Информация об авторах и публикации		
	На русском языке	На английском языке
Фамилия	Кригер	Kriger
Имя	Александра	Alexandra
Отчество	Борисовна	Borisovna
Учёная степень	кандидат физико-математических наук	Candidate of Physical and Mathematical Sciences
Учёное звание	доцент	Associate Professor
Место работы	Научно-образовательный центр «Искусственный интеллект» Владивостокского государственного университета, Владивосток, Российская Федерация	Scientific and Educational Center "Artificial Intelligence" of Vladivostok State University, Vladivostok, Russian Federation
должность	доцент	Associate Professor
E-mail	aleksandra.kriger@vvsu.ru	aleksandra.kriger@vvsu.ru
Фамилия	Клановец	Klanovets
Имя	Артём	Artyom
Отчество	Дмитриевич	Dmitryevich
Учёная степень	магистр	master
Учёное звание	нет	no
Место работы	Научно-образовательный центр «Искусственный интеллект» Владивостокского государственного университета, Владивосток, Российская Федерация	Scientific and Educational Center "Artificial Intelligence" of Vladivostok State University, Vladivostok, Russian Federation
должность	аспирант	Postgraduate student
E-mail		
Адрес для отправки журнала с ИНДЕКСОМ		
Название статьи	Прогнозирование расходов муниципальных бюджетов на благоустройства методами машинного обучения	Forecasting municipal budget expenditures on improvements by machine learning methods
Аннотация	*Благоустройство муниципалитетов (муниципалитет: город, поселок, сельское поселение или несколько таких территорий) является необходимым условием для их развития. Данная задача является особенно важной для муниципальных округов Дальневосточного федерального округа (далее ДФО). Среди городов дальневосточного региона нет «милионников»,*	*Improvement of municipalities (municipality: city, town, rural settlement or several such territories) is a necessary condition for their development. This task is especially important for municipal districts of the Far Eastern Federal District (hereinafter FEFD). Among the cities of the Far Eastern region there are no "millionaires", the infrastructure is not sufficiently developed, the level of population outflow to other regions is high.*

	инфраструктура развита не достаточно, высок уровень оттока населения в другие регионы. Под благоустройством понимается создание комфортной и функциональной среды в населённых пунктах. Оно включает в себя работы по озеленению, инженерные решения, архитектурное оформление и организационные меры. *Целью данной работы является создание модели для прогнозирования необходимого объема финансирования проектов по благоустройству территорий. Анализ закономерностей и моделирование выполнены на примере данных Приморского края.* *В результате исследования: выявлены факторы, в наибольшей степени определяющие расходы на благоустройство, на основе методов машинного обучения построены и протестированы модели прогнозирования расходов. Для моделей подобраны оптимальные параметры. Ожидаемо, лучшие результаты дали ансамблевые модели, построенные на основе нейронных сетей.*	*Improvement is understood as the creation of a comfortable and functional environment in populated areas. It includes landscaping work, engineering solutions, architectural design and organizational measures.* *The purpose of this work is to create a model for forecasting the required amount of funding for projects on improvement of territories. Analysis of patterns and modeling are performed on the example of data from Primorsky Krai.* *As a result of the study: the factors that determine the costs of improvement to the greatest extent were identified, models for forecasting costs were built and tested based on machine learning methods. Optimal parameters were selected for the models.* *As expected, the best results were obtained by ensemble models built on the basis of neural networks.*
Ключевые слова	методы машинного обучения, прогнозирование расходов, муниципальные бюджеты, благоустройство, городская среда	machine learning methods, cost forecasting, municipal budgets, improvement, urban environment

Научная статья
УДК 332.14

Кригер Александра Борисовна
кандидат физико-математических. наук, доцент
доцент Научно-образовательного центра «Искусственный интеллект» Владивостокского государственного университета, Владивосток, Российская Федерация
aleksandra.kriger@vvsu.ru

Kriger Alexandra Borisovna
Candidate of Physical and Mathematical Sciences, Associate Professor
Associate Professor, Scientific and Educational Center "Artificial Intelligence", Vladivostok State University, Vladivostok, Russian Federation
aleksandra.kriger@vvsu.ru

Клановец Артем Дмитриевич
аспирант по направлению 09.04.03 «Прикладная информатика»
Научно-образовательного центра «Искусственный интеллект» Владивостокского государственного университета, Владивосток, Российская Федерация
Klanovets.AD@vvsu.ru

Klanovets Artem Dmitrievich
Postgraduate student in the direction 09.04.03 "Applied Informatics"
Graduate of the Scientific and Educational Center "Artificial Intelligence" of Vladivostok State University, Vladivostok, Russian Federation
Klanovets.AD@vvsu.ru

-

ПРОГНОЗИРОВАНИЯ РАСХОДОВ МУНИЦИПАЛЬНЫХ БЮДЖЕТОВ НА БЛАГОУСТРОЙСТВО МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

5.2.3 – Региональная и отраслевая экономика

Research Article

FORECASTING MUNICIPAL BUDGETS EXPENDITURES ON IMPROVEMENTS BY MACHINE LEARNING METHODS

5.2.3 – Regional and sectoral economics

Благоустройство муниципалитетов (муниципалитет: город, поселок, сельское поселение или несколько таких территорий) является необходимым условием не только для их развития, но и для снижения оттока населения. Городская среда представляет собой сложную и многогранную систему.

Под благоустройством понимается комплекс инженерных, архитектурных, озеленительных и организационных мер, направленных на улучшение внешнего облика, функциональности и экологической обстановки в населённых пунктах. Объекты благоустройства выступают важнейшими компонентами городской инфраструктуры и играют значительную роль в обеспечении высокого качества жизни населения.

Система бюджетирования в Российской Федерации состоит из федерального, регионального и муниципального уровней, обеспечивает комплексное финансирование государственных и муниципальных задач, включая развитие городской среды и благоустройство территорий. Расходы на благоустройство составляют не только финансовые ресурсы собственно муниципальных бюджетов, но финансовые потоки федерального бюджета (различные программы по благоустройству, межбюджетные трансферты на конкретные проекты).

Последние годы особое внимание в государственной политике уделено вопросам устойчивого развития городских территорий и повышения качества жизни населения. Одним из ключевых направлений стал проект «Жильё и городская среда» [1, 2]. Ключевые задачи проекта: приведение состояния дворовых территорий в соответствие с установленными нормами, развитие системы общественного участия в выборе приоритетных объектов благоустройства, повышение уровня экологичности, безопасности и доступности городских пространств.

При всех положительных тенденциях, выделенные ресурсы не всегда используются в полной мере или расходуются неэффективно. Так среднее значение индекса качества городской среды для Приморского края за 2024 год составляет всего 188 [3], а это достаточно низкий уровень (в Сахалинской области индекс составляет 199, так же в Хабаровском крае 201, Московская область – 238).

