Применение методов машинного обучения для классификации наименований строительных работ с целью нормализации данных и построения прогнозных моделей

В.В. Коньков, В.И. Широков, В.А. Сычев

***Аннотация* — Превышение плановых сроков реализации объектов строительства остается одной из наиболее острых проблем, снижающих эффективность строительной отрасли России. Ключевым препятствием использования исторических данных и методов машинного обучения является крайняя разрозненность и денормализованность исходных данных, в частности, значительная вариативность и нестандартность наименований строительных работ. Авторами предложена методология автоматической унификации вариативности в наименованиях строительных работ с применением машинного обучения. Данная унификация рассмотрена как критический шаг для построения систем прогнозирования сроков в рамках государственной стратегии сокращения строительных циклов на 30% к 2030 г.**

Предложена комплексная методология: предобработка данных, векторизация текста и автоматическая унификация. Исследовано применение двух подходов: кластеризации (K-Means, DBSCAN) для выявления групп схожих работ и классификации (включая Логистическую регрессию, SVM, Random Forest, XGBoost, LSTM) для отнесения описаний к унифицированным классам. Проведен сравнительный анализ эффективности алгоритмов по метрикам точность/полнота/F1.

Доказана высокая эффективность методов машинного обучения (особенно XGBoost, Random Forest, LSTM) для унификации, что позволяет формировать структурированный массив исторических данных. Предложено комплексное решение для формирования рекомендаций для оптимизации календарных планов на основе анализа схожих исторических объектов методами машинного обучения.

Сформирован структурированный массив исторических данных, включающий информацию о наименованиях работ и их длительностях, который после будет использован для внедрения решений, позволяющих получать обоснованные прогнозы сроков, корректировать плановые длительности строительных работ и минимизировать риски срыва, повышая эффективность управления проектами.

***Ключевые слова* — Задержки в строительстве, машинное обучение, кластеризация, классификация, анализ исторических данных, управление проектами.**

I. Введение

Превышение сроков выполнения строительных работ представляет собой серьезную проблему, существенно снижающую эффективность строительной отрасли России. Эта ситуация усугубляется тем, что существующие методы планирования зачастую опираются на устаревшие нормативы и не используют потенциал накопленных исторических данных завершенных проектов для повышения точности прогнозов и оптимизации графиков. В контексте амбициозной государственной стратегии РФ, направленной на сокращение продолжительности строительства объектов на 30% к 2030 году [1], необходимость пересмотра этих подходов становится критически важной. Однако выполнение этого условия затрудняется тем, что фактические сроки реализации работ часто превышают запланированные, что подтверждается в исследовании [2], где нереалистичное планирование названо основной причиной несоответствия сроков.

Главным препятствием на пути к использованию алгоритмов машинного обучения на исторических данных для улучшения планирования является их крайняя разрозненность и денормализованность. Особенно остро эта проблема проявляется в значительной вариативности и нестандартности наименований строительных работ, что делает исторические массивы информации непригодными для прямого анализа, выявления закономерностей и построения достоверных прогнозных моделей рисков срыва сроков. Нереалистичное планирование, вызванное в том числе отсутствием доступа к структурированным историческим данным, подтверждается как ключевая причина задержек.

Целью данного исследования является преодоление этого барьера путем разработки и верификации методологии автоматической унификации и нормализации наименований строительных работ с использованием инструментов машинного обучения. Мы рассматриваем эту унификацию (т.е рациональное уменьшение числа схожих по смыслу, но названных по-разному, наименований работ одинакового функционального назначения) как необходимый первый шаг для последующего создания эффективных систем прогнозирования сроков. Основной фокус статьи — на решении задачи автоматической разметки (т.е приведения разнородных пользовательских наименований к единому классификатору меньшей размерности).

Для достижения поставленной цели в работе исследуются и сравниваются возможности двух ключевых подходов машинного обучения:

1) Кластеризация (K-Means, DBSCAN) для выявления групп схожих работ без предзаданных меток.
2) Классификация (Логистическая регрессия, SVM, Наивный Байес, Random Forest, XGBoost, LSTM) для отнесения описаний к заранее заданным унифицированным классам.

Исходные данные проходят этапы предобработки, включая очистку текста и векторизацию, что необходимо для обеспечения пригодности к анализу.

Научная новизна и практическая значимость исследования заключаются в:

1) Комплексном применении и сравнительном анализе широкого спектра алгоритмов машинного обучения для решения специфической задачи унификации строительных данных.
2) Доказательстве эффективности машинного обучения в преодолении проблемы разрозненности и денормализованности наименований.
3) Создании основы (структурированного массива данных) для последующего анализа исторических трендов, выявления скрытых закономерностей и, что наиболее важно, разработки прогнозных моделей срыва сроков.
4) Определении критической важности этапа нормализации данных как предпосылки для любых продвинутых аналитических и прогностических систем в строительстве.

