель будет минимизировать на протяжении всего процесса обучения.

Помимо обычной модели градиентного бустинга, также были рассмотрены ее модификации: XGBoost, LightGBM, CatBoost.

##### 1.3 Модель XGBoost

XGBoost (Extreme Gradient Boosting) - это оптимизированная реализация алгоритма градиентного бустинга, которая имеет более сложную структуру, позволяющей улучшить производительность модели. Она имеет высокую скорость обучения, благодаря параллельной обработке. Вместо последовательного построения деревьев, как это происходит в обычной модели градиентного бустинга, XGBoost разбивает данные на "гистограммы" и строит дерево на основе этих объединенных данных. Стоит отметить, что еще в отличии от обычной модели градиентного бустинга, XGBoost включает регуляризацию L1 (Lasso) и L2 (Ridge), благодаря этому модель избегает переобучения, а также поддерживает обработку пропущенных значений, принимая решение о том, как их обрабатывать еще на этапе построения дерева. Однако, она может быть менее эффективной на небольших наборах данных по сравнению с обычным градиентным бустингом. XGBoost использует градиент второго порядка, то есть, вторую производную для обновления весов.

Функция потерь. XGBoost использует функцию потерь, которая включает в себя регуляризацию:

,

Где:

* - функция потерь
* - фактическое значение
* - предсказанное значение
* - k-тое дерево
* - регуляризационный член k-го древа, который контролирует сложность модели

Регуляризационный член записывается так:

L1 (Lasso):

L2 (Ridge):

XGBoost использует градиенты и гессианы для обновления предсказаний:

, , где - градиенты, - гессианы.

Построение деревьев:

, где S - это текущий узел, а L и R - это левые и правые дочерние узлы.

Итоговая формула модели представляется как сумма предсказаний всех деревьев:

, где — m-ое дерево.

##### 1.4 Модель LightGBM

LightGBM - это фреймворк с открытым исходным кодом для машинного обучения, первоначально разработанный Microsoft. Он известен благодаря своей высокой скоростью обучения, хорошей точностью с параметрами по умолчанию, параллельным обучением, а также малым использованием памяти. LightGBM использует стратегию построения деревьев Leaf-wise (пол листьям), которая выбирает наибольший прирост информации на каждом шаге обучения, что позволяет эффективно уменьшать ошибку. Это отличается от обычного метода, где деревья строятся по принципу (level-wise), то есть по уровням.  
 При построении дерева, модель выбирает лист с максимальным приростом, что описывается в виде формулы:

, где - градиент, а - гессиана.

LightGBM применяет такую же регуляризацию как и модель XGBoost для предотвращения переобучения, L1 и L2. Стоит сказать, что LightGBM использует метод гистограмм для объединения данных. Данные разбиваются на фиксированное количество бинов, где бин - это способ группировки непрерывных значений в дробные категории, что уменьшает объем информации, которая будет обрабатываться.

По каждому признаку x, строится гистограмма:

, где - ширина бина.

Итоговая формула представляется как сумма предсказаний всех деревьев:

, где — m-ое дерево.

##### 1.5 Модель CatBoost

CatBoost - это алгоритм машинного обучения, разработанный компанией «Яндекс». Он предназначен для работы с табличными данными и эффективен для работы с категориальными данными. Вместо обычного градиентного бустинга, CatBoost использует метод ordered boosting, который создает предсказания на основе предыдущих данных, что минимизирует влияние утечки информации. Этот алгоритм считается универсальным, так как хорошо работает как с категориальными, так и с числовыми данными. Говоря более подробно о методе ordered boosting, стоит сказать, что он создает специальные статистики для категориальных признаков, используя кросс-валидацию для вычисления значений.  
 Для некоторого категориального признака , со значениями вычисляется статистика:

, учитывая условие, что =, где - это количество примеров, а - соответственно целевая переменная.

Аналогично моделям LightGBM и XGBoost, алгоритм CatBoost применяет регуляризацию, которая включает L2 (Ridge), для предотвращения переобучения.

Итоговая формула представляется как сумма предсказаний всех деревьев:

, где — m-ое дерево.

1.6 Настройка гиперпараметров

Для настройки гиперпараметров моделей градиентного бустинга применялась К-блочная кросс-валидация.  
 Кросс-валидация - это метод оценки производительности модели машинного обучения, при котором данные делятся на несколько несколько подмножеств, которые называются фолдами, и модель обучается и тестируется несколько раз на разных фолдах, в зависимости от их количества.  
 Кросс-валидация позволяет выявить переобучение модели, а также определить наилучшие настройки гиперпараметров для модели.  
 К-блочная кросс валидация работает следующим образом:  
 Датасет делится на K равных блоков, один из которых используется для валидации, а все остальные используются для обучения. Так процесс повторяется K раз, меняя блок валидации. В свою очередь, валидационная выборка показывает как ведет себя модель с разными данными.

##### 1.7 Модель ARIMA

ARIMA (АвтоРегрессионная Интегрированная Скользящая Средняя) описывает временные ряды с помощью трех компонентов:

* AR (авто-регрессия): зависит от предыдущих значений ряда.
* I (интеграция): разности ряда для достижения стационарности.
* MA (скользящая средняя): зависит от предыдущих ошибок прогноза.

Модель ARIMA обозначается как ARIMA(p, d, q), где:

* *p* – порядок авто-регрессии,
* *d* – степень интеграции (количество разностей),
* *q* – порядок скользящей средней.

Временной ряд - это последовательность данных, которые были собраны в определенные моменты времени.  
 ARIMA подходит для прогнозирования временных рядов, учитывая временные зависимости и сезонные колебания. Предполагается, что она подойдет для предсказания числа пациентов на ближайшие дни, основываясь на данных за предыдущие года. Однако, данная модель не учитывает нелинейные зависимости, из-за чего может показать плохие результаты и по итогу не подойти для задачи прогнозирования количества пациентов.

##### 1.8 Нейронная сеть LSTM

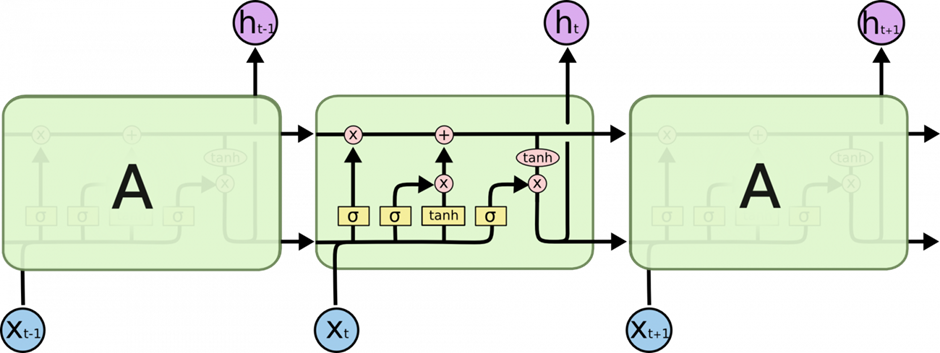


Рис. 1 - Схема ячейки LSTM.

LSTM (Long Short-Term Memory) — это тип рекуррентной нейронной сети, созданный для работы с последовательными данными. Основной элемент LSTM — это ячейка памяти, которая может сохранять информацию на длительное время.   
 Ключевым элементом LSTM является состояние ячейки (cell state), представленное в виде горизонтальной линии, проходящей по верхней части схемы. Она проходит напрямую через всю цепочку, участвуя лишь в нескольких линейных преобразованиях.

LSTM эффективно управляет информацией с помощью фильтров, которые решают, что следует пропустить, а что забыть в зависимости от условий. Эти фильтры состоят из слоя сигмоидальной нейронной сети и операции поточечного умножения.

Сигмоидальный слой играет ключевую роль, возвращая значения от 0 до 1, что определяет, какую часть информации передать по сети. Здесь 0 означает "не передавать ничего", а 1 – "передавать все".

В LSTM реализованы три фильтра, каждый с уникальной задачей состояния ячейки

1. Фильтр забывания (Forget Gate Layer):

* Этот фильтр оценивает значения h{t-1} и x{t}, возвращая для каждого элемента состояния ячейки C{t-1} значения в диапазоне от 0 до 1. Значение 1 указывает на полное забывание предыдущей информации, тогда как 0 означает ее полное сохранение.

2. Фильтр входных данных (Input Layer Gate):

* Этот фильтр отвечает за определение значений, которые необходимо обновить. Затем tanh-слой генерирует вектор новых значений-кандидатов, которые будут добавлены в состояние ячейки. Старое состояние умножается на значения забывания, после чего к нему прибавляются новые значения-кандидаты, умноженные на коэффициент обновления t для каждого элемента состояния.

3. Выходной фильтр (Output Gate Layer):

* Этот фильтр играет ключевую роль в извлечении информации из состояния ячейки. Он анализирует входные данные h{t-1} и x{t}, и с помощью сигмоидального слоя формирует весовые коэффициенты для каждого элемента текущего состояния ячейки C{t}. Затем с использованием функции tanh создается вектор кандидатов для нового выхода. После выполнения поточного умножения значений от сигмоидального слоя и tanh-слоя формируется новый вектор вывода, который становится выходом на текущий момент времени.

Использование фильтров в нейронной сети LSTM, позволяет эффективно управлять информацией внутри состояния каждой ячейки, этот подход хорошо подойдет для обработки последовательных данных, особенно в контексте предсказания количества пациентов на будущие дни.   
 Настраиваемые гиперпараметры в нейронной сети:

1. Скорость обучения
2. Количество эпох обучения
3. Количество слоев
4. Общая архитектура: Количество нейронов на слой и связь между слоями
5. Функция активации

##### 1.9 Линейная регрессия

Линейная регрессия моделирует зависимость между зависимой переменной *Y* (число пациентов) и независимыми переменными

𝑋1,𝑋2,...,𝑋𝑛(климатические данные и другие факторы) с помощью линейной функции:

* *Y* — зависимая переменная (число пациентов),
* 𝛽0 — свободный член (константа),
* 𝛽𝑖​ — коэффициенты регрессии, показывающие влияние каждой независимой переменной,
* 𝑋𝑖 — независимые переменные (климатические данные),
* — ошибка модели.

Линейная регрессия позволяет выявлять линейные зависимости между климатическими факторами и числом пациентов. Однако, она не учитывает нелинейные зависимости, из-за чего может упустить из виду скрытые зависимости между климатическими данными и количеством пациентов в определенные дни.   
 Для оценки предсказаний регрессионной модели градиентного бустинга используется несколько метрик, которые также подходят для оценки качества моделей LSTM и ARIMA для задачи регрессии:

1. Среднеквадратичная ошибка (Mean Squared Error):

где - истинное значение, - предсказанное значение, n - количество наблюдений.

1. Корень из среднеквадратичной ошибки (Root Mean Squared Error):

RMSE - интерпретируется в тех же единицах измерения, что и исходные данные.

1. Средняя абсолютная процентная ошибка (Mean Absolute Percentage Error):

MAPE - показывает средний процент ошибки предсказания относительно истинного значения.

4. Средняя абсолютная ошибка (Mean Absolute Error):

MAE - измеряет среднюю абсолютную ошибку между фактическими и прогнозируемыми значениями.

5. Коэффициент детерминации (R² score):

где - среднее значение истинных значений. Коэффициент детерминации показывает, какая доля дисперсии зависимой переменной объясняется моделью. Более подробно, score - это метрика, которая показывает, насколько хорошо данные соответствуют модели. Это значение находится в диапазоне от 0 до 1. Значение близкое к 0, означает, что данные случайны или не могут соответствовать модели. Значение близкое к 1, показывает, что данные хорошо соответствуют модели.

Для наглядной разницы моделей, ниже составлена таблица 1, сравнения с преимуществами и недостатками каждой модели:

Таблица 1

| Модель | Тип | Преимущества | Недостатки |
| --- | --- | --- | --- |
| Градиентный бустинг | Ансамблевое обучение | 1. Хорошо работает с пропущенными данными  2. Справляется с нелинейными зависимостями | 1. Риск переобучения  2. Чувствительность к выбросам  3. Требует настройки гиперпараметров |
| XGBoost | Ансамблевое обучение | 1. Высокая скорость обучения  2. Регуляризация L1 и L2  3. Обработка пропущенных значений | 1. Менее эффективен на малых выборках  2. Сложнее в настройке по сравнению с обычным градиентным бустингом |
| LightGBM | Ансамблевое обучение | 1. Высокая скорость и низкое потребление памяти  2. Эффективная обработка больших данных | 1. Может плохо работать с небольшими наборами данных  2. Сложность настройки |
| CatBoost | Ансамблевое обучение | 1. Отличная работа с категориальными данными  2. Минимизация утечки информации | 1. Может требовать больше времени на обучение  2. Меньше документации по сравнению с XGBoost |
| ARIMA | Временные ряды | 1. Хорошо подходит для стационарных временных рядов  2. Учитывает временные зависимости | 1. Не учитывает нелинейные зависимости  2. Требует стационарности ряда |
| LSTM | Реккурентные нейронные сети | 1. Эффективно работает с последовательными данными  2. Способность запоминать долгосрочные зависимости | 1. Сложность в настройке  2. Требует больших объемов данных для обучения |
| Линейная регрессия | Регрессия | 1.Интерпретируемость данных | 1. Не учитывает нелинейные зависимости  2. Чувствительность к выбросам |

В результате описания существующих моделей, были приведены теоретические сведения по каждой модели. Далее, в следующей главе, будет описано сравнение работы этих моделей на имеющихся данных, а также получены результаты прогнозирования количества пациентов, о которых пойдет речь в следующих главах.