Успешное выполнение проектов требует четкой координации между уровнями власти, привлечения внебюджетных источников финансирования, использования современных технологий и, самое главное, активного участия самих горожан. Эффективное прогнозирование и планирование бюджетных расходов, основанное на анализе прошлых данных и реалистичных оценках, является ключевым фактором устойчивости и долгосрочной успешности реализации проектов благоустройства.

Данное исследование, имеет своей целью решить проблему прогнозирования необходимого объема финансирования для проектов благоустройства территорий. Исследование проводилось на примере данных Приморского края.

Для достижения указанной цели решены следующие задачи:

- собраны и трансформированы данные об объектах благоустройства муниципалитетов Приморского края;
- выявлены факторы, влияющие на уровень расходов на благоустройство отдельных объектов городской среды;
- построены и протестированы модели на основе методов машинного обучения.

Основной задачей данного исследования авторы видят в ответе на вопрос: возможно ли использовать методы машинного обучения для оценки необходимого уровня расходов на благоустройство муниципалитетов с достаточной степенью точности

Публикации с результатами исследований по прогнозированию параметров бюджетов методами машинного обучения немногочисленны (пример [4], [5], [6]) и они не затрагивают прогнозы расходов бюджетов на благоустройство территорий.

Результатом данной работы является интеллектуальная модель, которая позволит прогнозировать уровень бюджетных расходов муниципалитетов на благоустройство.

Необходимо заметить, что применение таких моделей требует существенной работы по поиску и подготовке данных для обучения. Таким образом, другим результатом нашего исследования является *создание двух рабочих датафреймов для анализа закономерностей и обучения моделей*.

Исходные данные, источники данных, отбор наблюдений для формирования датафреймов

Объекты благоустройства, которые рассматривались для создания датафрейма включают в себя следующие элементы городской планировки:

1) Скверы;
2) Парки;
3) Набережные;
4) Площади.

Принципиальными характеристиками объектов являются:

1) Местоположение (Рядом с побережьем, в центре города);
2) Площадь;
3) Назначение.

Прогнозирование расходов муниципальных бюджетов на благоустройство требует использования разнообразных источников информации, которые позволяют не только оценить объём выделенных и фактических средств, но и учесть пространственные, функциональные и проектные особенности объектов благоустройства. Формирование качественного датафрейма является ключевым этапом построения моделей прогнозирования, поскольку от полноты и точности исходных данных зависит репрезентативность результатов и надёжность принимаемых решений.

Для целей исследования были выбраны следующие основные источники информации:

- Единый портал бюджетной системы — источник данных по бюджетным ассигнованиям и их исполнению [7];
- Государственный кадастровый реестр — источник непространственной и количественной информации об объектах благоустройства [8];
- Сайт городской администраций города Владивосток — источник данных по работам над конкретными объектами благоустройства города Владивостока [9];
- Сайт городской администраций города Уссурийск — источник данных по работам над конкретными объектами благоустройства города Уссурийска [10];
- Сайт городской администраций города Артём — источник данных по работам над конкретными объектами благоустройства города Артёма [11];
- Сайт городской администраций города Находка — источник данных по работам над конкретными объектами благоустройства города Находка [12];
- Официальный сайт государственных закупок — источник данных о контрактах, подрядчиках и финансировании конкретных объектов благоустройства [13].

Эти источники обеспечивают комплексное описание как финансовых, так и пространственно-функциональных характеристик объектов благоустройства, что делает возможным анализ тенденций, построение аналитических моделей и прогнозирование бюджетных затрат. Для формирования пригодного для анализа датафрейма, нужно определить, какие показатели важны для определения объема бюджетного финансирования.

Поскольку данные по общим расходам на благоустройство и конкретным расходам на объекты трудно объединить вместе, поскольку неизвестно, по каким конкретно объектам проводились работы в тот или иной день, было принято решение сформировать два датафрейма:

1) Датафрейм общего расхода бюджетов по муниципалитету;
2) Датафрейм с расходами на конкретный объект благоустройства за год.

Таким образом, необходимо определить показатели для двух датафреймов.

Первый датафрейм является временным рядом с двумя показателями: плановый и фактически исполненный бюджет. Этих показателей будет достаточно для того, чтобы доказать факт того, что расходы на благоустройство не являются случайными и в этих расходах имеется историчность (инерционность) данных.

Второй датафрейм требует более подробного пояснения, как следует отбирать для него показатели. Для определения этих показателей следует использовать технические задания и сметы проектов по благоустройству.

Поскольку документы, прикладываемые к тендеру (пример документа [14]), могу различаться в плане описания проводимых работ, их подробностях описания и обозначениям

для того или иного типа работ, необходимо найти общие совпадения по работам и использовать только «универсальные» виды работ. В качестве показателей можно взять пункты сметной документации и сделать их бинарными (был ли или не был проведён данных тип работ).

Таким образом, на основании анализа проектных документов, был выделен следующий список показателей:

1) Выделенный бюджет на благоустройство;
2) Площадь благоустроенной территории;
3) Год благоустройства;
4) Местность благоустройства;
5) Благоустройство пешеходных тропинок;
6) Строительство детских площадок;
7) Строительство спортивных площадок;
8) Строительство освещения;
9) Озеленение;
10) Строительство фонтанов;
11) Строительство парковок;
12) Строительство доступной среды для инвалидов;
13) Размещение малых архитектурных форм;
14) Сезонный уход
15) Обслуживание территории

Первый датафрейм был собран исключительно с использованием данных с сайта [7], так как располагает данными по всем необходимым показателям. Ресурсы городских администраций были использованы для проверки данных. В итоге был получен датафрейм по четырём городам Приморского края:

1) Владивосток
2) Уссурийск
3) Артём
4) Находка

Эти муниципалитеты были выбраны потому, что именно в городах находится наибольшее количество объектов благоустройства в Приморского края. Таким образом, получилось собрать датафрейм объемом 290 наблюдений (см. таблицу 1).

Таблица 1 – Показатели первого датафрейма

Название показателя	Описание показателя	Тип показателя

date	Дата фиксирования	Время
city	Город	Категориальный
plan	Плановый бюджет	Непрерывный
done	Фактический бюджет	Непрерывный
differ	Разница между следующим и предыдущим месяцем	Непрерывный

В итоговый вариант первого датафрейма включено 5 переменных: 4 получены из отчетов (date, city, plan, done, differ); пятая (differ) вычислена как разница между значением за текущий и предыдущий месяцы, т.к. в отчетах показатель представлен нарастающим итогом.

Второй датафрейм был собран на основе интернет-ресурса [13], откуда были извлечены данные по работам над объектами благоустройства, а также документация, прикреплённая к работам. Анализ данной документации позволил сопоставить её с теми показатели, которые были определены для датафрейма. Таким образом, после сборки датафрейма, был получен набор данных размеров 102 объекта (строки) с количеством показателей 18. В таблице 2 представлены используемые переменные (показатели) и их описание.