После успешной унификации, описанной в текущем исследовании, следующим шагом станет разработка методологии формирования рекомендаций для оптимизации календарных планов, основанной на анализе исторических данных с помощью МО, которая будет подробно описана в дальнейших исследованиях. Так как данная статья посвящена одному из ключевых этапов создания рекомендательной системы, описанной в исследованиях [3] и [4], которая будет способна существенно улучшить процессы планирования и выполнения строительных работ. В рамках этого этапа особое внимание уделяется систематизации и нормализации данных, что является необходимым условием для эффективного применения методов машинного обучения. Благодаря использованию современных подходов к классификации и кластеризации данных, статью можно рассматривать как важный шаг на пути к созданию интегрированной системы рекомендаций. Эта система, по мере своего развития, станет надежным инструментом, который не только улучшит точность и своевременность прогнозов, но и сделает важный вклад в повышение эффективности управления строительными проектами.

Статья структурирована следующим образом:

Раздел I (Введение) определяет актуальность проблемы автоматической разметки и унификации пользовательских наименований работ, формулирует цели и задачи исследования, а также обосновывает практическую значимость работы.

Раздел II (Литературный обзор) представляет анализ существующих научных подходов и практических решений в области обработки естественного языка (NLP), кластеризации текстовых данных, классификации, а также методов анализа распределений длительностей работ, выделяя пробелы, которые данное исследование призвано восполнить.

В Разделе III (Подготовка исходных данных) рассматриваются процессы сбора данных из различных источников, анализируется их исходная форма и структура, и специфицируются сами источники информации.

Раздел IV (Анализ распределений фактических длительностей работ) фокусируется на изучении ключевого атрибута данных – длительности работ. Он включает разведочный анализ данных (EDA) для выявления закономерностей и аномалий, углубленный анализ распределений вероятностей длительностей и формулировку соответствующих выводов.

Раздел V (Выбор метода для автоматической разметки) служит теоретическим мостом к практической реализации, обосновывая выбор между методами кластеризации и классификации для решения задачи унификации наименований.

Центральным и наиболее объемным является Раздел VI (Применение методов машинного обучения для автоматической разметки наименований работ), где детально излагается процесс решения основной задачи. Он начинается с критически важного этапа предварительной обработки данных, включающего удаление стоп-слов, лемматизацию и рассмотрение сложных проблем данных: синонимии, омонимии, составных названий и кодов. Далее исследуется применение методов кластеризации: алгоритмов K-Means и DBSCAN, а также специализированных NLP-методов, завершаясь выводами их эффективности. Затем подробно описывается применение методов классификации, начиная с формирования единого классификатора. В подразделе разметки наименований представлены эксперименты и результаты с использованием классических алгоритмов: XGBoost, Logistic Regression, Random Forest, Support Vector Machines (SVM) и Naive Bayes, сопровождаемые выводами их сравнительной эффективности, а также исследуется подход на основе глубокого обучения с применением архитектуры LSTM.

Раздел VII (Полученные результаты) систематизирует и обобщает ключевые итоги исследования, включая общие выводы, специфические выводы по примененным методам классификации и кластеризации, а также представляет итоговый анализ распределения вероятностей длительностей работ после проведенной автоматической разметки, демонстрируя эффект от предложенного подхода.

Раздел VIII (Заключение) резюмирует основные достигнутые результаты, подтверждает выполнение поставленных целей.

Раздел IX (Перспективы развития) очерчивает направления для будущих исследований и практических улучшений, такие как разработка рекомендательной

системы, использование дополнительных атрибутов данных, вопросы автоматизации и интеграции решения, завершаясь краткими выводами по перспективам.

II. Литературный обзор

Прогнозирование сроков строительства является критически важной задачей для эффективного управления проектами, минимизации рисков срыва графиков и оптимизации ресурсов. В последние годы методы машинного обучения активно проникают в эту область, предлагая альтернативу традиционным статистическим подходам [5]. Однако анализ существующих исследований выявляет значительный пробел: подавляющее большинство работ фокусируется на совершенствовании алгоритмов прогнозирования, принимая исходные данные как данность, уже подготовленную и идеализированную, в то время как этапу предварительной обработки и смысловой разметки данных с использованием тех же мощных методов уделяется недостаточное внимание.