#### ГЛАВА 2. РЕАЛИЗАЦИЯ

##### 2.1 Постановка задачи

Разработка web-приложения, которое будет являться инструментом

для прогнозирования числа пациентов первичного приема на основе климатических данных, что позволит руководителям медицинских учреждений заранее знать о возможных нагрузках в определенные дни.

Таким образом, для реализации проекта были поставлены следующие задачи:

1. Провести разведочный анализ данных   
   Разведочный анализ данных — это анализ основных свойств данных, нахождение в них закономерностей их распределения и взаимосвязи признаков.
2. Исследовать возможность применения моделей машинного обучения для прогнозирования числа обращений пациентов. (Изучение эффективности классических подходов)
3. Выбрать модель прогнозирования числа пациентов на основании выполненного исследования
4. Оценка эффективности модели
5. Сформировать функционал web-приложения
6. Спроектировать архитектуру web-приложения, согласно функциональным требованиям и клиент-серверному взаимодействию
7. Разработать серверную часть web-приложения (backend)
8. Спроектировать пользовательский интерфейс web-приложения
9. Разработать клиентскую часть web-приложения (frontend)
10. Провести тестирование web-приложения на известных данных

##### 2.2 Подбор стека технологий

Для разработки серверной части web-приложения (backend), также будет использован язык программирования Python, фреймворк FastAPI и Redis.

FastAPI – веб-фреймворк для создания API, написанный на Python. FastAPI активно использует декораторы, аннотации типов и интроспекцию кода, что позволяет уменьшить количество шаблонного кода в веб-приложении. Также фреймворк автоматически генерирует и отображает документацию согласно спецификации OpenAPI.

Для разработки клиентской части web-приложения (frontend), предполагается использовать язык программирования JavaScript фреймворк Nuxt.js. Он написан на базе другого фреймворка — Vue.js.

Nuxt.js 3 — это фреймворк для разработки серверных и клиентских приложений на основе Vue.js. Он упрощает создание универсальных (исходно рендерящихся на сервере) приложений, поддерживает статическую генерацию и серверный рендеринг (SSR), а также улучшает производительность и SEO. Nuxt 3 основан на Vue 3 и использует новые возможности, такие как Composition API и Vite для сборки.

Некоторые возможности Nuxt.js:

* возможность создания SPA, PWA, JAMstack-сайтов, а также статически сгенерированных сайтов;
* поддержка SSR и CSR;
* поддержка СSS-препроцессоров: Sass, Less, Stylus;
* технология автоматического разделения JavaScript-кода на страницы и компоненты;
* система маршрутизации с асинхронными данными;
* обработка статических файлов;
* объединение и минимизация JavaScript- и CSS-кода.

NGINX **–** это программное обеспечение с открытым исходным кодом для создания легкого и мощного веб-сервера. Также его используют в качестве почтового или обратного прокси-сервера. Nginx решает проблему падения производительности с ростом трафика и является самым популярным веб-сервером в России ивторым в мире.

NGINX действует как обратный прокси-сервер, принимающий входящие HTTP-запросы от клиентов и перенаправляющий их на сервер FastAPI. Он управляет маршрутизацией запросов между клиентской и серверной частями приложения. Кроме того, NGINX используется для обслуживания статических файлов, таких как HTML, CSS, JavaScript и изображения, генерируемые Nuxt.js. NGINX также добавляет дополнительный уровень безопасности для приложения и предоставляет возможности для мониторинга и логирования.  
 Redis – этосистема управления базами данных, которая хранит информацию в виде пар ключ-значение. Ключ – это название какого-то поля, а значение – его содержание. Используется для кеш хранилища.  
 Для реализации моделей машинного обучения, использовался язык программирования Python и его библиотеки:

1. Для реализации моделей и метрик: Scikit-learn, tensorflow, xgboost, catboost, lightgbm, statsmodels.
2. Для расчетов: numpy
3. Для работы с датасетом: pandas
4. Для сохранения модели: joblib
5. Для построения графиков: matplotlib, shap, seaborn.

2.3 Предобработка данных

В качестве исходных данных использовались данные о посещениях (были предоставлены научным руководителем) медицинских учреждений в период с ноября 2022 года по октябрь 2023 года города Тюмени, а также климатические данные за указанный период (получены с сайта RP5).  
 Данные о посещаемости предоставляли следующую информацию:

* Пациент
* Дата рождения
* Пол
* Диагноз
* Услуга
* Дата обращени
* Сотрудник медицинского учреждения
* Причина посещения
* Результат обращения

Климатические данные предоставляли следующую информацию:

* Дату и время замера
* Температура
* Давление
* Влажность
* Скорость ветра

Во время этапа предобработки был сформирован следующий датасет, который был использован для обучения моделей:

* День недели (Day\_of\_Week)
* Количество пациентов, посетивших медицинские учреждения в этот день
* Средняя температура воздуха (T\_mean)
* Максимальная температура воздуха (T\_max)
* Минимальная температура воздуха (T\_min)
* Среднее давление (Po\_mean)
* Максимальное давление (Po\_max)
* Минимальное давление (Po\_min)
* Средняя влажность воздуха (U\_mean)
* Максимальная влажность воздуха (U\_max)
* Минимальная влажность воздуха (U\_mean)
* Средняя скорость ветра (Ff\_mean)
* Максимальная скорость ветра (Ff\_max)
* Минимальная скорость ветра (Ff\_min)
* Временной лаг в 3 дня для каждого климатического параметра (в датасете отображен с постфиксом \_lagN, где N – нумерация лага от 1 до 3)

В результате предобработки данные о посещаемости использовались только для подсчета количества пациентов, посетивших медицинское учреждение, поскольку остальные параметры, на данном этапе проекта, оказали слабое влияние на итоговый результат предсказания и могут использованы в будущем для детального предсказания (например, прогнозировать количество пациентов по полу, отдельным возрастным группам). Для предобработки данных использовалась библиотек pandas и Python.

##### 2.3 Реализация моделей

Перед тем как произвести выбор конкретной модели для нашего проекта, необходимо было реализовать каждую модель, выполнить предобработку данных, и настроить гиперпараметры. Настройка гиперпараметров моделей выполнялась с применением кросс-валидации и эмпирического анализа. Для этого использовалась библиотека Scikit-learn. Scikit-learn - это библиотека для машинного обучения на языке Python.  
 Также стоит внести пояснение по термину модель. Модель - это абстрактное представление системы (ее математическое описание), которое используется для прогнозирования и решения математических задач.   
 Обычная модель градиентного бустинга GradientBoostingRegressor. Для этой модели были определены соответствующие настройки гиперпараметров:

1. n\_estimators = 600 - количество деревьев, которые будут построены в модели.
2. learning\_rate = 0.01 - скорость обучения модели, такая скорость обосновывается большим количество деревьев для модели.
3. max\_depth = 5 - максимальную глубина деревьев, помимо применения кросс-валидации, такая глубина обосновывается тем, что более глубокие деревья улучшают качество модели.
4. random\_state = 42 - случайное начальное значение для воспроизводимости результатов, данный параметр, особо не влияет на результат, обычно всегда выставляется значение 42, но оно может любым.

Модель XGBoost. Для этой модели были определены следующие гиперпараметры:

1. n\_estimators = 500 - количество деревьев.
2. learning\_rate = 0.01 - скорость обучения модели.
3. max\_depth = 5 - максимальную глубина деревьев.
4. random\_state =42 - случайное начальное значение для воспроизводимости результатов.

Модель LightGBM. Для этой модели были определены следующие гиперпараметры:

1. n\_estimators = 600 - количество деревьев.
2. learning\_rate = 0.01 - скорость обучения модели.
3. max\_depth = 5 - максимальную глубина деревьев.
4. random\_state = 42 - случайное начальное значение для воспроизводимости результатов.

Модель CatBoost. Для этой модели были определены следующие настройки гиперпараметров:

1. iterations = 2000 - количество итераций деревьев, которые будут построены.
2. learning\_rate = 0.01 - скорость обучения модели также невысокая, чтобы избежать переобучения
3. depth = 5 - глубина деревьев.
4. random\_state = 42 - случайное начальное значение для воспроизводимости результатов.

Модель LSTM. Создание нейронной сети:  
 В модели применяется среднеквадратичная ошибка (MSE), как функция активации.  
 Adam — это оптимизатор, который применяется для обучения нейронных сетей. Он объединяет подходы двух оптимизаторов: Momentum Optimization и RMSprop (Root Mean Square Propagation). Adam адаптивно регулирует скорость обучения для каждого параметра.   
 Эпоха — это один полный проход по всему набору данных в процессе обучения модели.  
 Далее, происходила настройка гиперпараметров:

1. learning\_rate = 0.01 - скорость обучения модели.
2. loss = "mean\_squared\_error" - функция потерь (MSE).
3. optimizer = Adam - оптимизатор Adam.
4. epochs = 2000 - количество эпох обучения модели.
5. batch\_size = 16 - это параметр, который определяет количество образцов, обрабатываемых моделью за одну итерацию обучения. Это означает, что модель обновляет свои веса после обработки каждого пакета из 16 образцов данных.

Нейронная сеть для анализа климатических и временных данных построена таким образом, что сначала создается слой LSTM (Long Short-Term Memory) с 100 нейронами для учета долгосрочных зависимостей в последовательности климатических данных. Это позволяет модели улавливать последовательные зависимости.

За ним идет слой Dropout - этот слой используется для регуляризации модели, он случайным образом обнуляет часть векторов на каждом шаге обучения, в нашем случае 20%, то есть 0.2. Это также помогает модели избежать переобучения.   
 Далее идет еще один слой LSTM, в котором 100 нейронов, за ним еще слой Dropout(0.2). После этого создается финальный слой - Dense Layer. Слой Dense - это полносвязный слой с одним нейроном и линейной функцией активации.  
 Настройка параметров модели ARIMA:  
 Для задачи прогнозирования количества пациентов, были подобраны параметры p, d, q со значениями:

1. p = 1.
2. d = 1.
3. q = 1.

Перед обучением модели было необходимо проверить, является ли временной ряд стационарным. Для этого использовался тест Дики-Фуллера (ADF). Если значение больше 0.05, то временной ряд считается не стационарен и в таком случае, потребуется применить дифференцирование.   
 Стационарный ряд - это временной ряд, свойства которого не изменяются со временем. То есть, среднее значение остается постоянным, а дисперсия не меняется.   
 Тест Дики-Фуллера - это методика, которая используется в прикладной статистике для анализа временных рядов для проверки на стационарность. То есть, проверяет гипотезу о том, что временной ряд либо стационарен (имеет единичный корень), либо нет.

##### 2.4 Разработка серверной части приложения

Серверная часть реализована при использовании фреймворка FastAPI и Redis (кеш система). Система кэширования потребовалась для временного хранения климатических данных с сервиса Яндекс.Погода, поскольку частые запросы могут привести к временной блокировке со стороны сервиса Яндекс.Погода. Изначально для получения климатических данных использовался сервис RP5, однако, он не предоставляет прогноз на срок дольше 7 дней. Вследствии этого, сервис Яндекс.Погода был выбран для получения актуального прогноза погоды.   
 Серверная часть приложения предоставляет результат прогноза в структуре массива, передаваемого в формате JSON, со следующими данными:

* Дата
* Спрогнозированное число
* Бинарный признак - рабочий или выходной день (true/false)

##### 2.5 Интерфейс пользователя

Пользовательский интерфейс (фронтенд) выполнен с использование фреймворка NuxtJS3 и библиотеки Tailwind. Фронтенд получает результат прогноза с серверной части приложения, после чего, после его обработки, отображает его пользователю.  
 На рисунке представлен интерфейс приложения, где имеется блок для указания длины прогноза и кнопка “Получить прогноз”, по нажатию которой прогноз будет запрошен с бекенд части приложения и отображен в блоке прогнозных элементов справа.  
 Каждый элемент прогноза предоставляет для пользователя следующею информацию:

* День недели
* Дата
* Процент, на который отличается спрогнозированное число от среднего числа пациентов за день, в зависимости от дня - рабочий или выходной.
* Спрогнозированное кол-во пациентов



Рис. 2 – Страница “Прогноз”

Также имеется раздел “О проекте”. Кроме информации о самом проекте, данный раздел содержит описание того, как рассчитываются конечные значения, которые видит пользователь. Так же из данного раздела можно узнать актуальные средние значения по посещаемости пациентов.

##### 

Рис. 3 – Страница “О проекте”

##### 2.6 Функционал системы

Система предоставляет возможность получения прогноза количества пациентов на каждый день, количество дней прогноза выбирает сам пользователь. Длина прогноза ограничена от 3 до 30 дней, однако, конечное число зависит от имеющегося прогноза погоды и может сокращаться до 28 дней.