Таблица 2 – Показатели второго датафрейма

Название показателя	Описание показателя	Тип показателя
city	Город объекта	Категориальный
park_name	Название объекта	Идентификатор
park_type	Тип объекта (парк, сквер, площадь)	Категориальный
Address	Адрес объекта	Идентификатор
year_renovated	Год благоустройства	Дата
area_sqm	Площадь	Непрерывный
has_slope	Наличие неровностей ландшафта	Бинарный
has_pedestrian_walkaways	Благоустройство пешеходных дорожек	Бинарный
has_children_area	Благоустройство детских площадок	Бинарный
has_sports_area	Благоустройство спортивных площадок	Бинарный
has_lighting	Благоустройство освещения	Бинарный
has_greenery	Озеленение	Бинарный
has_water_features	Благоустройство фонтанов	Бинарный
has_parking	Благоустройство парковки	Бинарный
has_accessible_environment	Благоустройство доступной	Бинарный

	среды	
has_maf	Благоустройство объектов малой архитектуры	Бинарный
has_seasonal_maintance	Сезонное обслуживание	Бинарный
Budget_total_rub	Расходы на объект	Непрерывный

Выбор инструмента, анализ качества данных и закономерностей

Анализ данных в современных исследованиях стал неотъемлемой частью процесса принятия обоснованных решений, особенно в задачах прогнозирования. Эффективность такого анализа во многом зависит от выбора подходящего программного обеспечения и вычислительных инструментов, которые обеспечивают не только точность и надёжность обработки информации, но и гибкость, масштабируемость и воспроизводимость результатов.

В рамках настоящего исследования был выбран язык программирования Python [15], который на сегодняшний день является одним из самых популярных и функциональных инструментов в области анализа данных и машинного обучения.

Для выполнения задач предварительной обработки, визуализации, статистического анализа и построения моделей прогнозирования были использованы следующие библиотеки Python:

NumPy — обеспечивает поддержку многомерных массивов и матричных операций [16];

Pandas — используется для работы с табличными данными [17];

Matplotlib.pyplot и Seaborn — позволяют строить графики и диаграммы для наглядного представления полученных результатов;

SciPy.stats и Statsmodelstsa.stattools — предоставляют набор статистических тестов и функций, необходимых для оценки закономерностей и значимости выявленных взаимосвязей (в частности тест Дики – Фуллера).

Документация для указанных библиотек доступна в открытом доступе на различных ресурсах.

Анализ датафрейма с динамическими данными (расходы бюджета)

Датафрейм с динамическими данными создан для оценки принципиальной возможности использовать данные по расходам бюджетов для прогнозирования. Для этой цели необходимо исключить гипотезу о так называемом «случайном блуждании». Моделью случайного блуждания (Random Walk) — стохастический процесс, в котором каждое следующее значение формируется как предыдущее значение плюс случайная величина:

$$y_t = y_{t-1} + \varepsilon_t$$

где y_t— значение переменной в момент времени t, ε_t— белый шум.

Случайное блуждание относится к классу нестационарных процессов, поскольку его дисперсия со временем возрастает, а математическое ожидание остаётся постоянным. Это свойство создает серьёзные трудности при попытках моделирования и прогнозирования динамических показателей. В данном случае, прогноз определяется как $\hat{y}_{t+1} = y_t$, что сводит его ценность к минимуму.

Если временной ряд бюджетных расходов на благоустройство соответствует гипотезе случайного блуждания, это означает, что финансирование объектов благоустройства зависит от случайных факторов, не подчиняется никаким долгосрочным закономерностям, не может быть приведён к стационарному виду. Под стационарностью понимается, что математическое ожидание, дисперсия и автоковариационная функция ряда — не меняются со временем.

Для оценки характера динамического ряда авторы использовали автокорреляционную функцию и тест Дики–Фуллера.

Анализ автокорреляционной функции позволяет выявить:

1) скрытые закономерности: тенденции, сезонность или цикличность в данных, порядок цикличности;

2) степень случайности колебаний: если автокорреляция близка к нулю на всех лагах, значит, данные могут быть случайными (например, следовать случайному блужданию).

На рисунках 1, 2 и 3 представлены автокорреляционные функции.

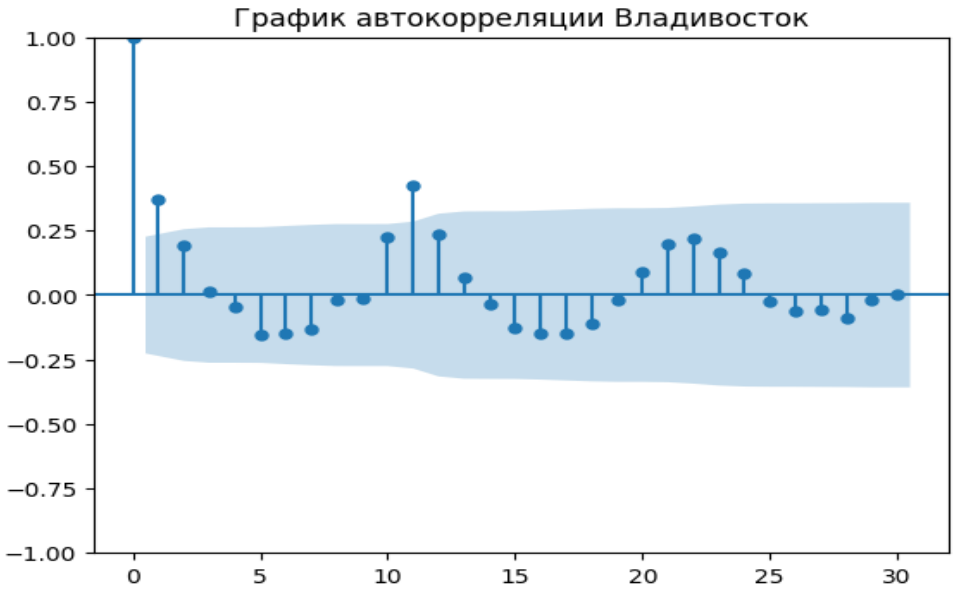


Рисунок 1 – график автокорреляционой функции для Владивостока

Рисунок 2 – график автокорреляционой функции для Артёма

Рисунок 3 – график автокорреляционой функции для Уссурийска

Как очевидно из графиков, расходы бюджетов во Владивостоке и Артёме обладают явной цикличностью. Для Уссурийска цикличность явно не выражена. Анализ данных графиков, позволяет сделать вывод, что временные ряды являются нестационарными с цикличностью («сезонностью») и очевидно не являются случайным блужданием.

Для дальнейшего анализа стационарности ряда и проверки наличия случайного блуждания были выполнены тесты Дики–Фуллера.