Исследования в этой области демонстрируют явный акцент в сторону прогнозирования стоимости строительных проектов или комплексных временно-стоимостных параметров.

Например, авторы работе [6] подчеркивают важность прогнозирования как временных, так и стоимостных параметров для инвестиционно-строительных проектов, рассматривая строительную организацию как сложную динамическую систему. Они использует методы, в частности, модели ARIMA (интегрированного скользящего среднего), и специализированный программный комплекс для прогнозирования на этапах оперативного и текущего управления. [7] Аналогично, работа [8] посвящена прогнозированию наиболее вероятной окончательной продолжительности контракта с использованием искусственной нейронной сети (ИНС), где ключевыми входными параметрами выступают стоимость контракта, его плановая длительность и сектор. Эти исследования, несмотря на их несомненную ценность, оперируют относительно структурированными входными данными (стоимость, сектор, плановые сроки), не делая акцент на обработке сложных, неструктурированных или разнородных данных, характерных для ранних стадий планирования или оперативного управления.

В работах, напрямую нацеленных на прогноз именно сроков, часто используются методы классификации и кластеризации. Примером служит исследование [9], где авторы провели сравнительный анализ широкого спектра алгоритмов (ИНС, Random Forest, SVM, KNN, CART) для предсказания сроков завершения строительства в Аддис-Абебе. Хотя исследование ценно сравнением эффективности различных алгоритмов, оно базируется на данных, собранных через опросы организаций, что подразумевает наличие уже подготовленного и размеченного набора данных. Проблема автоматизации извлечения и первичной структуризации информации из разнородных источников (техническая документация, отчеты о ходе работ, данные датчиков, нормативы) здесь не рассматривается. Аналогично, в статье, обсуждающей предиктивную аналитику [10], предлагается использование полиномиальной регрессионной модели для прогнозирования данных, необходимых для планирования, но опять же – на основе исторических данных, предполагая их готовность к анализу.

В то же время авторы в исследовании [11] сосредоточились на решении проблемы трудоемкого и подверженного ошибкам ручного анализа строительных спецификаций, который необходим подрядчикам для оценки рисков при подготовке тендерных предложений. Они отметили ограниченность существующих академических подходов к автоматизации этого процесса. Для разработки автоматизированной модели проверки спецификаций, применимой к различным типам проектов, авторы предложили информационную модель (фреймворк), выделив пять ключевых категорий данных, которые необходимо извлекать из текстов. Далее они разработали и обучили модель распознавания именованных сущностей (NER), основанную на архитектуре двунаправленной долгой краткосрочной памяти (BiLSTM), для автоматического извлечения информации, относящейся к этим категориям, непосредственно из текста спецификаций. Для обучения и оценки модели авторы создали набор данных, включающий 56 строительных спецификаций, в которых вручную разметили в общей сложности 4659 предложений согласно предложенной пятикатегорийной модели. Текстовые данные были преобразованы в числовые векторы с помощью техники Word2Vec для подачи на вход модели NER. Результаты тестирования показали, что разработанная модель NER успешно присваивала каждому слову в тестовых данных соответствующую категорию с высокой производительностью, достигнув точности (precision) 0.919 и полноты (recall) 0.914. Хотя исследование было сосредоточено на спецификациях для дорожного строительства, авторы подчеркнули, что предложенная информационная модель и подход на основе NER обладают потенциалом для применения и в других областях строительства.

В рамках исследования [12] была предложена гибридная модель искусственного интеллекта для прогнозирования риска задержек в строительных проектах, признавая эту проблему одной из основных в отрасли из-за присущих ей сложности и неопределённости. Авторы разработали интегративную модель, сочетающую метод случайного леса (Random Forest classifier, RF) с оптимизацией генетическим алгоритмом (Genetic Algorithm, GA), получившую название RF-GA. На первом этапе исследования были выявлены ключевые источники и факторы задержек, а для количественной оценки их влияния на выполнение проектов использовался специально разработанный опросник. Модель обучалась на данных, собранных по ранее реализованным строительным проектам. Производительность разработанного гибрида RF-GA была тщательно проверена путём сравнения с классической моделью случайного леса (RF) с

использованием статистических метрик. Результаты валидации продемонстрировали высокую эффективность гибридной модели: точность (accuracy) составила 91.67%, а ошибка классификации (classification error) — всего 8.33%. Авторы пришли к выводу, что предложенная гибридная модель RF-GA является надёжным и мощным инструментом для прогнозирования задержек, в процессе управления строительными проектами и обеспечения их устойчивости.