##### 2.7 Архитектура веб-приложения

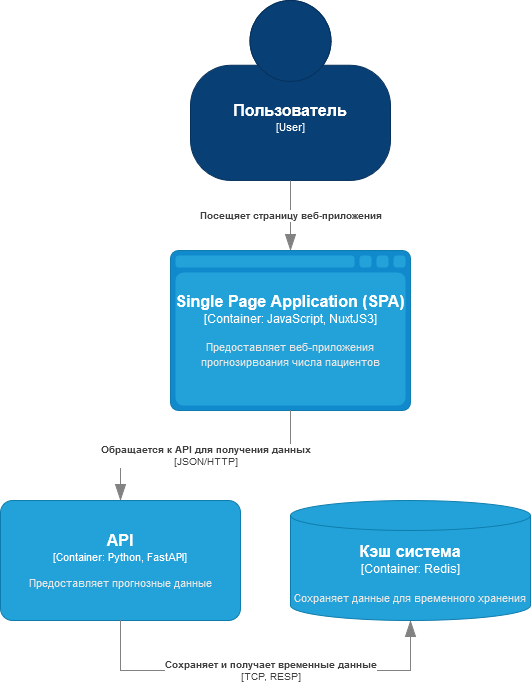


Рис. 4 – Архитектура приложения

Архитектура приложения, представленная на рисунке 4, включает в себя SPA-приложение, созданное при использовании фреймворка NuxtJS3, серверную часть приложения, выполненную с использованием фреймворка FastAPI, а также Redis в качестве кеш системы для временного хранения данных.

#### ГЛАВА 3. СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ И ОЦЕНКА ИХ КАЧЕСТВА

Для обучения и тестирования модели использовались климатические данные и данные по посещениям медучреждений первичного приема. В климатических данных учитывался немало важный параметр "Временной лаг" (временное запаздывание), если говорить в контексте прогнозирования количества пациентов - это период времени, в течении которого изменения климатических условий начинает оказывать влияние на количество обращений в медучреждения.   
 Для обучения моделей использовались данные за период с ноября 2022 года по сентябрь 2023 года. Тестирование же проводилось на данных за октябрь 2023 года. Ниже представлены результаты предсказаний каждой модели.

##### 3.1 Модель Gradient Boosting

Первый столбец отображает дату, второй столбец показывает реальные данные посещений за октябрь 2023 года, а последний столбец показывается предсказанное значение:

Результаты посещений за октябрь 2023 года представлены в таблице 2:

Таблица 2

| Дата | Реальное кол-во пациентов | Предсказанное кол-во |
| --- | --- | --- |
| 2023-10-01 | 2 | 11.92 |
| 2023-10-02 | 435 | 371.53 |
| 2023-10-03 | 376 | 399.84 |
| 2023-10-04 | 409 | 384.71 |
| 2023-10-05 | 368 | 316.89 |
| 2023-10-06 | 386 | 355.14 |
| 2023-10-07 | 78 | 95.74 |
| 2023-10-08 | 6 | 5.79 |
| 2023-10-09 | 428 | 377.57 |
| 2023-10-10 | 417 | 382.81 |
| 2023-10-11 | 410 | 384.82 |
| 2023-10-12 | 408 | 375.01 |
| 2023-10-13 | 402 | 322.47 |
| 2023-10-14 | 96 | 105.21 |
| 2023-10-15 | 4 | 23.14 |
| 2023-10-16 | 429 | 393.99 |
| 2023-10-17 | 376 | 401.13 |
| 2023-10-18 | 415 | 393.49 |
| 2023-10-19 | 360 | 366.95 |
| 2023-10-20 | 365 | 352.72 |
| 2023-10-21 | 88 | 90.06 |
| 2023-10-22 | 6 | 2.96 |
| 2023-10-23 | 362 | 382.75 |
| 2023-10-24 | 371 | 382.32 |
| 2023-10-25 | 449 | 318.65 |
| 2023-10-26 | 383 | 313.89 |
| 2023-10-27 | 337 | 364.19 |
| 2023-10-28 | 104 | 100.40 |
| 2023-10-29 | 2 | 12.32 |
| 2023-10-30 | 355 | 384.19 |
| 2023-10-31 | 391 | 395.35 |

Результаты по метрикам:

* Mean Squared Error: 1547.058
* R^2 Score: 0.9412
* Mean Absolute Error: 28.524
* Root Mean Squared Error: 39.332
* Mean Absolute Percentage Error: 57.563%

Далее на рисунке 5 представлен график обучения модели, как видно, модель показала хорошие результаты обучения.

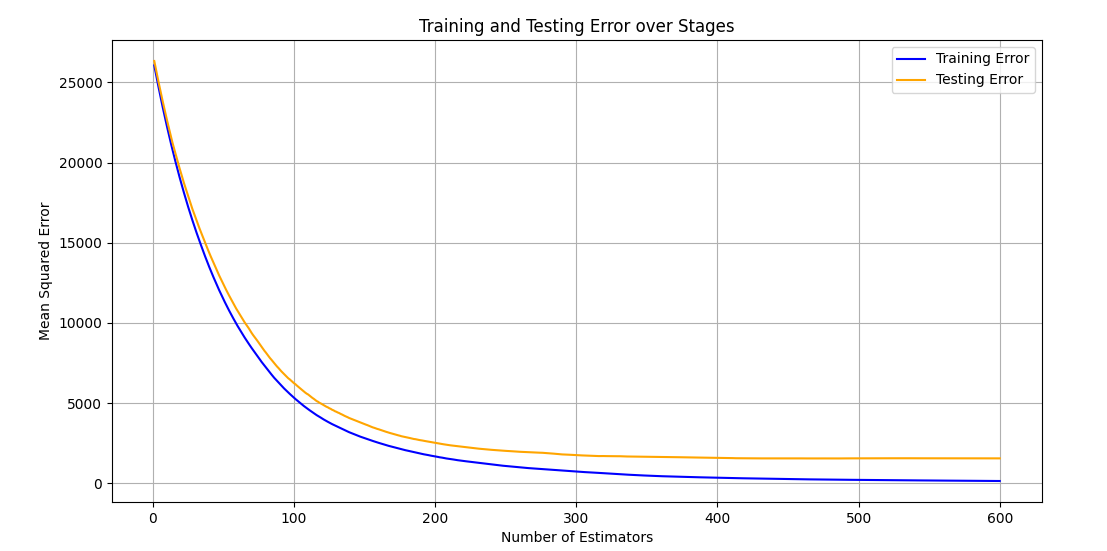


Рис. 5 - Процесс обучения модели Gradient Boosting.

Ниже на рисунке 6 представлен график гистограмм со всеми признаками и их степенью важности в предсказании количества пациентов. На нем видно, что наиболее значимым признаком является признак Day\_of\_Week.

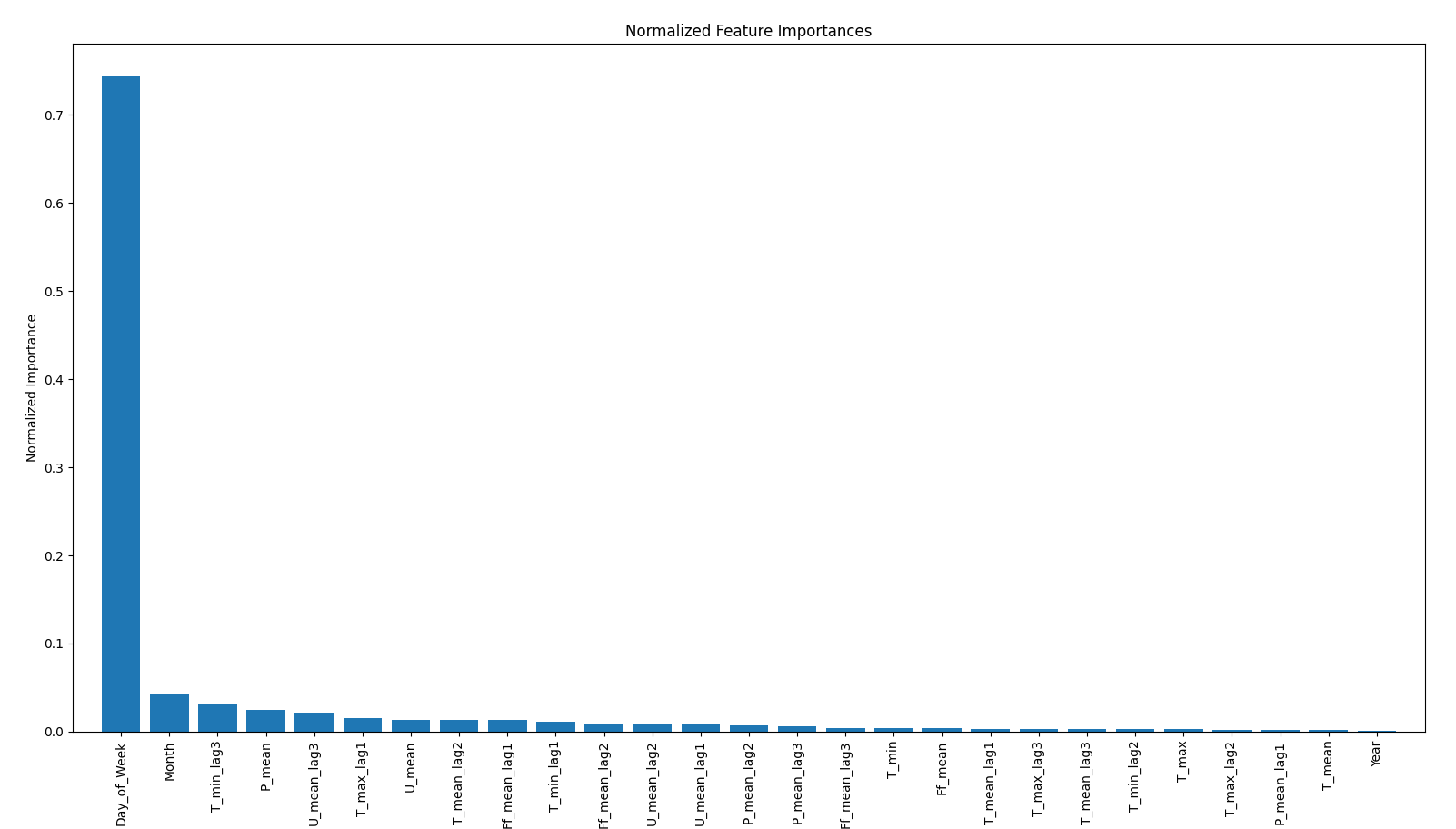


Рис. 6 - График гистограмм, нормализованных значений параметров каждого признака модели Gradient Boosting.

На рисунке 7 представлена диаграмма рассеяния, то есть насколько предсказанные значения, близки к истинным. Из него видно, что предсказанные значения сильно приближены к реальным.

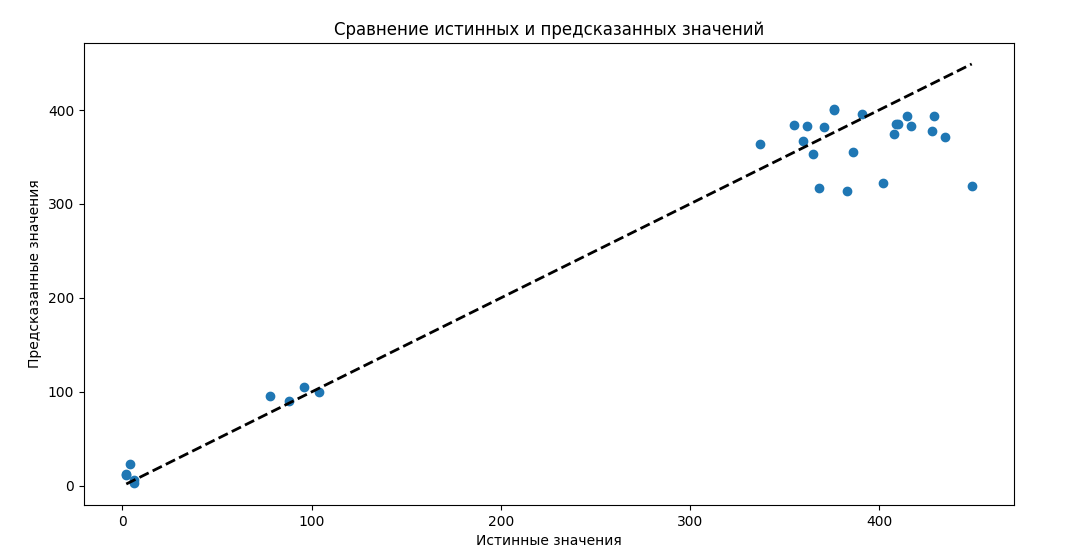


Рис. 7 - Диаграмма рассеяния предсказанных и истинных значений модели Gradient Boosting.

В таблице 3 представлены нормализованные значения (от 0 до 1) по каждому признаку.