Тест Дики–Фуллера (Augmented Dickey-Fuller Test, ADF-тест) — это один из ключевых статистических тестов, используемых для анализа временных рядов [18]. Он позволяет определить, является ли временной ряд стационарным или содержит единичный

корень, что указывает на его нестационарность и потенциальную зависимость от времени. Для прогнозировании важно учитывать характер поведения временного ряда. Если ряд не стационарен (например, имеет тренд или сезонность), то любые модели, построенные без учёта этого факта, могут давать смещённые и ненадёжные прогнозы.

В нашем исследовании ADF-тест используется:

- для проверки гипотезы о наличии единичного корня в авторегрессионном процессе, чтобы убедиться, что данные имеют закономерную динамику и не являются случайным блужданием
- для принятия решения о необходимости дифференцирования ряда перед моделированием;

Тесты выполнены для данных каждого из городов: классический тест (без константы и тренда), с константой, с трендом. В таблице 3 приведены результатов необходимых тестов (p-value) для трёх крупных городов Приморского края.

Таблица 3 – Результаты тестов Дикки-Фуллера

Город	ADF тест без константы и тренда	ADF тест с константой	ADF тест с линейным трендом	ADF тест с квадратичным трендом
Владивосток	0.847169	0.000535	0.000317	0.002600
Артём	0.155064	0.000000	0.000000	0.068901
Уссурийск	0.211062	0.000854	0.004911	0.022343

Результаты теста демонстрируют, что для моделей с константой или линейным трендом p-value меньше 0.05, следовательно ряд можно привести к стационарному виду (в данном случае просто используя первую разность). Таким образом, можно сделать вывод, данные по расходам на благоустройство обладают историчность, на их основе можно построить модель прогнозирования.

Построение алгоритма машинного обучения

«Машинное обучение» включает широкий спектр моделей и алгоритмов: от классических методов прикладной статистики до ансамблевых моделей глубокого обучения.

Для целей настоящего исследования использовалась библиотека scikit-learn (sklearn), которая предоставляет широкий набор функций для подготовки данных, построения моделей, их оценки и сравнения. Основным критерием выбора именно этой библиотеки стало её удобство использования, хорошее документирование и высокое качество реализации стандартных алгоритмов машинного обучения.

Для анализа прогностической ценности построенных моделей исходный набор данных случайным образом разбивали на две части (обучающую и тестовую) в заданном соотношении (например, 80% на обучение и 20% на тестирование). Такой подход обеспечивает объективную оценку качества модели на «новых», участвующих в обучении данных.

Для оценки устойчивости результатов и качества модели использовались методы кросс-валидации. В работе применялись два варианта:

1) алгоритм, реализующий разделение данных на K фолдов, где модель обучается на K–1 частях и тестируется на одной. Это позволяет получить усреднённую оценку эффективности модели.

2) алгоритм реализующий сохранения пропорций в каждом фолде. Используется в случае, когда целевая переменная имеет выраженную неоднородность или распределение.

Основной метрикой, использованной в исследовании, стал коэффициент детерминации R^2 (R-squared score). Метрика позволяет точно оценить адекватность каждой модели перед её финальным тестированием на отложенной выборке.

Для построения прогноза бюджетных расходов на благоустройство были выбраны три различных типа моделей. Выбор моделей был обусловлен стремлением сравнить как «базовые», так и более сложные методы, чтобы определить наиболее эффективный подход для данной задачи. Выбор указанных методов машинного обучения связан с изначально достаточно короткой выборкой, и отсутствием предположений о нелинейных взаимодействиях параметров.

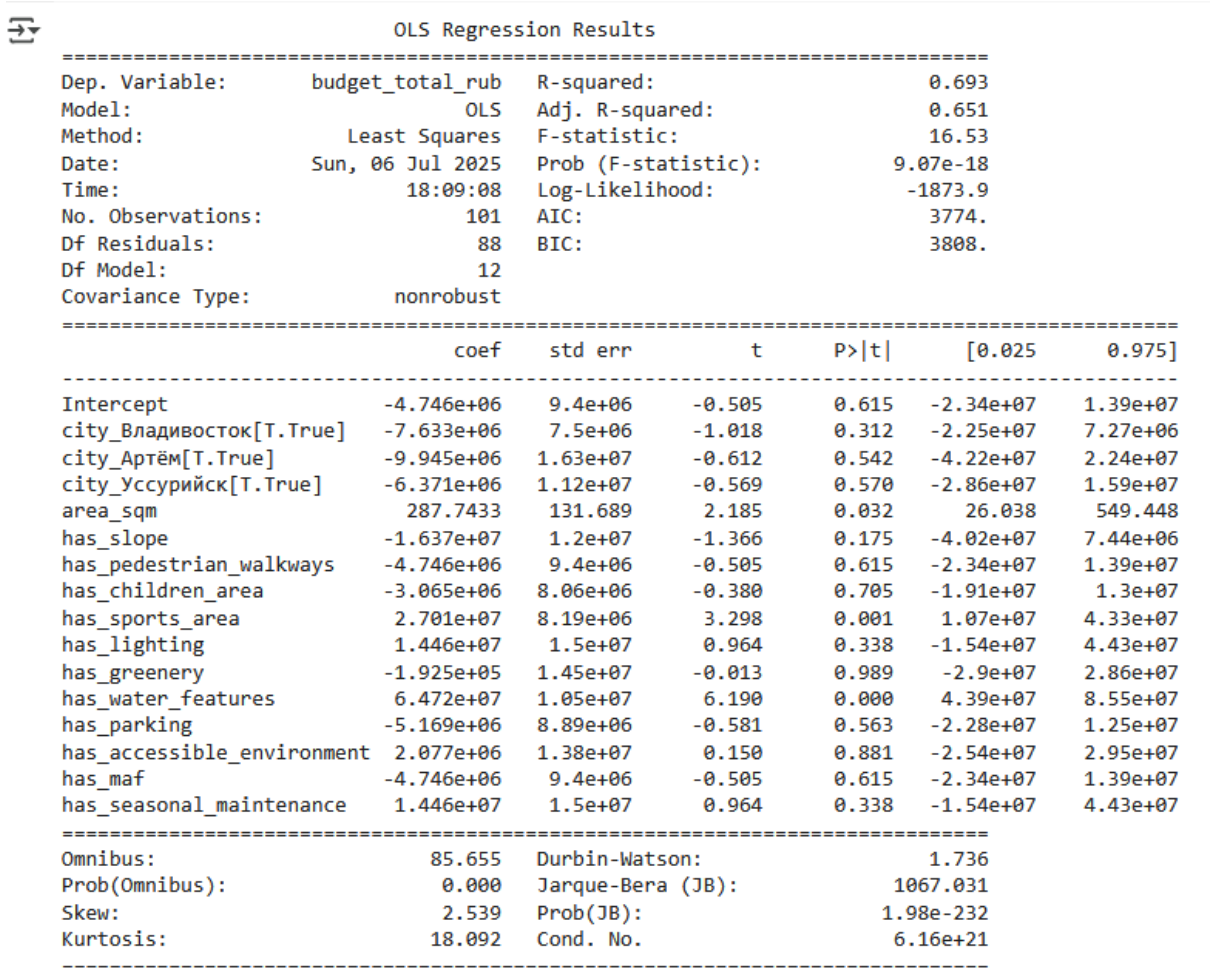
Модель линейной регрессии представляет собой один из базовых алгоритмов машинного обучения. В данной работе модель линейной регрессии использовалась как базовая интерпретируемая модель, с которой сравнивались более сложные алгоритмы.