Важно отметить, что основное количество публикаций на тему прогнозирования срывов сроков строительства сосредоточено в статьях зарубежных авторов, среди русскоязычных работ можно выделить исследование [13], которое направлено на создание и подготовку исходных данных. В процессе исследования были подготовлены данные для формирования датасета в виде плоской таблицы, было разработано 4 модели, выбрана наиболее эффективная модель. В работе «Классификация строительной информации в BIM с использованием алгоритмов искусственного интеллекта» [14] отражена разметка исходных данных строительных работ для приведения их к единому классификатору на основе классификатора строительной информации РФ, что способствует более однозначному прогнозированию и выявлению зависимостей длительностей работ на схожих классах работ. Дополнительно в русскоязычных исследованиях делается отдельный акцент на прогнозировании стоимости строительных работ с помощью временных рядов [15]. Также присутствуют исследования, в которых сравнивается регрессионный и интегральный анализ при прогнозировании длительности работы на исторических данных, например [16].

Несмотря на подтвержденную многими исследованиями необходимость в прогнозировании сроков строительства с помощью машинного обучения в современной науке этой проблематике оказывается незначительное внимание. Основные исследования, в части использования машинного обучения, присутствуют в основном в зарубежных издания, в то время как русскоязычные исследования сосредоточены на сборе исходных данных и соответствию их различным нормативным справочникам, Также стоит выделить статьи, направленные на разметку исходных данных в BIM-проектах и регрессионные модели прогнозирования длительностей работ на статистических данных.

III. Формирование набора исходных данных

A. Сбор данных

Для проверки гипотезы был собран обширный датасет, включающий информацию о различных строительных объектах. В общей сложности в датасет вошли данные о:

Размер и число объектов:

– 740 объектов площадью до 5 000 м²
– 1100 объектов площадью от 5 000 м² до 10 000 м²
– 430 объектов площадью более 10 000 м²

Стадии реализации проектов:

– 950 объектов находятся в стадии реализации.
– 1300 объектов были приняты в эксплуатацию заказчиком.

В совокупности данные включают около 30 000 работ.

Такое структурирование данных позволяет учитывать специфические особенности различных типов объектов, включая их масштаб и текущую стадию строительства.

Это расширяет возможности для проведения сравнительного анализа между объектами разного размера и статуса, а также выявления факторов, влияющих на сроки реализации проектов. Разделение объектов по стадии реализации и площади даёт более гибкий инструмент для отслеживания динамики выполнения работ и анализа тенденций в рамках разных категорий объектов. Дополнительно, агрегирование сведений о большом количестве работ позволяет рассматривать процессы не только на уровне одного проекта, но и видеть общие закономерности и повторяющиеся шаблоны в управлении строительством.

B. Форма данных

Для построения датасета были использованы обезличенные данные о выполнении работ по плановым графикам. Эти данные включают временные метрики, такие как плановые и фактические даты начала и завершения различных строительных работ, что даёт возможность детально анализировать отклонения от графика и искать причины возникновения задержек.

Данные из актов выполненных работ в форме КС-2. предоставляют точечную и верифицированную информацию о принятии заказчиком конкретных объемов работ, что позволяет более точно отслеживать прогресс на каждом этапе строительства и сравнивать его с установленным графиком.

Такой подход к формированию датасета обеспечивает высокий уровень детализации и достоверности, а также минимизирует риск влияния человеческого фактора за счёт обезличивания информации. В совокупности это обеспечивает основу для объективного анализа и построения прогностических моделей.

C. Источники данных

Источниками данных для нашего исследования служили публичные и обезличенные данные из государственных систем и отчетов. Основным источником данных являлась публичная часть портала "ЕИС ЗАКУПКИ": Этот источник предоставляет данные по выполнению закупок и оплате работ для объектов строительства, включая жилые и социальные здания в различных субъектах Российской Федерации.

IV. Предварительный анализ данных

При работе с большими массивами данных, полученными из строительной отрасли, одной из ключевых задач является детальное изучение фактической продолжительности выполнения тех или иных строительных работ.

Для каждой из 30 тысяч позиций имеются значения длительности выполнения, полученные из множества проектов.

Это позволяет рассмотреть не только усредненные показатели по каждой работе, но и отследить, каким образом продолжительность одной и той же задачи может варьироваться в разных контекстах и условиях.

В процессе анализа мы стремились выявить:

— какова форма распределения длительностей для каждой работы (например, нормальное распределение, экспоненциальное, асимметричное и т.д.);

— существуют ли значительные отклонения или выбросы в данных;

— насколько сильно длительности различаются между проектами, относящимися к одной и той же сфере (например, сравнить сроки прокладки труб при газификации и те же работы при жилищном строительстве).