Таблица 3

| Признак | Важность |
| --- | --- |
| Day\_of\_Week | 0.7437 |
| Month | 0.0421 |
| T\_min\_lag3 | 0.0311 |
| P\_mean | 0.0244 |
| U\_mean\_lag3 | 0.0213 |
| T\_max\_lag1 | 0.0153 |
| U\_mean | 0.0134 |
| T\_mean\_lag2 | 0.0132 |
| Ff\_mean\_lag1 | 0.013 |
| T\_min\_lag1 | 0.0108 |
| Ff\_mean\_lag2 | 0.0089 |
| U\_mean\_lag2 | 0.008 |
| U\_mean\_lag1 | 0.0078 |
| P\_mean\_lag2 | 0.007 |
| P\_mean\_lag3 | 0.0062 |
| Ff\_mean\_lag3 | 0.0044 |
| T\_min | 0.0041 |
| Ff\_mean | 0.0041 |
| T\_mean\_lag1 | 0.0031 |
| T\_max\_lag3 | 0.0031 |
| T\_mean\_lag3 | 0.0031 |
| T\_min\_lag2 | 0.0026 |
| T\_max | 0.0025 |
| T\_max\_lag2 | 0.0024 |
| P\_mean\_lag1 | 0.0021 |
| T\_mean | 0.0016 |
| Year | 0.0006 |

По результатам обучения модели, как оказалось, наиболее значимым признаком, который влияет на прогнозирование - это Day\_of\_Week (признак определения дня недели), он имеет значение 0.7437 или 74.37%, что является высоким показателем и говорит о том, что день недели существенно влияет на количество обращений пациентов. Это, скорее всего, связано с тем, что в определенные дни, наблюдается повышенная активность в обращении за медицинской помощью. Следующим по значимости идет признак Month (месяц), со значением 0.0421 или 4.21%. Это говорит о том, что помимо признака Day\_of\_Week, признак месяца, также влияет на прогнозирование, однако в меньшей степени, чем предыдущий. Далее идут погодные признаки и значение с временным запаздыванием. T\_min\_lag3 (минимальное значение температуры с временным лагом 3) имеет значение 0.0311 или 3.11%. Признак P\_mean (среднее значение атмосферного давления) имеет значение 0.0244 или 2.44%.

Исходя из этого, можно сказать, что все остальные признаки также влияют на прогнозирование количества пациентов, однако в меньшей степени, чем признак Day\_of\_Week, поскольку он оказывается ключевым. Это говорит о том, что модель лучше находит зависимость между днем недели и целевым признаком, по сравнению с остальными признаками. Тем не менее, стоит отметить, что несмотря на низкую значимость, остальные признаки корректируют итоговое значение прогнозирования, повышая его точность.

##### 3.2 Модель LightGBM

Результаты посещений за октябрь 2023 года представлены в таблице 4:

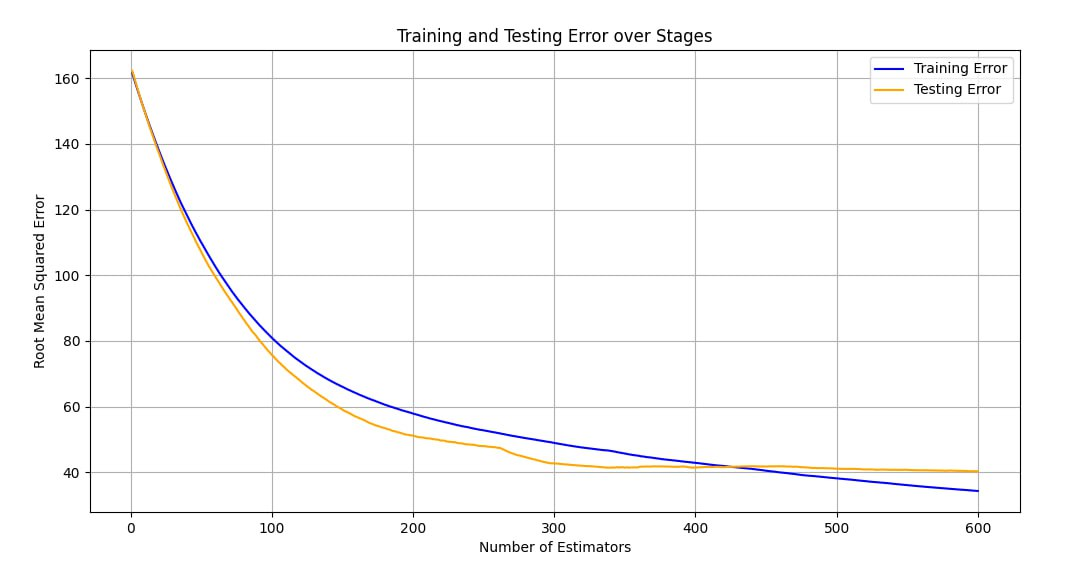
Таблица 4

| Дата | Реальное кол-во пациентов | Предсказанное кол-во |
| --- | --- | --- |
| 2023-10-01 | 2 | 5.83 |
| 2023-10-02 | 435 | 363.47 |
| 2023-10-03 | 376 | 376.17 |
| 2023-10-04 | 409 | 397.34 |
| 2023-10-05 | 368 | 370.6 |
| 2023-10-06 | 386 | 336.42 |
| 2023-10-07 | 78 | 91.34 |
| 2023-10-08 | 6 | 5.55 |
| 2023-10-09 | 428 | 399.23 |
| 2023-10-10 | 417 | 376.6 |
| 2023-10-11 | 410 | 363.1 |
| 2023-10-12 | 408 | 403.66 |
| 2023-10-13 | 402 | 331.1 |
| 2023-10-14 | 96 | 84.54 |
| 2023-10-15 | 4 | 12.04 |
| 2023-10-16 | 429 | 394.03 |
| 2023-10-17 | 376 | 387.76 |
| 2023-10-18 | 415 | 406.76 |
| 2023-10-19 | 360 | 356.65 |
| 2023-10-20 | 365 | 364.04 |
| 2023-10-21 | 88 | 76.03 |
| 2023-10-22 | 6 | 6.94 |
| 2023-10-23 | 362 | 386.17 |
| 2023-10-24 | 371 | 340.4 |
| 2023-10-25 | 449 | 330.56 |
| 2023-10-26 | 383 | 316.26 |
| 2023-10-27 | 337 | 384.89 |
| 2023-10-28 | 104 | 92.97 |
| 2023-10-29 | 2 | 17.34 |
| 2023-10-30 | 355 | 270.51 |
| 2023-10-31 | 391 | 422.35 |

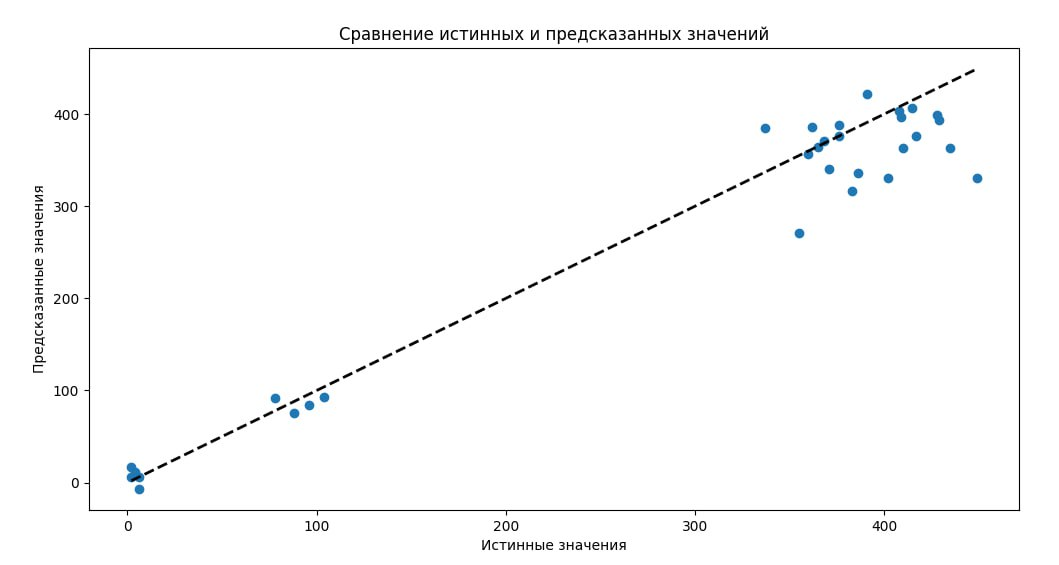
Результаты по метрикам:

* Mean Squared Error: 1623.609
* R^2 Score: 0.9383
* Mean Absolute Error: 28.328
* Root Mean Squared Error: 40.294
* Mean Absolute Percentage Error: 52.761%

Далее на рисунке 8 представлен график обучения модели LightGBM, как видно, модель показала хорошие результаты обучения, однако видны некоторые скачки с 300-ой эпохи, но на эпохе 600 график выравнивает значение ошибки на тестовой выборке.

****Рис. 8 - Процесс обучения модели LightGBM.

На рисунке 9 представлена диаграмма рассеяния модели LightGBM, из него видно, что предсказанные значения сильно приближены к реальным. Однако на значениях ближе к 400, есть несколько предсказаний, которые сильно отклонились от реальных значений.

****Рис. 9 - Диаграмма рассеяния предсказанных и истинных значений модели LightGBM.

Далее на рисунке 10 представлен график гистограмм нормализованных значений параметров каждого признака, и их степень важности в предсказании. На нем видно, что наиболее значимым признаком также остается признак Day\_of\_Week, хотя и с меньшим весом, чем в результатах предыдущей модели.

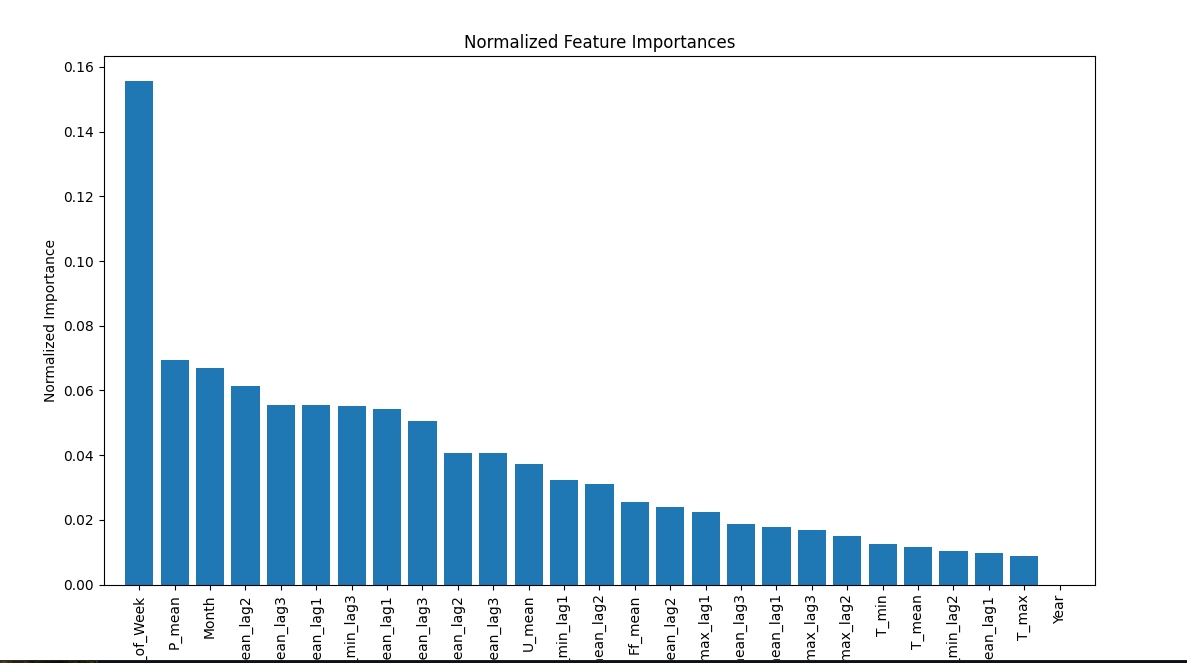


Рис. 10 - График гистограмм, нормализованных значений параметров каждого признака модели LightGBM.