«Случайный лес» — это ансамблевый метод, основанный на множестве деревьев решений, каждое из которых обучается на случайной подвыборке данных. Финальный прогноз получается усреднением предсказаний всех деревьев. Модель «случайный лес» характеризуется высокой точностью, устойчивостью к переобучению.

Градиентный бустинг представляет собой ещё один мощный ансамблевый метод, заключающийся в последовательном построении моделей, каждая из которых исправляет ошибки предыдущей.

Благодаря возможности настройки параметров, таких как скорость обучения, количество деревьев и глубина, градиентный бустинг может достигать высокой точности прогноза, хотя и требует внимательного подхода к предотвращению переобучения.

Результат оценки параметров линейной регрессии представлен на рисунке 4.



OLS Regression Results

Dep. Variable:	budget_total_rub	R-squared:	0.693
Model:	OLS	Adj. R-squared:	0.651
Method:	Least Squares	F-statistic:	16.53
Date:	Sun, 06 Jul 2025	Prob (F-statistic):	9.07e-18
Time:	18:09:08	Log-Likelihood:	-1873.9
No. Observations:	101	AIC:	3774.
Df Residuals:	88	BIC:	3808.
Df Model:	12		
Covariance Type:	nonrobust		

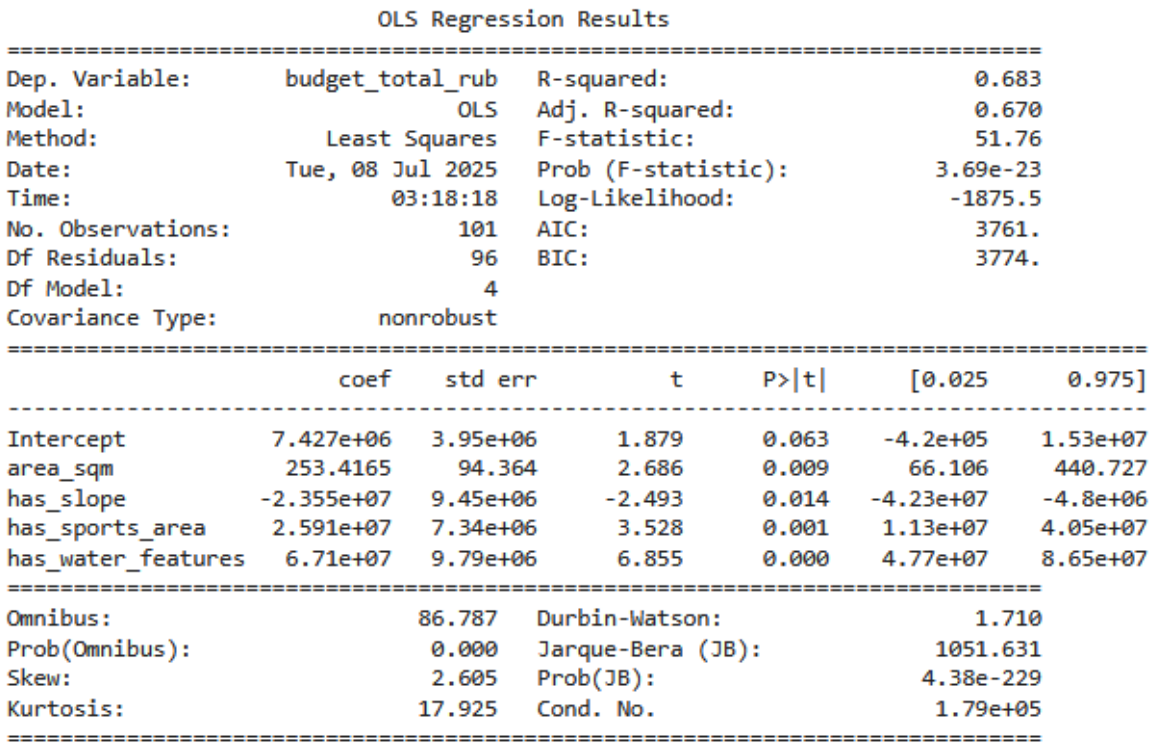
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.746e+06	9.4e+06	-0.505	0.615	-2.34e+07	1.39e+07
city_Владивосток[T.True]	-7.633e+06	7.5e+06	-1.018	0.312	-2.25e+07	7.27e+06
city_Артём[T.True]	-9.945e+06	1.63e+07	-0.612	0.542	-4.22e+07	2.24e+07
city_Уссурийск[T.True]	-6.371e+06	1.12e+07	-0.569	0.570	-2.86e+07	1.59e+07
area_sqm	287.7433	131.689	2.185	0.032	26.038	549.448
has_slope	-1.637e+07	1.2e+07	-1.366	0.175	-4.02e+07	7.44e+06
has_pedestrian_walkways	-4.746e+06	9.4e+06	-0.505	0.615	-2.34e+07	1.39e+07
has_children_area	-3.065e+06	8.06e+06	-0.380	0.705	-1.91e+07	1.3e+07
has_sports_area	2.701e+07	8.19e+06	3.298	0.001	1.07e+07	4.33e+07
has_lighting	1.446e+07	1.5e+07	0.964	0.338	-1.54e+07	4.43e+07
has_greenery	-1.925e+05	1.45e+07	-0.013	0.989	-2.9e+07	2.86e+07
has_water_features	6.472e+07	1.05e+07	6.190	0.000	4.39e+07	8.55e+07
has_parking	-5.169e+06	8.89e+06	-0.581	0.563	-2.28e+07	1.25e+07
has_accessible_environment	2.077e+06	1.38e+07	0.150	0.881	-2.54e+07	2.95e+07
has_maf	-4.746e+06	9.4e+06	-0.505	0.615	-2.34e+07	1.39e+07
has_seasonal_maintenance	1.446e+07	1.5e+07	0.964	0.338	-1.54e+07	4.43e+07

Omnibus:	85.655	Durbin-Watson:	1.736
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1067.031
Skew:	2.539	Prob(JB):	1.98e-232
Kurtosis:	18.092	Cond. No.	6.16e+21

Рисунок 4 – отчет по оценке линейной регрессии с использованием всех доступных показателей

Как видно из отчета, точность модели регрессии составляет R2 0.693 и adjusted R2 score 0.651 (с учетом «штрафа» за количество переменных) Исходя из значений p-value, можно сделать вывод, что большинство показателей являются статистически незначимыми (p-value больше 0.05). Значимыми являются показатели: Площади объекта (area_sqm), строительство или благоустройство спортивных объектов (has_sports_area) и строительство или благоустройство фонтанов и иных водных объектов (has_water_features).