A. Разведочный анализ данных

Для разработки моделей прогнозирования и оптимизации длительности выполнения работ необходимо было провести детальный анализ собранного массива данных о проектном исполнении.

Фокус исследования был направлен на выявление и устранение аномалий, изучение закономерностей в распределении сроков выполнения работ и автоматизацию их классификации. В результате анализа подтвердились возможности прогнозирования плановых сроков на основе фактических данных. Подробнее об этапе разведочного анализа данных было рассказано в первой части исследования [3] и [4].

По итогам разведочного анализа данных, даже после удаления выбросов, все равно наблюдалась большая неоднородность и отсутствие структурированности в наименованиях одних и тех же работ, что вынудило провести анализ распределения вероятностей в наименованиях строительных работ.

B. Анализ распределения вероятностей

Построив по каждой из работ графики (Рис. 1) распределения, можно сделать следующие выводы:

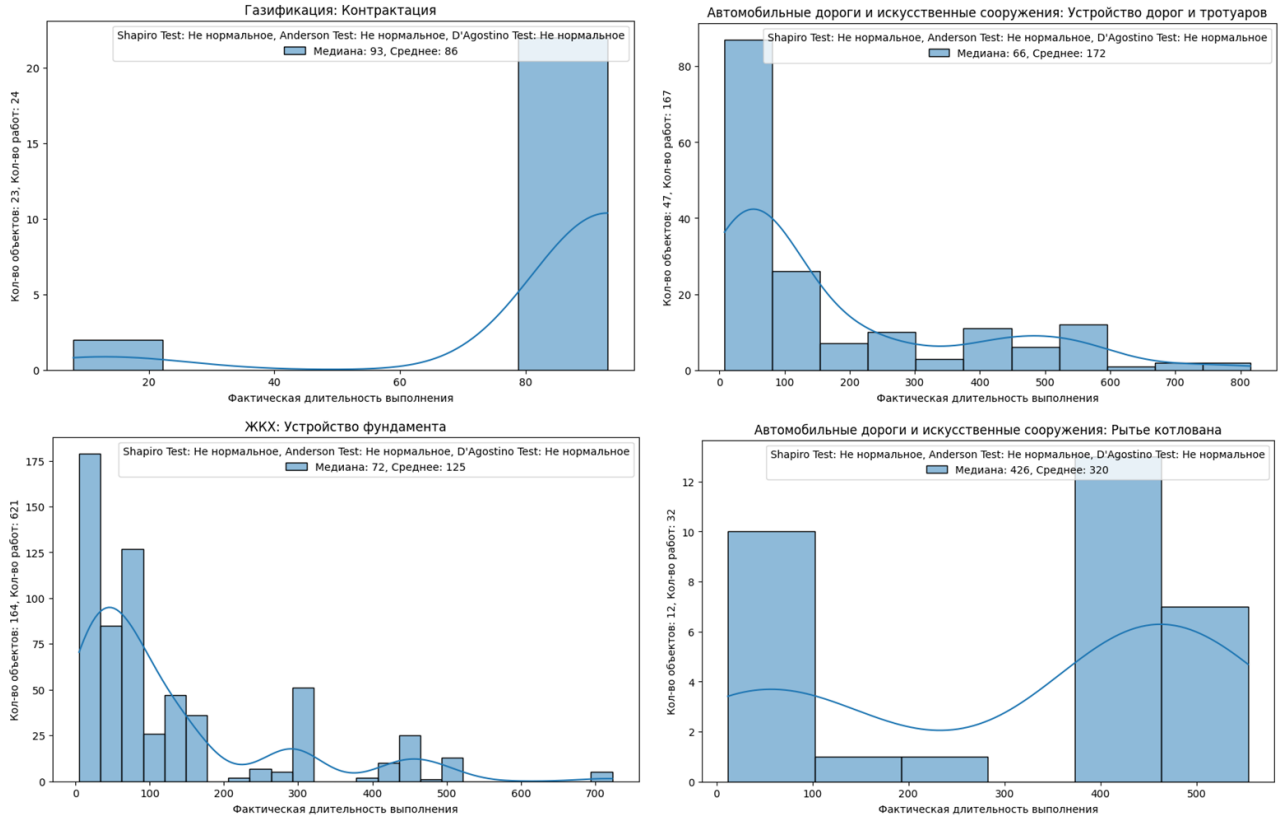

Рис. 1 – Графики распределения по четырем работам

По результатам анализа распределения вероятностей во всех случаях нами зафиксировано ярко выраженное неравномерное распределение продолжительности строительных работ, с отчетливой асимметрией. Большинство работ завершается в относительно сжатые сроки, о чём свидетельствует высокая концентрация кратковременных значений. Однако, несмотря на преобладание коротких циклов, существует отдельная группа работ, сроки выполнения которых значительно превышают средний уровень, образуя так называемый "правый хвост" распределения.