Нормализованные значения представлены в таблице 5:

Таблица 5

| Признак | Важность |
| --- | --- |
| Day\_of\_Week | 0.1555 |
| P\_mean | 0.0693 |
| Month | 0.0671 |
| U\_mean\_lag2 | 0.0615 |
| P\_mean\_lag3 | 0.0556 |
| Ff\_mean\_lag1 | 0.0556 |
| T\_min\_lag3 | 0.0553 |
| U\_mean\_lag1 | 0.0543 |
| U\_mean\_lag3 | 0.0505 |
| P\_mean\_lag2 | 0.0408 |
| Ff\_mean\_lag3 | 0.0406 |
| U\_mean | 0.0371 |
| T\_min\_lag1 | 0.0322 |
| T\_mean\_lag2 | 0.0311 |
| Ff\_mean | 0.0255 |
| Ff\_mean\_lag2 | 0.0239 |
| T\_max\_lag1 | 0.0225 |
| T\_mean\_lag3 | 0.0186 |
| T\_mean\_lag1 | 0.0178 |
| T\_max\_lag3 | 0.0169 |
| T\_max\_lag2 | 0.0151 |
| T\_min | 0.0126 |
| T\_mean | 0.0116 |
| T\_min\_lag2 | 0.0105 |
| P\_mean\_lag1 | 0.0097 |
| T\_max | 0.0088 |
| Year | 0 |

По результатам обучения модели, наиболее значимым признаком, который влияет на прогнозирование, оказался Day\_of\_Week (признак определения дня недели), имеющий значение 0.1555 или 15.55%. Это указывает на то, что день недели имеет определенное влияние на количество обращений пациентов, однако его значимость значительно ниже по сравнению с предыдущими результатами модели GradientBoosting. Это может быть связано с тем, что в данной модели другие факторы стали более важными для прогнозирования.

Следующим по значимости является признак P\_mean (среднее значение атмосферного давления), который имеет значение 0.0693 или 6.93%. Это свидетельствует о том, что атмосферное давление также оказывает влияние на обращения за медицинской помощью, возможно, из-за его связи с метеорологическими условиями, которые могут влиять на здоровье людей.

Признак Month (месяц) имеет значение 0.0671 или 6.71%, что говорит о его значимости в контексте сезонных изменений в заболеваемости. Это может указывать на то, что в определенные месяцы года наблюдается увеличение обращений, например, в зимний период из-за простудных заболеваний.

Далее идут признаки, связанные с климатическими условиями, такие как U\_mean\_lag2 (средняя влажность с временным лагом 2) с значением 0.0615 и P\_mean\_lag3 (среднее атмосферное давление с временным лагом 3) с значением 0.0556. Эти показатели указывают, что погодные условия оказывают влияние на прогнозирование количества пациентов.

Исходя из этого, можно сказать, что все остальные признаки также влияют на прогнозирование количества пациентов, но гораздо меньше, чем признак Day\_of\_Week, который остается ключевым, хотя и с значительно меньшим весом. Однако, несмотря на низкую значимость, кроме признака Year (год), который имеет значение 0.0 и не влияет на прогнозирование, остальные признаки корректируют итоговое прогнозируемое количество пациентов.

##### 3.3 Модель CatBoost

Результаты посещений за октябрь 2023 года представлены в таблице 6:

Таблица 6

| Дата | Реальное кол-во пациентов | Предсказанное кол-во |
| --- | --- | --- |
| 2023-10-01 | 2 | 17.29 |
| 2023-10-02 | 435 | 375.1 |
| 2023-10-03 | 376 | 382.19 |
| 2023-10-04 | 409 | 381.81 |
| 2023-10-05 | 368 | 370.49 |
| 2023-10-06 | 386 | 350.81 |
| 2023-10-07 | 78 | 88.33 |
| 2023-10-08 | 6 | 12 |
| 2023-10-09 | 428 | 381.09 |
| 2023-10-10 | 417 | 374.1 |
| 2023-10-11 | 410 | 370.48 |
| 2023-10-12 | 408 | 375.13 |
| 2023-10-13 | 402 | 352.24 |
| 2023-10-14 | 96 | 108.24 |
| 2023-10-15 | 4 | 17.92 |
| 2023-10-16 | 429 | 400.11 |
| 2023-10-17 | 376 | 405.72 |
| 2023-10-18 | 415 | 384.95 |
| 2023-10-19 | 360 | 372.52 |
| 2023-10-20 | 365 | 355.92 |
| 2023-10-21 | 88 | 73.56 |
| 2023-10-22 | 6 | 3.72 |
| 2023-10-23 | 362 | 384.89 |
| 2023-10-24 | 371 | 364.42 |
| 2023-10-25 | 449 | 331.06 |
| 2023-10-26 | 383 | 338.15 |
| 2023-10-27 | 337 | 355.51 |
| 2023-10-28 | 104 | 99.72 |
| 2023-10-29 | 2 | 25.51 |
| 2023-10-30 | 355 | 397.35 |
| 2023-10-31 | 391 | 404.25 |

Результаты по метрикам:

* Mean Squared Error: 1223.595
* R^2 Score: 0.9535
* Mean Absolute Error: 26.509
* Root Mean Squared Error: 34.979
* Mean Absolute Percentage Error: 85.498%

Далее на рисунке 11 представлен график обучения модели CatBoost. Модель показывает плавный процесс обучения, снижая с каждой эпохой значение ошибки. До 300 эпохи происходит резкое снижение значения ошибки. После 2000 эпохи изменения ошибки становятся незначительные.

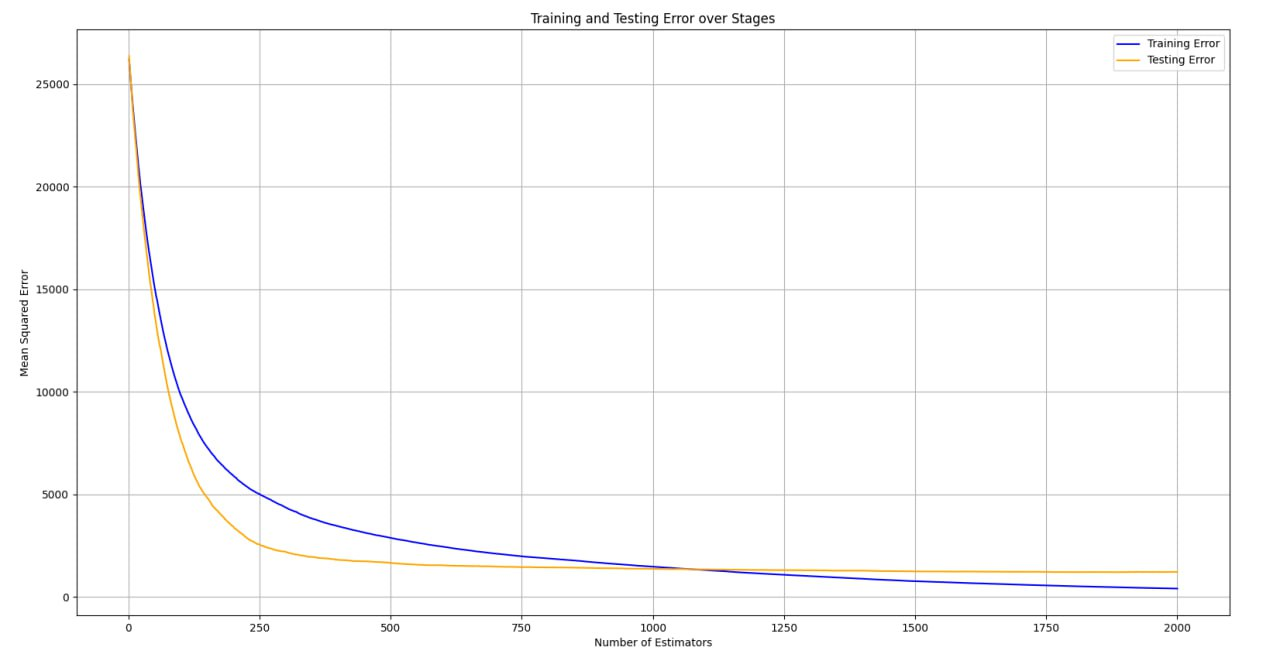


Рис. 11 - Процесс обучения модели CatBoost.

На рисунке 12 представлена диаграмма рассеяния, Исходя из результатов, показанных на диаграмме, можно сказать, что предсказанные значения сильно приближены к реальным, особенно это видно до значения 400.

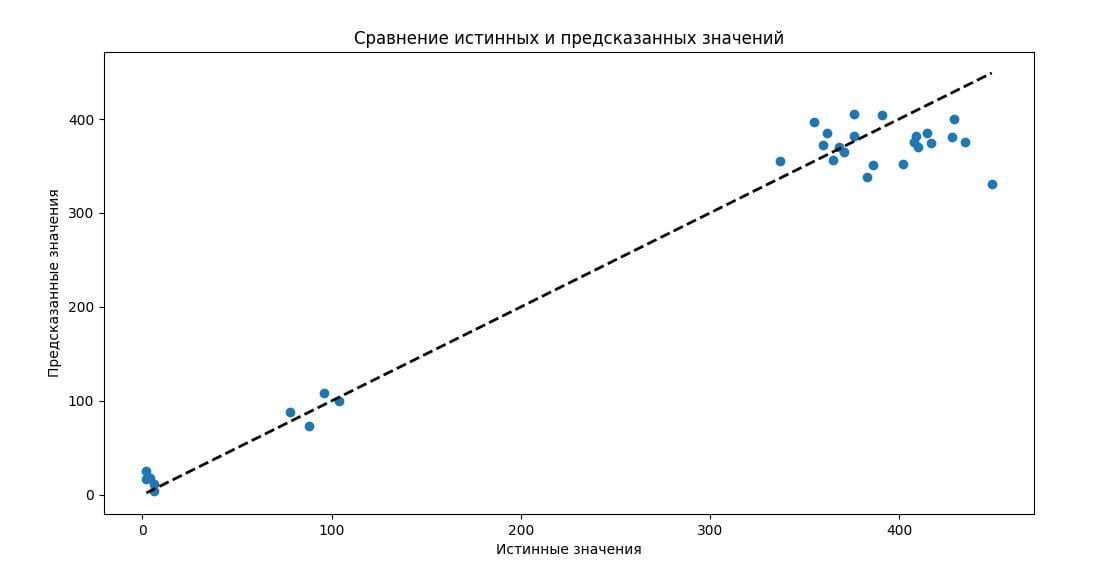


Рис. 12 - Диаграмма рассеяния предсказанных и истинных значений модели CatBoost.

Далее на рисунке 13 показан график гистограмм нормализованных значений. Исходя из него можно отметить, что значение признака Day\_of\_Week является самым высоким и приближено к 0.7, по сравнению с другими признаками, которые имеют значение, ниже 0.1.

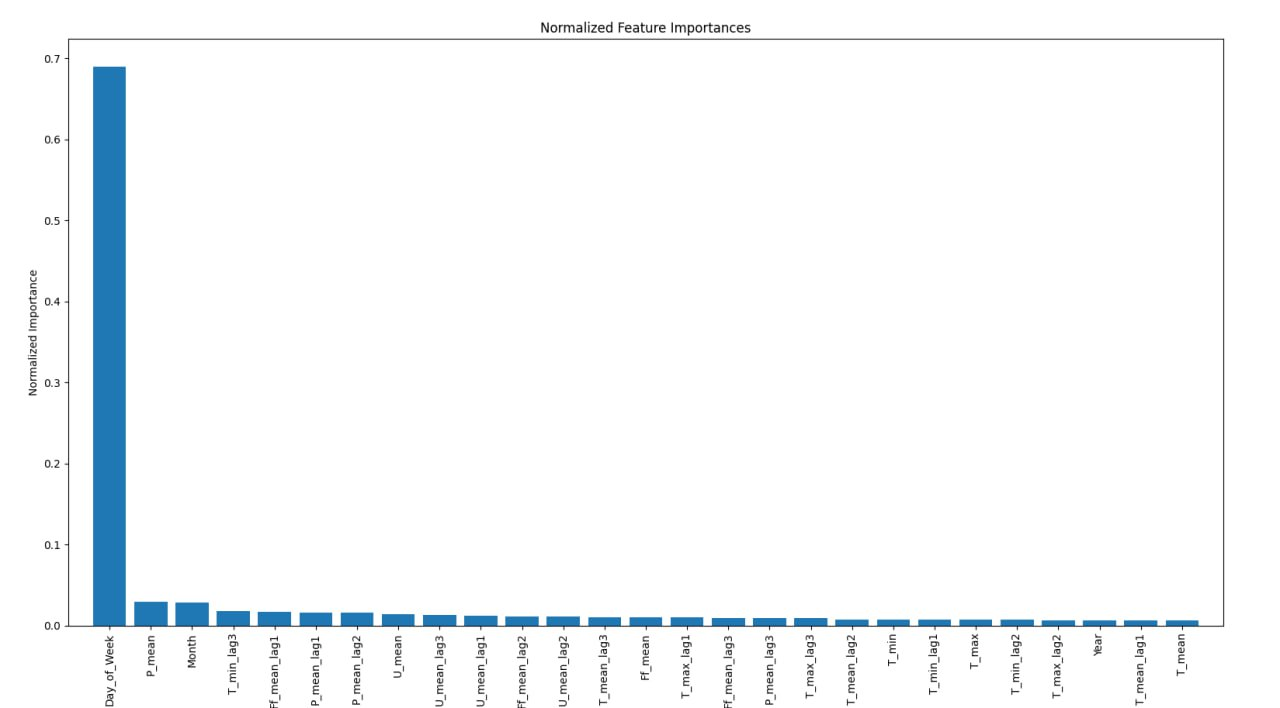


Рис. 13 - График гистограмм, нормализованных значений параметров каждого признака модели CatBoost.