Процедура пошагового улучшения линейной регрессии позволяет достичь оптимального набора показателей (переменных). Наилучшая модель включает показатели: площадь объекта (area_sqm), наличие неровностей (has_slope), благоустройство или строительство спортивных площадок (has_sports_area) и строительство или благоустройство фонтанов и других водных объектов (has_water_features). На рисунке 5 показан результат оценки параметров «короткой» регрессии.



OLS Regression Results

Dep. Variable:	budget_total_rub	R-squared:	0.683
Model:	OLS	Adj. R-squared:	0.670
Method:	Least Squares	F-statistic:	51.76
Date:	Tue, 08 Jul 2025	Prob (F-statistic):	3.69e-23
Time:	03:18:18	Log-Likelihood:	-1875.5
No. Observations:	101	AIC:	3761.
Df Residuals:	96	BIC:	3774.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.427e+06	3.95e+06	1.879	0.063	-4.2e+05	1.53e+07
area_sqm	253.4165	94.364	2.686	0.009	66.106	440.727
has_slope	-2.355e+07	9.45e+06	-2.493	0.014	-4.23e+07	-4.8e+06
has_sports_area	2.591e+07	7.34e+06	3.528	0.001	1.13e+07	4.05e+07
has_water_features	6.71e+07	9.79e+06	6.855	0.000	4.77e+07	8.65e+07

Omnibus:	86.787	Durbin-Watson:	1.710
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1051.631
Skew:	2.605	Prob(JB):	4.38e-229
Kurtosis:	17.925	Cond. No.	1.79e+05

Рисунок 5 – отчет по оценке линейной регрессии с отбором переменных

Метрика R2 score не значительно снизилась 0.683, adjusted R2 score повысился до 0.67. Все переменные, включенные в модель, статистически значимы, имеют p-value меньше 0.05.

Обучение моделей случайного леса и градиентного бустинга выполнены сначала со всеми параметрами, затем с отобранными, наиболее важными параметрами (далее feature_importance) [19]. Отбор по важности выполнен с помощью алгоритма (модуь sklearn SequentialFeatureSelector), который позволяет автоматически подобрать только значимые параметры для выбранной модели машинного обучения. На рисунках 6 и 7 изображены диаграммы «feature_importance» для случайного леса и градиентного бустинга соответственно.

Как видно на диаграммах, для обоих моделей наиболее значимы три переменных: *строительство или благоустройство спортивных объектов (has_sports_area), площадь объекта (area_sqm) и строительство или благоустройство фонтанов и иных водных объектов (has_water_features).* Вероятно, объекты благоустройства, обладающие большим размером или наличием фонтанов или спортивных площадок, как правило, имеют наибольшие расходы на благоустройство. Значимость других переменных для моделей существенно различаются.