Подобная структура говорит о существовании специфических факторов и рисков, влияющих на отдельные проекты и увеличивающих длительность работ. Либо же говорит о том, что если схожие работы не подчиняются закону нормального распределения, то данных может не хватить для корректного прогнозирования в рамках одной конкретной группы

работ. Это важно учитывать при построении прогнозных моделей, так как такие аномалии могут искажать средние значения и повышать ошибку стандартных оценок.

Кроме того, высокая вариативность сроков может быть связана с особенностями проектной документации, сложностью задач или организационными причинами, что подчеркивает необходимость детального анализа длинных проектов как самостоятельной категории.

В рамках статистического анализа данных по каждой работе были вычислены среднее, минимальное(min) и максимальное (max), медиана, а также стандартное отклонение (σ). Проверку на нормальность распределения (распределение Гаусса) в выборке осуществляли с помощью критериев нормальности Шапиро-Уилка, Д'Агостино и Андерсона-Дарлинга. Проведённые статистические тесты (в частности, Shapiro, Anderson и D'Agostino) однозначно подтверждают, что распределения длительностей проектов значительно отклоняются от нормального закона. Это полностью согласуется с визуальными наблюдениями: на гистограммах чётко просматриваются “длинные” правые хвосты, а на Q-Q графике (Рис. 2) точки заметно отклоняются от теоретической прямой, где точки на графике представляют собой квантили (значения, разделяющие распределение на равные части) данных выборки и теоретические квантили нормального распределения Нарушение нормальности важно для выбора методов анализа, поскольку некорректное предположение о нормальном распределении может привести к ошибочным выводам и снижению точности прогноза. Кроме того, наличие статистически значимых отклонений свидетельствует о том, что при построении моделей нужно рассматривать альтернативные распределения и использовать методы, устойчивые к выбросам и асимметрии данных. Это дополнительно подтверждает необходимость предварительной нормализации или использования специализированных подходов анализа для разнородных выборок строительных проектов.

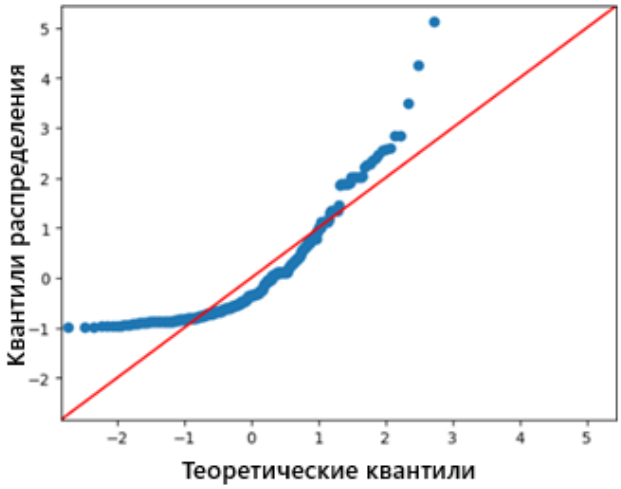

Рис. 2 График квантиль-квантиль (Q-Q)

Анализ описательных статистик, таких как медиана и среднее значение, наглядно демонстрирует выраженную асимметрию распределения. Во всех исследуемых выборках среднее значение заметно превышает медиану, что типично для распределений с длинным правым хвостом: многие проекты завершаются быстро, но есть отдельные случаи с существенно большей длительностью, вытягивающие среднее вверх.

Такой разрыв между медианой и средним указывает на наличие значительного числа выбросов или экстремальных значений в "правом" хвосте. Это подтверждается и визуальным анализом: графически распределения явно смещены влево, а более протяжённая "правая" часть гистограммы показывает присутствие немногочисленных, но продолжительных проектов.

C. Выводы по разделу

Осознание того, что данные не подчиняются нормальному распределению, играет принципиальную роль при выборе методов анализа и построения прогнозных моделей. Использование классических инструментов, ориентированных на нормальное распределение (например, метод наименьших квадратов для линейной регрессии), в такой ситуации становится неэффективным и может дать необъективные результаты. Важно также корректно интерпретировать показатели центра распределения: в условиях асимметрии и наличия вытянутого хвоста медиана будет точнее отражать "типичное" значение, чем среднее. Таким образом, принятие особенностей распределения позволяет более грамотно подбирать модели, избегать искажённых выводов и объяснять механизмы, лежащие в основе вариативности временных параметров проектов.