Нормализованные значения представлены в таблице 7:

Таблица 7

| Признак | Важность |
| --- | --- |
| Day\_of\_Week | 0.6896 |
| P\_mean | 0.0291 |
| Month | 0.0281 |
| T\_min\_lag3 | 0.0181 |
| Ff\_mean\_lag1 | 0.0175 |
| P\_mean\_lag1 | 0.0161 |
| P\_mean\_lag2 | 0.016 |
| U\_mean | 0.0147 |
| U\_mean\_lag3 | 0.0136 |
| U\_mean\_lag1 | 0.0123 |
| Ff\_mean\_lag2 | 0.0117 |
| U\_mean\_lag2 | 0.0113 |
| T\_mean\_lag3 | 0.0108 |
| Ff\_mean | 0.0101 |
| T\_max\_lag1 | 0.0101 |
| Ff\_mean\_lag3 | 0.0097 |
| P\_mean\_lag3 | 0.0093 |
| T\_max\_lag3 | 0.0092 |
| T\_mean\_lag2 | 0.0078 |
| T\_min | 0.0076 |
| T\_min\_lag1 | 0.0075 |
| T\_max | 0.0074 |
| T\_min\_lag2 | 0.0073 |
| T\_max\_lag2 | 0.0067 |
| Year | 0.0062 |
| T\_mean\_lag1 | 0.0061 |
| T\_mean | 0.0061 |

По результатам обучения модели, наиболее значимым признаком, который влияет на прогнозирование, остается Day\_of\_Week (признак определения дня недели), имеющий значение 0.6896 или 68.96%.

Следующим по значимости является признак P\_mean (среднее значение атмосферного давления), имеющий значение 0.0291 или 2.91%. Это указывает на то, что атмосферное давление оказывает влияние на количество обращений.

Признак Month (месяц) идет следом с значением 0.0281 или 2.81%. Указывая на то, что сезонные изменения также влияют на количество обращений пациентов.

По итогу, главным признаком также остается признак дня недели, а все остальные признаки также влияют на прогнозирование количества пациентов, однако меньше оказывают влияние, чем признак Day\_of\_Week.

##### 3.4 Модель XGBoost

Результаты посещений за октябрь 2023 года представлены в таблице 8:

Таблица 8

| Дата | Реальное кол-во пациентов | Предсказанное кол-во |
| --- | --- | --- |
| 2023-10-01 | 2 | 5.17 |
| 2023-10-02 | 435 | 367.87 |
| 2023-10-03 | 376 | 379.98 |
| 2023-10-04 | 409 | 369.06 |
| 2023-10-05 | 368 | 343.26 |
| 2023-10-06 | 386 | 351.23 |
| 2023-10-07 | 78 | 85.84 |
| 2023-10-08 | 6 | 2.59 |
| 2023-10-09 | 428 | 383 |
| 2023-10-10 | 417 | 373.03 |
| 2023-10-11 | 410 | 370.29 |
| 2023-10-12 | 408 | 384.76 |
| 2023-10-13 | 402 | 318.83 |
| 2023-10-14 | 96 | 98.12 |
| 2023-10-15 | 4 | 12.19 |
| 2023-10-16 | 429 | 401.02 |
| 2023-10-17 | 376 | 406.22 |
| 2023-10-18 | 415 | 408.96 |
| 2023-10-19 | 360 | 368.98 |
| 2023-10-20 | 365 | 351.23 |
| 2023-10-21 | 88 | 92.69 |
| 2023-10-22 | 6 | 8.47 |
| 2023-10-23 | 362 | 376.13 |
| 2023-10-24 | 371 | 375.73 |
| 2023-10-25 | 449 | 345.76 |
| 2023-10-26 | 383 | 327.4 |
| 2023-10-27 | 337 | 359.76 |
| 2023-10-28 | 104 | 94.1 |
| 2023-10-29 | 2 | 14.44 |
| 2023-10-30 | 355 | 385.56 |
| 2023-10-31 | 391 | 398.81 |

Результаты по метрикам:

* Mean Squared Error: 1253.239
* R^2 Score: 0.9523
* Mean Absolute Error: 25.345
* Root Mean Squared Error: 35.401
* Mean Absolute Percentage Error: 41.693%

На рисунке 14 показан график процесса обучения модели XGBoost, на котором видно, что после четырехсотой эпохи имеются скачки значений ошибки. Процесс обучения прекращается на 500 эпохе, для того, чтобы избежать переобучения модели



Рис. 14 - Процесс обучения модели XGBoost.

На рисунке 15 представлена диаграмма рассеяния предсказанных значений, на котором можно отметить, что значения, которые идут до 100, сильно приближены к реальным значениям. Однако значения, приближенные к 400, имеют больше разброс по отношению к реальным значениям.

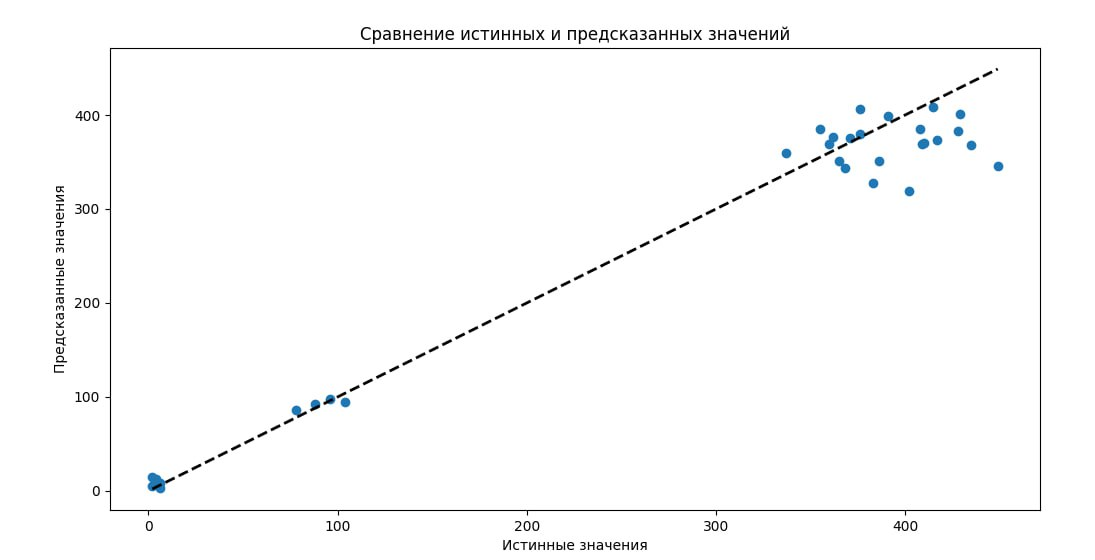
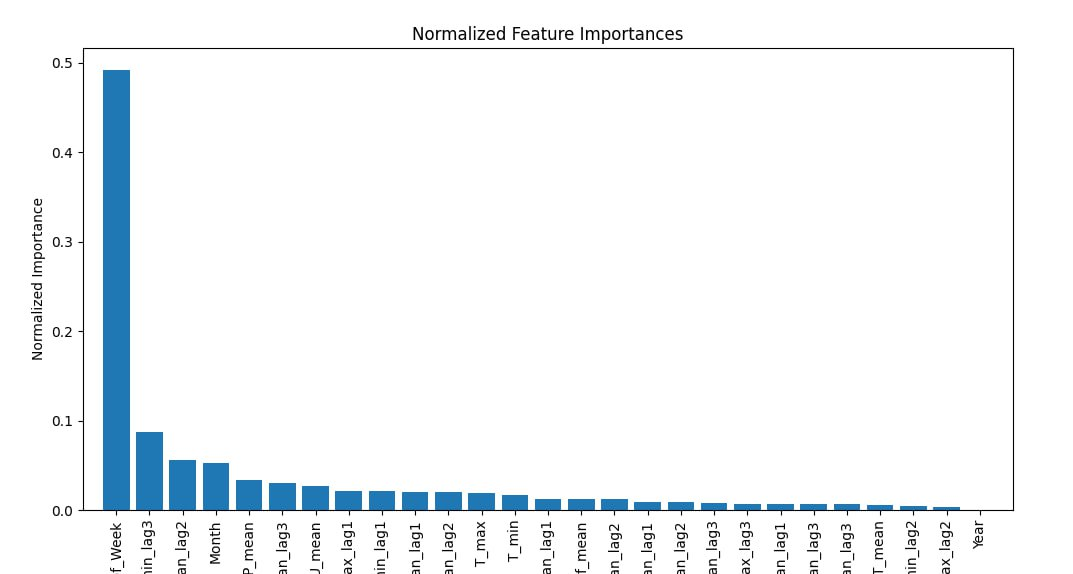


Рис. 15 - Диаграмма рассеяния предсказанных и истинных значений модели XGBoost.

Исходя из рисунка 16, можно сделать вывод, что как и в других моделях, наиболее важным признаком, оказался признак Day\_of\_Week, однако его значение равняется 0.492, не превышая отметку 0.5, по сравнению с моделью CatBoost, у которой значение признака Day\_of\_Week равняется 0.6896.

  
Рис. 16 - График гистограмм, нормализованных значений параметров каждого признака модели XGBoost.

Нормализованные значения представлены в таблице 9:

Таблица 9

| Признак | Важность |
| --- | --- |
| Day\_of\_Week | 0.492 |
| T\_min\_lag3 | 0.0873 |
| T\_mean\_lag2 | 0.056 |
| Month | 0.0524 |
| P\_mean | 0.0332 |
| U\_mean\_lag3 | 0.0299 |
| U\_mean | 0.0271 |
| T\_max\_lag1 | 0.0214 |
| T\_min\_lag1 | 0.0209 |
| Ff\_mean\_lag1 | 0.0206 |
| U\_mean\_lag2 | 0.0199 |
| T\_max | 0.0191 |
| T\_min | 0.0167 |
| T\_mean\_lag1 | 0.0122 |
| Ff\_mean | 0.0121 |
| P\_mean\_lag2 | 0.0118 |
| P\_mean\_lag1 | 0.0094 |
| Ff\_mean\_lag2 | 0.0086 |
| T\_mean\_lag3 | 0.0083 |
| T\_max\_lag3 | 0.0071 |
| U\_mean\_lag1 | 0.0068 |
| Ff\_mean\_lag3 | 0.0067 |
| P\_mean\_lag3 | 0.0066 |
| T\_mean | 0.0059 |
| T\_min\_lag2 | 0.0049 |
| T\_max\_lag2 | 0.0031 |
| Year | 0 |

По результатам определено, что ключевым признаком остается признак дня недели (0.492 или 49.2%), остальные же признаки в меньшей степени влияют на прогнозирование. Тем не менее, в этом варианте, признак года никак не влияет на результат, поскольку его значение равно 0.0.

##### 3.5 Модель LSTM

Предсказанные и реальные значения представлены в таблице 10:

Таблица 10

| Реальное | Предсказанное |
| --- | --- |
| 435 | 351.55 |
| 376 | 411.04 |
| 409 | 377.21 |
| 368 | 360.48 |
| 386 | 350.75 |
| 78 | 82.3 |
| 6 | 8.01 |
| 428 | 294.91 |
| 417 | 394.77 |
| 410 | 409.71 |
| 408 | 401.73 |
| 402 | 371.46 |
| 96 | 87.08 |
| 4 | 3.46 |
| 429 | 391.38 |
| 376 | 414.26 |
| 415 | 393.66 |
| 360 | 377.61 |
| 365 | 359.4 |
| 88 | 83.1 |
| 6 | 15.01 |
| 362 | 331.59 |
| 371 | 403.64 |
| 449 | 391.81 |
| 383 | 414.74 |
| 337 | 367.27 |
| 104 | 86.22 |
| 2 | 5.85 |
| 355 | 66.7 |
| 391 | 409.44 |

Результаты по метрикам:

* Mean Squared Error: 4151.233
* R^2 Score: 0.8293
* Mean Absolute Error: 34.873
* Root Mean Squared Error: 64.43
* Mean Absolute Percentage Error: 22.8%

На рисунке 17 представлен график обучения нейронной сети LSTM, исходя из результатов можно отметить то, что примерно первые 100 эпох, значение потерь падает с 80000 до 25000 и держится в диапазоне 25000 до 200 эпохи. После 200 эпохи, потери снова снижаются.

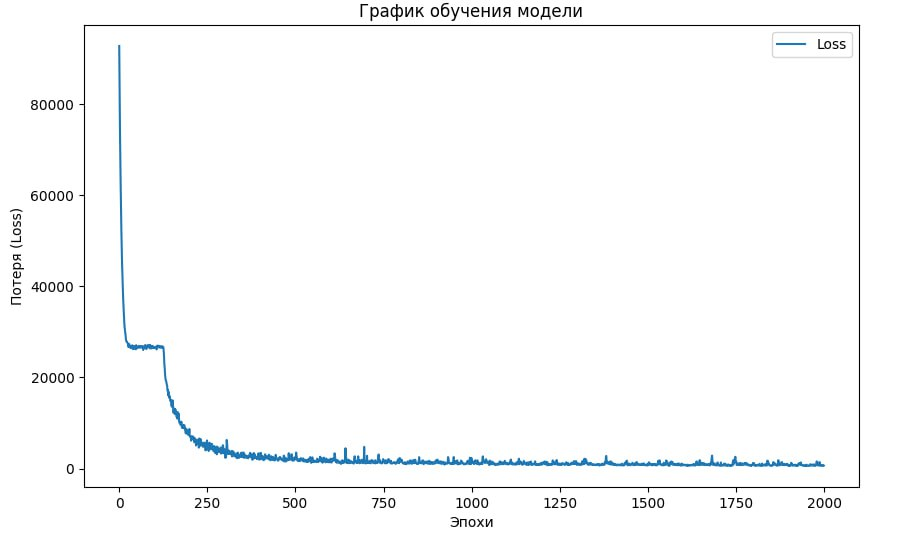


Рис. 17 - Процесс обучения нейронной сети LSTM.