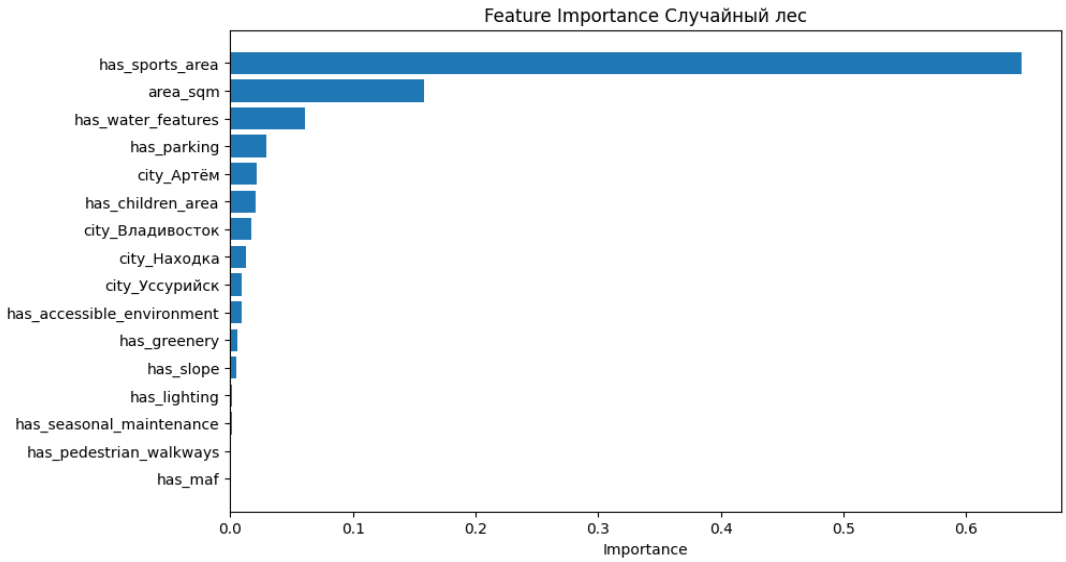


Рисунок 6 – Диаграмма Feature importance для случайного леса

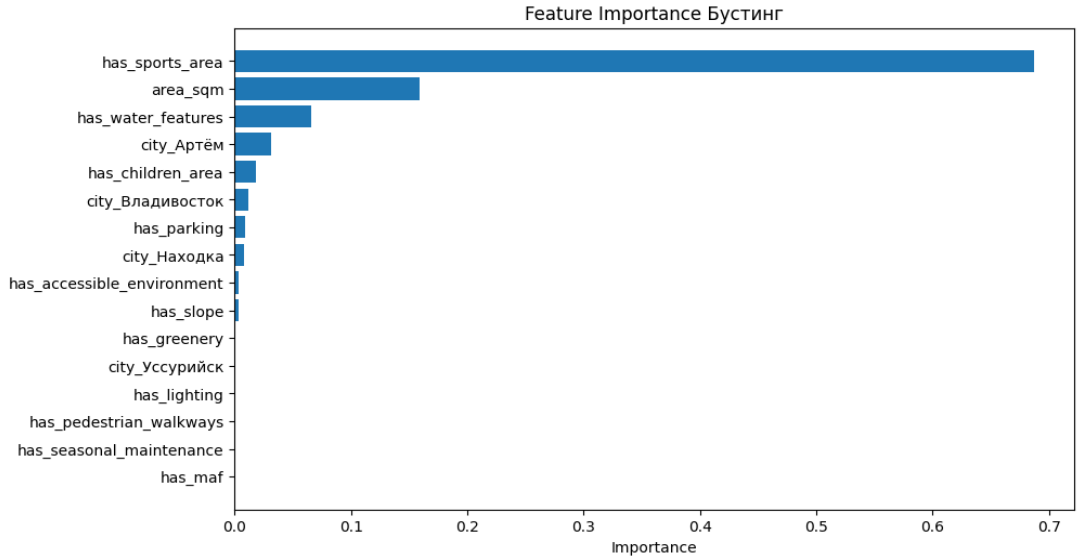


Рисунок 7 – Диаграмма Feature importance для градиентного бустинга

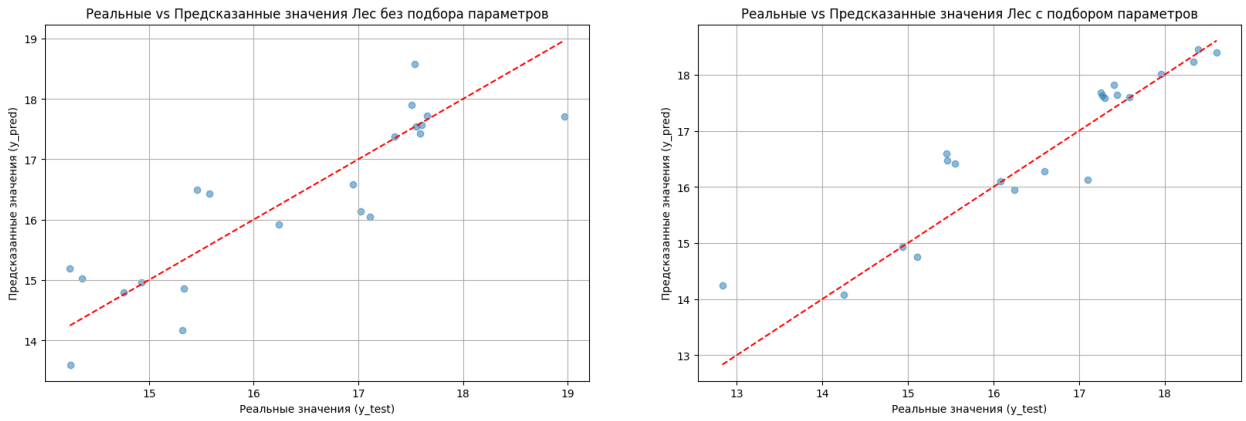


Рисунок 8 – График рассеивания для модели случайного леса без подбора параметров и с подбором параметров

На рисунке 8 представлен график рассеивания для обученных моделей случайного леса с подбором параметров и со всеми параметрами.

На представленных графиках ось X отображает значение реального целевой переменной (в нашем случае уровня расходов). Ось Y отображает значение предсказанного выходного параметра. Чем ближе точка к красной линии, тем более точна было предсказано её значение. Как видно на графиках, модель с подбором параметров имеет меньшее рассеяние, более высокую точность предсказания.

На рисунке 9 изображена такая же пара графиков рассеяния для обученных моделей, но уже моделей градиентного бустинга, также с подбором и без подбора параметров.

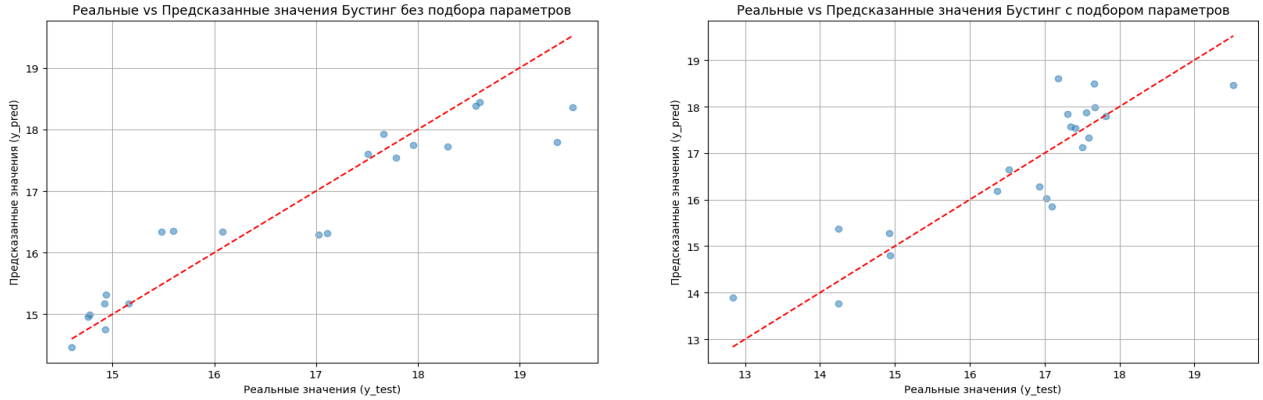


Рисунок 9 – График рассеивания для модели градиентного бустинга с подбором и без подбора параметров

Анализ графиков рассеяния позволяет сделать вывод, что модели имеют примерно одинаковую точность.

В таблице 5 показаны результаты выполнения кросс-валидации и оценки по метрике R2 score, а также доверительный интервал этих результатов (stats.t.interval()).

Таблица 5 – результаты обучения моделей машинного обучения

Название модели	Значения R2 score кросс-валидации	Среднее значение R2 score	Доверительный интервал	Список параметров
Случайный лес	[0.82 0.75 0.85, 0.78, 0.77]	0.79	[0.743, 0.840]	Все доступные
Случайный лес с подбором параметров	[0.83 0.75 0.84 0.80 0.78]	0.80	[0.756, 0.844]	Все, кроме has_lighting и has_pedestrian_ walkaways
Градиентный бустинг	[0.74 0.70 0.80 0.74 0.73]	0.74	[0.696, 0.787]	Все доступные

Градиентный бустинг с подбором параметров	[0.74 0.70 0.80 0.74 0.73]	0.75	[0.706, 0.789]	Все, кроме has_ pedestrian_walka ways

Исходя из данных, представленных в таблице, можно сделать вывод, что модель случайного леса способна более точно предсказывать расходы бюджета. Подбор же параметров, о чём можно судить, исходя из данных в таблице, не привёл к качественному изменению в точности, но всё же дал небольшое увеличение точности.

С целью оптимизации моделей (увеличения точности), было принято решение подобрать оптимальные гиперпараметры моделей (процедура выполнена помощью модулей sklearn RandomizedSearchCV и GridSearchCV). Гиперпараметров являются техническими характеристиками моделей машинного обучения и в данной работе не приводятся.

После оптимизации параметров обучения (подбора гиперпараметров), модели были вновь были обучены и проверены методом кросс-валидации. В таблице 6 показаны результаты обучения оптимизированных моделей. За основу были взяты модели с уже отобранными переменными, так как они показали более точный результат, и поэтому были использованы для составления таблица.