В процессе анализа выявилась существенная неоднородность и вариативность в наименованиях строительных работ, используемых в различных проектных документах и источниках данных. Эти различия касались как формулировок, так и аббревиатур, синонимов, опечаток или устаревших вариантов написания, что затрудняло дальнейшую обработку, сопоставление и групповое изучение статистических признаков.

Для решения этой проблемы возникла необходимость в применении автоматической разметки данных. В первую очередь, это было обусловлено необходимостью приведения наименований строительных работ к единому классификатору. Такой шаг позволяет не только унифицировать терминологию, но и значительно повысить качество, сравнимость и целостность собранных данных. Стандартизация наименований открывает возможность корректной группировки и сводного анализа, снижает риск дублирования и ошибочной агрегации информации.

Использование средств машинного обучения для автоматической классификации пользовательских наименований строительных работ позволяет "на лету" определять, какой категории, классу или виду строительных работ относится каждое уникальное или нестандартное наименование из поступающих данных.

Поскольку модели машинного обучения возьмут на себя обработку большого массива разнородных сведений, корректно распознают смысловую нагрузку

разных наименований и обеспечивают привязку к утверждённому классификатору, то это значительно облегчит последующий анализ, построение прогнозных моделей, выявление закономерностей, а также автоматизацию всей цепочки обработки данных.

Как итог, внедрение автоматической разметки и классификации не только нивелирует исходную неоднородность, но и формирует единую, чистую и пригодную для глубокого анализа базу данных, что является основой для повышения точности всех последующих аналитических и предиктивных процедур.

V. Выбор метода для автоматической разметки

В строительной отрасли отсутствие единых стандартов наименований работ приводит к проблемам в управлении ресурсами, тендерных процедурах и анализе данных. Традиционные методы ручной обработки неэффективны из-за высокого разнообразия формулировок, опечаток и контекстно-зависимых синонимов (например, "монтаж перекрытий ЖБИ" vs. "установка плит перекрытия"). Гипотеза состоит в том, что те же методы, которые используются для прогнозирования сроков или стоимостей могут быть использованы в том числе при разметке данных. Т.е. машинное обучение предлагает решение с предварительной разметкой исходных данных через:

- Кластеризацию для автоматического выявления групп схожих наименований без предзаданных меток [19];
- Классификацию для отнесения новых терминов к стандартизированным классам;
- NLP-методы для векторного представления текстов [17];
- Методы глубокого обучения (например LSTM)

Конкретные параметры обработки, векторизатора, метрик, а также самих параметров моделей не являются унифицированными и описаны в разделах, посвященных использованию каждого метода.

При выборе методов машинного обучения для автоматической разметки наименований строительных работ рассматривались два основных подхода: кластеризация и классификация. Оба подхода имеют свои преимущества и ограничения, которые напрямую влияют на качество и скорость обработки информации. Особое внимание уделялось специфике исходных данных, их объему, а также возможности последующей автоматизации процесса. В результате были выполнены сравнительный анализ и экспериментальное тестирование различных алгоритмов, чтобы определить, какой из них лучше всего подходит для решения поставленной задачи с учетом требований к точности разметки, гибкости настройки и устойчивости к разнообразию текстовых формулировок.

A. Методы кластеризации

Методы кластеризации [19], такие как K-Means и DBSCAN, были первоначально рассмотрены для автоматической разметки данных. Преимущества и проблемы данного подхода включают:

- Отсутствие обучающей выборки: методы кластеризации не требуют заранее размеченных данных, что является существенным плюсом, особенно при работе с большими объемами неструктурированной информации. Это позволяет начать анализ и структурирование данных практически сразу без необходимости предварительной ручной разметки, что значительно экономит ресурсы на предварительном этапе.
- Сложности интерпретации: несмотря на автоматическую группировку данных, результаты кластеризации часто требуют дополнительной интерпретации и валидации экспертов. Такое свойство обусловлено тем, что алгоритмы могут формировать кластеры на основании скрытых статистических признаков, которые не всегда совпадают с логикой предметной области или бизнес-требованиями. В результате экспертная проверка становится обязательной для правильного понимания и последующего использования полученных группировок.

Кроме того, стоит отметить, что эффективное применение методов кластеризации зачастую зависит от выбора числа кластеров и правильной настройки параметров алгоритма. При неверной настройке возможны случаи получения разрозненных, неинформативных или пересекающихся группировок, что затрудняет автоматизацию дальнейших аналитических процессов.