На рисунке 18 изображена диаграмма рассеяния предсказанных значений по отношению к истинным. Можно отметить, что предсказанные значения довольно близки к истинным. Однако на графике есть одно значение, которое находится в диапазоне от 300 до 400 и сильно отдалено от истинного. Такое значение можно считать выбросом, это говорит о том, что несмотря на большое количество предсказанных значений, которые близки к истинным, существуют и такие, которые модель не может предсказать. Это может быть связано с тем, что модель не смогла найти зависимость для данного значения в определенный день.

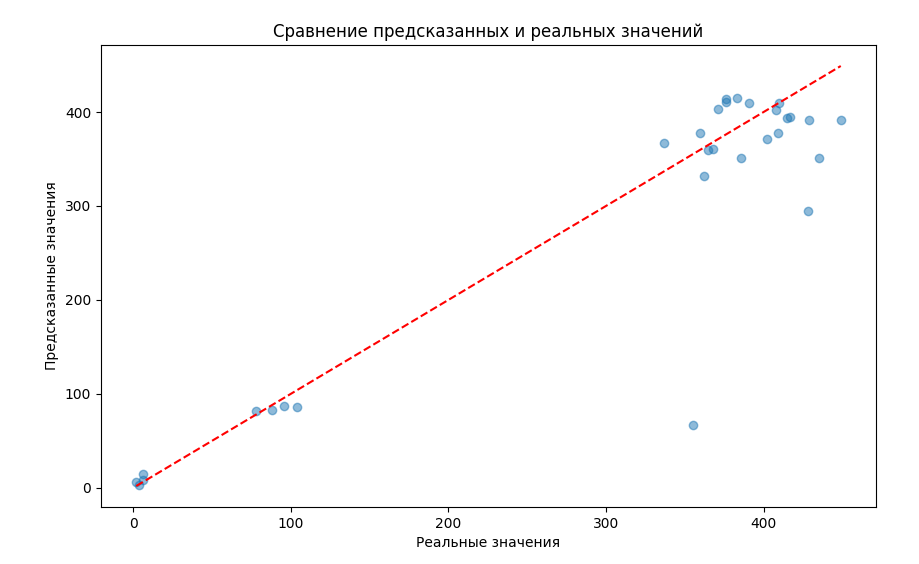


Рис. 18 - Диаграмма рассеяния предсказанных и истинных значений нейронной сети LSTM.

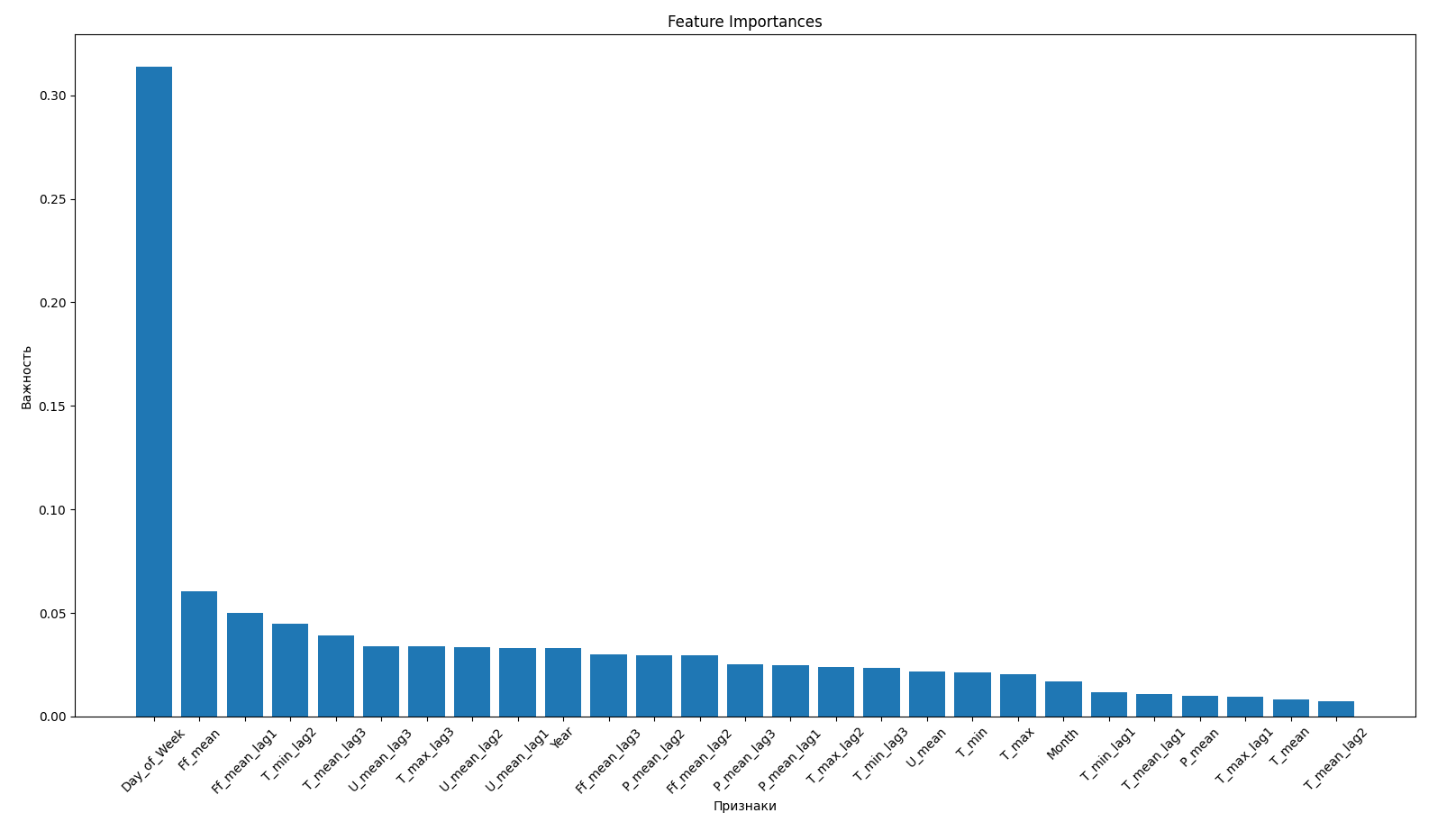


Рис. 19 - График гистограмм, нормализованных значений параметров каждого признака нейронной сети LSTM.

Исходя из рисунка 19, где представлен график нормализованных значений каждого признака, можно отметить, что как и в других случаях с предыдущими моделями, признак дня недели остается ключевым в контексте прогнозирования количества пациентов, в отличии от остальных, которые имеют меньшую значимость, но также корректируют итоговое значение прогнозирования.

Нормализованные значения представлены в таблице 11:

Таблица 11

| Признак | Важность |
| --- | --- |
| Day\_of\_Week | 0.3138 |
| Ff\_mean | 0.0605 |
| Ff\_mean\_lag1 | 0.0502 |
| T\_min\_lag2 | 0.0446 |
| U\_mean\_lag2 | 0.0333 |
| U\_mean\_lag1 | 0.0329 |
| Year | 0.0329 |
| T\_max\_lag3 | 0.034 |
| U\_mean\_lag3 | 0.034 |
| T\_mean\_lag3 | 0.0391 |
| P\_mean\_lag2 | 0.0294 |
| Ff\_mean\_lag2 | 0.0294 |
| Ff\_mean\_lag3 | 0.0298 |
| P\_mean\_lag3 | 0.0252 |
| P\_mean\_lag1 | 0.0247 |
| T\_min\_lag3 | 0.0236 |
| T\_max\_lag2 | 0.0239 |
| T\_min | 0.0215 |
| U\_mean | 0.0216 |
| T\_max | 0.0205 |
| Month | 0.0171 |
| T\_mean\_lag1 | 0.011 |
| P\_mean | 0.0101 |
| T\_max\_lag1 | 0.0095 |
| T\_mean\_lag2 | 0.0073 |
| T\_mean | 0.0083 |

##### 3.6 Модель ARIMA

Предсказанные и реальные значения представлены в таблице 12:

Таблица 12

| Реальное | Предсказанное |
| --- | --- |
| 2 | 199.88 |
| 435 | 243.6 |
| 376 | 259.68 |
| 409 | 265.59 |
| 368 | 267.76 |
| 386 | 268.56 |
| 78 | 268.85 |
| 6 | 268.96 |
| 428 | 269 |
| 417 | 269.02 |
| 410 | 269.02 |
| 408 | 269.02 |
| 402 | 269.02 |
| 96 | 269.03 |
| 4 | 269.03 |
| 429 | 269.03 |
| 376 | 269.03 |
| 415 | 269.03 |
| 360 | 269.03 |
| 365 | 269.03 |
| 88 | 269.03 |
| 6 | 269.03 |
| 362 | 269.03 |
| 371 | 269.03 |
| 449 | 269.03 |
| 383 | 269.03 |
| 337 | 269.03 |
| 104 | 269.03 |
| 2 | 269.03 |
| 355 | 269.03 |
| 391 | 269.03 |

Результаты по метрикам:

* Mean Squared Error: 26166.644
* R^2 Score: 0.0059
* Mean Absolute Error: 152.234
* Root Mean Squared Error: 161.761
* Mean Absolute Percentage Error: 1294.18%

На рисунке 20, можно заметить, что предсказанные значения, сильно отличаются от реальных и это значение равняется 269.03, имея лишь в начале низкое значение, равное 199.88. Это говорит о том, что модель не способна давать адекватных результатов предсказания.



Рис. 20 - Сравнение предсказанных и истинных значений модели ARIMA.

Сравнение результатов работы моделей по метрикам приведены ниже в таблице 13:

Таблица 13

| Метрика | Gradient Boosting | XGBoost | LightGBM | CatBoost | LSTM | ARIMA |
| --- | --- | --- | --- | --- | --- | --- |
| RMSE | 39.3326 | 35.4011 | 40.294 | **34.9799** | 64.43 | 161.761 |
| R^2 Score | 0.9412 | 0.9523 | 0.9383 | **0.9535** | 0.8293 | 0.0059 |
| MAE | 28.5241 | **25.3453** | 28.3283 | 26.5099 | 34.8734 | 152.2346 |
| MAPE | 57.5636% | 41.6932% | 52.7615% | 85.4983% | **22.8%** | 1294.18% |

Описывая результаты каждой модели, исходя из таблицы сравнения, стоит подробно остановиться на каждой модели:

1. Модель Gradient Boosting

Модель показала себя хорошо на тестовых данных, не имея при этом выбросов, она справляется со всеми предсказаниями числа пациентов за каждый день октября 2023 года. Однако по сравнению с другими моделями, результаты по метрикам показывают, что она хуже других и имеет высокое значение среднего процента ошибки по метрике MAPE.

1. Модель ARIMA

Данная модель показала наихудший результат среди всех других, в задачи прогнозирования количества пациентов. Результаты получились плохими как по предсказаниям, так и по метрикам. Модель оказалась не способна выдавать адекватные значения по прогнозам как на будние, так и выходные дни. Модель имеет самое высокое среднего процента ошибки, а также самое низкое значение по метрике R^2 Score.

1. Нейронная сеть LSTM

Нейронная сеть показала хорошие результаты по всем метрикам, относительно других моделей. Она имеет наименьший средний процент ошибки MAPE (22.8%) среди всех остальных моделей. Однако имеются выбросы в предсказаниях (это означает, что предсказанное значение меньше 0), что критически важно при прогнозировании числа пациентов, это является неудовлетворительным результатом, поскольку отрицательное число пациентов не может существовать. Такой результат может быть связан с переобучением модели, однако, смотря на график обучения модели, видно что она не переобучается и значение loss получилось низким, как ранее было отмечено, результаты по метрикам также хорошие, и несмотря на это, нейронная сеть не смогла найти зависимость в определенные дни и выдала отрицательные значения по предсказаниям.

1. Модель LightGBM

Модель имеет средние результаты работы по метрикам относительно других моделей, однако стоит отметить достаточно высокое значение по метрике MAPE (52.761%). Модель LightGBM, как и нейронная сеть при тестировании на данных месяца октября 2023 года получила выбросное значение, как ранее было отмечено, это критически важно при прогнозировании числа пациентов. Несмотря на результаты по метрикам и адекватному процессу обучения, модель не справилась и не смогла определить зависимость в определенный день, что привело к отрицательному значению.