Таблица 6 – Результаты обучения моделей с подобранными гиперпараметрами

Название модели	Значения R2 score кросс-валидации	Среднее значение R2 score	Доверительный интервал
Случайный лес	[0.88 0.67 0.85 0.74 0.74]	0.77	[0.668, 0.881]
Градиентный бустинг	[0.85 0.73 0.84 0.82 0.75]	0.80	[0.730, 0.865]

Исходя из данных, представленных в таблице, можно сделать вывод, что метрики случайного леса стали показывать результат хуже, но метрики градиентного бустинга, наоборот, значимо выросли по сравнению со стандартными гиперпараметрами.

Таким образом, при сопоставлении метрик качества моделей, наибольшую точность демонстрируют две модели: модель случайного леса с стандартными гиперпараметрами и подобранными переменными и модель градиентного бустинга с подобранными гиперпараметрами и подобранными переменными.

Данные результаты можно считать удовлетворительным. Основной проблемой создания модели для прогнозирования является невысокая точность исходных данных. Укажем некоторые проблемы формирования датафреймов:

1. Отсутствие данных по фактическим ежемесячным расходам на объекты благоустройства;
2. Отсутствие информации о корректировке данных по расходам в связи с непредвиденными обстоятельствами.

Таким образом, очевидно, что для качественного прогнозирования необходимо реализовать систему хранения данных, содержащую насколько возможно полные сведения о финансировании проектов благоустройства.

Выводы:

Прогнозирование расходов бюджетов является сложной информационной-математической задачей. Построение моделей, пригодных для прогнозирования, требует формирования качественных датафреймов с привлечение данных из разных источников, анализа региональных закономерностей (выявления показателей (факторов) определяющий уровень расходов).

Результатами нашего исследования являются:

1) Два рабочих датафрейма, для анализа закономерностей и обучения моделей. Полученные датафреймы могут служить образцом структуры данных для реализации прогнозирования в других регионах
2) Выявленные показатели, определяющие уровень расходов бюджетов. Для всех типов моделей такими факторами являются тип объекта (спортивные объекты, фонтаны или иные водные объекты), площадь объекта.
3) Интеллектуальные модели позволяющие получить максимально возможную точность прогнозирования.

Модели, полученные в результате нашего исследования (на данных для Приморского края) показывают, что, несмотря на недостаточный объем и качество данных, применение машинного обучения позволяет получить прогноз уровня бюджетных расходов муниципалитетов на благоустройство с высокой степенью точности. Ожидаемо, такими моделями являются ансамблевые модели, построенные на основе нейронных сетей.

Библиографический список:

1. Статья о национальном проекте «Жильё и городская среда» [Электронный ресурс] // Минстрой России – Режим доступа: https://minstroyrf.gov.ru/trades/natsionalnye-proekty/natsionalnyy-proekt-zhilye-i-gorodskaya-sreda/ (дата обращения: 06.03.2025).

2. Официальный сайт проекта «Формирование комфортной городской среды [Электронный ресурс] // Формирование комфортной городской среды - Режим доступа: https://gorodsreda.ru/ (дата обращения: 06.03.2025).
3. Формирование комфортной городской среды, портал проекта «Федеральный проект «Формирование комфортной городской среды» реализуется в рамках национального проекта» [Электронный ресурс] //«Инфраструктура для жизни» – Режим доступа: https://gorodsreda.ru/(дата обращения: 01.08.2025)
4. Гареев И.Ф., Ахметгалиев Т.А., Юнусов Р.Р. Источники данных и машинное обучение на прединвестиционой стадии проектов развития территорий. // Жилищные стратегии – 2024. – том 11. – № 3. – С. 409-426
5. Стерник С.Г. Прогнозирование доходов бюджета от внутреннего производства с использованием гибридных моделей машинного обучения // Менеджмент и бизнес-администрирование – 2024. – № 3. – С. 121-136
6. Плескачев Ю.А., Пономарев Ю.Ю., Сапрыкин М.А., Территориальное планирование и прогнозирование экономических показателей методами машинного обучения // Экономическое развитие России – 2023. – Том: 30. – № 9. – С. 46-57
7. Едиый портал бюджетной системы Российской Федерации [Электронный ресурс] // Элект ронный бюджет – Режим доступа: https://budget.gov.ru (дата обращения: 10.06.2024).
8. Кадастровая карта [Электронный ресурс] // Кадастр - Режим доступа: https://кадастр.сайт/ (дата обращения: 10.06.2024).
9. Страница сайта городской администрации с отчётом об благоустройстве общественных территорий [Электронный ресурс] // Официальный сайт администрации Владивостока – Режим доступа: https://www.vlc.ru/city-environment/blagoustroystvo-obshchestvennykh-prostranstv/ (дата обращения: 8.06.2024).
10. Сайт администрации города Уссурийска [Электронный ресурс] // Официальный сайт администрации Уссурийска – Режим доступа: https://adm-ussuriisk.ru/# (дата обращения: 8.06.2024).
11. Сайт администрации города Артёма [Электронный ресурс] // Официальный сайт администрации Артёма – Режим доступа: https://artemokrug.gosuslugi.ru/ (дата обращения: 8.06.2024).
12. Сайт администрации города Находка [Электронный ресурс] // Официальный сайт администрации Находки – Режим доступа: https://www.nakhodka-city.ru/ (дата обращения: 8.06.2024).
13. Официальный сайт Единой информационной системы в сфере закупок [Электронный ресурс] // ЕИС Закупки -

Режим доступа: https://zakupki.gov.ru/epz/main/public/home.html (дата обращения: 10.02.2025).

14. Сайт администрации города Находка [Электронный ресурс] // Официальный сайт администрации Находки – Режим доступа: https://www.nakhodka-city.ru/ (дата обращения: 8.06.2024).
15. Документация для языка программирования Python [Электронный ресурс] // Python – Режим доступа: https://docs.python.org/3.13/ (дата обращения: 20.04.2025).
16. Документация для библиотеки для Python Numpy [Электронный ресурс] // Numpy – Режим доступа: https://numpy.org/doc/stable/index.html (дата обращения: 20.04.2025).
17. Документация для библиотеки для Python Pandas [Электронный ресурс] // Pandas – Режим доступа: https://pandas.pydata.org/docs/ (дата обращения: 20.04.2025).
18. Демидова, О. А. Эконометрика: учебник и практикум для вузов / О. А. Демидова, Д. И. Малахов. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 398 с. — (Высшее образование). — ISBN 978-5-534-20392-9. — Текст: электронный // образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/560504 (дата обращения: 15.09.2025).
19. H. Wang Research on the Application of Random Forest-based Feature Selection Algorithm in Data Mining Experiments/ H. Wang, International Journal of Advanced Computer Science and Applications, Vol. 14, No. 10. Intelligence Engineering, Southwest Forestry University, Kunming, Yunnan, 650224, China - c. 505.