1. Модель XGBoost

Модель показала хорошие результаты по всем метрикам, а также имеет наилучший результат по метрике MAE, имея самое низкое значение (25.345). Также при тестировании модели, не было обнаружено выбросов, а предсказанные значения, получились близкими к реальным.

1. Модель CatBoost

Модель показала наилучшие результаты по метрикам RMSE (34.979) и R^2 Score (0.9535), среди всех остальных моделей, однако имеет высокое значение по метрике MAPE (85.498%), тем не менее это не сильно отразилось на результатах предсказания за октябрь 2023 года. Выбросные значения не были обнаружены, однако по сравнению с моделью XGBoost, эмпирическим путем было обнаружено, что предсказанные значения в выходные дни были выше, чем в предсказаниях модели XGBoost.

Рассматривая более подробно:

CatBoost:

2023-10-01 2.0 17.29

2023-10-08 6.0 12.0

2023-10-15 4.0 17.92

2023-10-22 6.0 3.72

2023-10-29 2.0 25.51

XGBoost:

2023-10-01 2.0 5.17

2023-10-08 6.0 2.59

2023-10-15 4.0 12.19

2023-10-22 6.0 8.47

2023-10-29 2.0 14.44

В данном случае, это не столь критично и не сильно влияет на конечный результат. Таким образом, исходя из сравнения результатов моделей, конечный выбор стоял между моделью XGBoost и CatBoost. Далее проводилось тестирование моделей в условиях, когда не были известны данные по посещениям за будущие дни, однако проверялась работоспособность моделей, способность улавливать нелинейные зависимости на более поздние периоды времени. Тестирование проводилось на климатических данных за период декабря 2024 года.

Результаты за 2024 год, декабрь модели XGBoost представлены в таблице 14:

Таблица 14

| Дата | Предсказанное значение |
| --- | --- |
| 2024-12-01 | -24.28 |
| 2024-12-02 | 328.85 |
| 2024-12-03 | 336.89 |
| 2024-12-04 | 286.17 |
| 2024-12-05 | 281.71 |
| 2024-12-06 | 240.36 |
| 2024-12-07 | 111.24 |
| 2024-12-08 | 16.23 |
| 2024-12-09 | 326.67 |
| 2024-12-10 | 337.23 |
| 2024-12-11 | 339.98 |
| 2024-12-12 | 305.42 |
| 2024-12-13 | 277.47 |
| 2024-12-14 | 73.97 |
| 2024-12-15 | 9.67 |
| 2024-12-16 | 244.55 |
| 2024-12-17 | 330.88 |
| 2024-12-18 | 314.61 |
| 2024-12-19 | 255.98 |
| 2024-12-20 | 243.79 |
| 2024-12-21 | 95.18 |
| 2024-12-22 | -25.55 |
| 2024-12-23 | 298.71 |
| 2024-12-24 | 293.15 |
| 2024-12-25 | 269.05 |
| 2024-12-26 | 285.41 |
| 2024-12-27 | 239.11 |
| 2024-12-28 | 63.01 |
| 2024-12-29 | -22.86 |
| 2024-12-30 | 272.3 |

Результаты за 2024 год, декабрь модели CatBoost представлены в таблице 15:

Таблица 15

| Дата | Предсказанное значение |
| --- | --- |
| 2024-12-01 | 3.51 |
| 2024-12-02 | 356.41 |
| 2024-12-03 | 356.45 |
| 2024-12-04 | 320.29 |
| 2024-12-05 | 320.23 |
| 2024-12-06 | 317.64 |
| 2024-12-07 | 54.14 |
| 2024-12-08 | -25.69 |
| 2024-12-09 | 343.75 |
| 2024-12-10 | 352.79 |
| 2024-12-11 | 331.65 |
| 2024-12-12 | 311.96 |
| 2024-12-13 | 307.13 |
| 2024-12-14 | 74.05 |
| 2024-12-15 | 12.88 |
| 2024-12-16 | 302.07 |
| 2024-12-17 | 353.35 |
| 2024-12-18 | 324.53 |
| 2024-12-19 | 297.99 |
| 2024-12-20 | 324.46 |
| 2024-12-21 | 93.46 |
| 2024-12-22 | 17.87 |
| 2024-12-23 | 327 |
| 2024-12-24 | 309.5 |
| 2024-12-25 | 274.08 |
| 2024-12-26 | 325.83 |
| 2024-12-27 | 282.21 |
| 2024-12-28 | 68.15 |
| 2024-12-29 | 8.62 |
| 2024-12-30 | 319.91 |

Как видно из результатов, модель XGBoost, справилась хуже, поскольку в результате определила больше выбросов (3), чем модель CatBoost (1 выброс). Такие результаты, скорее всего говорят о том, что модель XGBoost, хуже определяла зависимость между признаками модели. Вероятно это связано с иными зависимостями, которые характерны для этого периода времени. Каких конкретно, этого точно сказать нельзя, поскольку это может быть обусловлено разными факторами, на это могли повлиять не только климатические условия. Причина таких результатов также может быть связана с нехваткой данных, аномальных значений, слабой связанности (отсутствие зависимости), что в приводит к отрицательным предсказаниям.

MAPE XGBoost (41,69%) намного ниже, чем MAPE CatBoost (85,49%), что говорит о том, что XGBoost имеет лучшую общую точность с точки зрения процентной ошибки.

Однако MAPE может вводить в заблуждение при работе с малыми или нулевыми значениями целевой переменной, поскольку он чувствителен к таким значениям (поскольку это процентная ошибка). Если фактическое количество посетителей бывает низкое (близкое к нулю), MAPE может стать завышенным, создавая впечатление, что модель работает плохо, даже когда абсолютная ошибка низкая.

Несмотря на то, что XGBoost лучше справляется с метриками обучения, CatBoost, по-видимому, лучше справляется с реальными тестовыми данными на основе прогнозов. Это говорит о том, что CatBoost может лучше обобщать, возможно, потому, что он более эффективно обрабатывает базовые закономерности в данных. Более низкий MAPE для XGBoost, скорее всего, не обязательно означает, что он окажется лучше.

В качестве временного решения, отмечать данные выбросы аномальными, приводя результат к нулю, вместо отрицательного значения. В будущем, потребуется рассмотреть результаты с большим количеством данных, расширить их, чтобы впоследствии дообучить модель, возможно это приведет к улучшению результатов.

В результате была выбрана модель CatBoost для прогнозирования количества пациентов первичного приема на будущие дни. Однако, исследование показало невозможность предсказания количества пациентов только за счет климатических данных, поскольку на графиках важности признаков обучения моделей было отражено, что ключевым признаком в предсказании был день недели. Без этого признака, модель была не в состоянии выдавать адекватные результаты по метрикам. Остальные же признаки имели низкие значения значимости. Тем не менее, в ходе эмпирического анализа, было выявлено, что остальные признаки корректируют значение предсказаний модели, что отражает реальные результаты работы, и за счет этих признаков, также корректировалась оценка моделей по метрикам.

#### ГЛАВА 4. ТЕСТИРОВАНИЕ

##### 4.1 Тестирование модели

Ввиду специфики проекта тестирование предполагает тестирование модели на основе реальных данных, которые не были использованы в процессе обучения. Фокус группа была определена в виде авторов работы, а также 8-и студентов университета ТюмГУ направления МОиАИС.

**Имеется:**

* Данные по посещениям пациентов за октябрь 2023 года
* Климатические данные за октябрь 2023 года

**Ожидаемый результат:**

* Прогноз числа пациентов моделью, должны быть максимально близкими к реальному количеству посетителей за конкретный день

**Полученный результат:**

* Модель хорошо справляется с большей частью предсказаний числа пациентов за каждый день октября 2023 года, однако, имеются дни, предсказания в которых имеют сильную погрешность.

**Причины:**

* Это может быть связано с нехваткой данных, аномальными значениями в датасете для обучения и отсутствие зависимости.

**Способы решение:**

* В качестве временного решения, отмечать данные выбросы аномальными, приводя результат к нулю
* В будущем, рассмотреть данный случай с большим кол-во данных для обучения  
   Результаты посещений за октябрь 2023 года представлены в таблице 16:

Таблица 16

| Дата | Реальное кол-во | Предсказанное кол-во |
| --- | --- | --- |
| 2023-10-01 | 2 | 17.29 |
| 2023-10-02 | 435 | 375.1 |
| 2023-10-03 | 376 | 382.19 |
| 2023-10-04 | 409 | 381.81 |
| 2023-10-05 | 368 | 370.49 |
| 2023-10-06 | 386 | 350.81 |
| 2023-10-07 | 78 | 88.33 |
| 2023-10-08 | 6 | 12 |
| 2023-10-09 | 428 | 381.09 |
| 2023-10-10 | 417 | 374.1 |
| 2023-10-11 | 410 | 370.48 |
| 2023-10-12 | 408 | 375.13 |
| 2023-10-13 | 402 | 352.24 |
| 2023-10-14 | 96 | 108.24 |
| 2023-10-15 | 4 | 17.92 |
| 2023-10-16 | 429 | 400.11 |
| 2023-10-17 | 376 | 405.72 |
| 2023-10-18 | 415 | 384.95 |
| 2023-10-19 | 360 | 372.52 |
| 2023-10-20 | 365 | 355.92 |
| 2023-10-21 | 88 | 73.56 |
| 2023-10-22 | 6 | 3.72 |
| 2023-10-23 | 362 | 384.89 |
| 2023-10-24 | 371 | 364.42 |
| 2023-10-25 | 449 | 331.06 |
| 2023-10-26 | 383 | 338.15 |
| 2023-10-27 | 337 | 355.51 |
| 2023-10-28 | 104 | 99.72 |
| 2023-10-29 | 2 | 25.51 |
| 2023-10-30 | 355 | 397.35 |
| 2023-10-31 | 391 | 404.25 |

Стоит отметить, что модель для некоторых прогнозов сильно занижает, либо сильно завышает результат. Стоит отметить, что модель достаточно часто предсказывает близкие к реальным значениям.

##### 4.2 Тестирования пользовательского интерфейса

Для пользовательского интерфейса установлена следующая процедура тестирования:

**Ожидаемый результат:**

* При нажатии на кнопку “Получить прогноз” и выбранном интервале для предсказания, должен быть отображен прогноз, содержащий: разница предсказанного числа пациентов и среднего количества пациентов (рассчитывается на основе данных из датасета), предсказанное число пациентов, а также дата и день недели, для которых был выполнен прогноз.

**Полученный результат:**

* Пользовательский интерфейс корректно отобразил запрошенный прогноз

##### 4.3 Процедура тестирования серверной части

Для серверной части установлены следующая процедура тестирования:

**Ожидаемый результат:**

* При обращении на эндпоинт для получения прогнозных данных, должны быть получены следующие данные: дата дня, для которого был выполнен прогноз и спрогнозированное кол-во пациентов

**Полученный результат:**

* Серверная часть проекта предоставляет корректные данные в установленном формате.

##### 4.4 Результаты тестирования

На основе результатов тестирования можно отметить следующее:

* Серверная часть web-приложения исправно исполняет установленные требования
* Пользовательский интерфейс исправно исполняет установленные требования
* Модель, в большинстве случаев, исправно исполняет установленные требования, однако требуется обратить внимание на аномальные прогнозные результаты

#### ЗАКЛЮЧЕНИЕ

Было разработано web-приложение, которое, при помощи модели CatBoost, поможет руководителям медицинских учреждений заранее подготовиться к повышенным нагрузкам, выделить необходимое количество персонала на будущие дни, тем самым сохранить качество обслуживания пациентов.  
 Также стоит отметить исследование, которое показало, что прогнозирование количества пациентов первичного приема, только за счет климатических данных невозможно, так как есть более существенные признаки, влияющие на предсказание (День недели).  
 Рассматривая проект с перспективной точки зрения, можно отметить потенциал для дальнейшего расширения и улучшения функционала, а также расширение данных для дообучения модели.   
 Таким образом, разработанное web-приложение представляет собой инструмент, который имеет большой потенциал для будущего развития и применения на рынке здравоохранения и медицинских систем с возможностью интеграции в уже существующие информационные системы.

#### СПИСОК ЛИТЕРАТУРЫ

1. Министерство здравоохранения Кузбасса, Михаил Мурашко: здравоохранение с приходом пандемии научилось работать оперативно. Итоговое интервью телеканалу Россия 24, URL: <https://kuzdrav.ru/activity/news/34603/>
2. Федеральная служба государственной статистики, URL: <https://rosstat.gov.ru/folder/13721>
3. Effects of extreme temperatures on hospital emergency room visits for respiratory diseases in Beijing, China / Yuxia Ma, Jianding Zhou, College of Atmospheric Sciences, 2018
4. Air pollution and emergency department visits for cardiac and respiratory conditions: a multi-city time-series analysis / David M Stieb [и др.], Health Canada, 2009
5. Climate change and human health: Impacts, vulnerability and public health / A. Haines [и др.], London School of Hygiene and Tropical Medicine, 2005
6. Assessing Health Vulnerabilities and Adaptation to Climate Change: A Review of International Progress / Peter Berry [и др.], University of Waterloo, 2018