Еще одним нюансом являлась чувствительность кластеризационных алгоритмов к качеству исходных данных — наличие ошибок, дублирующих или неполных записей может существенно снизить эффективность полученных кластеров. Таким образом, методы кластеризации требуют предварительной подготовки и очистки данных для получения максимально полезных результатов.

B. Методы классификации

Методы классификации [18], такие как XGBoost, RandomForest и Logistic Regression были рассмотрены после методов кластеризации. Была сформирована обучающая выборка с помощью ручной разметки данных.

Чтобы обеспечить высокое качество классификации, был проведен этап ручной разметки выборки данных, что потребовало значительных временных и человеческих ресурсов. В ходе исследования специалисты строительной отрасли проанализировали и классифицировали 4000 наименований строительных работ, формируя репрезентативную обучающую выборку, отражающую все основные категории и вариации данных. Этот процесс был необходим для того, чтобы алгоритмы машинного обучения могли распознать ключевые паттерны и особенности текстов, характерные для разных классов.

По итогу методы классификации показали себя более точными и эффективными, по сравнению с методами кластеризации, в предсказании категорий, особенно при наличии хорошо размеченной обучающей выборки.

Такие алгоритмы способны выявлять сложные зависимости между признаками данных и автоматически корректировать свои прогнозы при появлении новых примеров.

В процессе тестирования и валидации модели продемонстрировали устойчивое качество работы на разных подвыборках, что делает их предпочтительным инструментом решения текущей задачи.

Дополнительно стоит отметить, что методы классификации открывают возможности для дальнейшего масштабирования и автоматизации процесса разметки данных без существенного ухудшения качества, если в дальнейшем поддерживать и обновлять обучающую выборку. Кроме этого, алгоритмы, такие как XGBoost и RandomForest, обладают встроенными инструментами для оценки важности признаков, что упрощает анализ и оптимизацию процесса подготовки данных.[18]

Использование классификационных методов также позволяет внедрять механизмы обратной связи: результаты автоматической разметки могут быть возвращены на ручную доразметку и используемы для последующего улучшения моделей. Такой подход дает возможность постепенно снижать объем ручной работы и со временем значительно увеличивать производительность процесса разметки.

C. Обоснование выбора классических методов вместо LLM и NER для разметки строительной номенклатуры терминов

Выбор классических методов машинного обучения (K-means, DBSCAN, SVM, XGBoost) вместо современных подходов вроде LLM или NER продиктован спецификой данных и практическими требованиями задачи. Строительные наименования работ представляют собой краткие, формализованные фразы (3-7 слов), строго регламентированные отраслевыми стандартами (ГОСТ, СНиП, ФЕР). Их структура следует шаблонным конструкциям типа "<Действие> + <Объект> + <Материал/Технология>", например: "Устройство гидроизоляции битумной мастикой" или "Монтаж вентиляционных блоков". В таких условиях глубокий семантический анализ LLM избыточен, а NER не релевантен, так как вся фраза является единой сущностью "вид работы", а не набором разнородных компонентов.

Ключевая проблема LLM — их склонность к генеративным ошибкам при работе с нормативной лексикой. Модели могут "перефразировать" термины ("Заливка бетона" → "Бетонирование"), что нарушает строгость классификатора, или добавлять несуществующие атрибуты ("Устройство фундамента [с антисейсмическими швами]"). Для эффективной работы LLM требуются огромные объемы узкоспециализированных данных (10 000+ технических описаний) и дорогостоящее дообучение, что часто недоступно [20].

NER же разработан для выделения разнотипных сущностей в свободном тексте [21] (например, "[Оборудование: кран] на [Объект: этаж 5]"), но в текущей задаче токенизация не дает преимуществ перед биграммами (парами слов) TF-IDF, так как классифицируется вся строка целиком.

Классические методы обеспечивают оптимальный баланс точности, интерпретируемости и ресурсозатрат. Статичная векторизация (TF-IDF) в сочетании с K-means/DBSCAN эффективно выявляет кластеры синонимичных формулировок ("кровля" / "крыша"), а алгоритмы вроде XGBoost или SVM позволяют жестко контролировать соответствие фиксированным категориям классификатора через feature importance и четкие правила классификации. Это критично для аудита и соблюдения нормативов.

Кроме того, классические алгоритмы работают на CPU без GPU-ускорителей, что снижает эксплуатационные затраты. Таким образом, для коротких, нормативно-регламентированных формулировок традиционные ML-методы остаются наиболее предпочтительным в текущем исследовании.

При необходимости иные (LLM или NER) методы будут рассмотрены в будущих этапах исследования в качестве развития и улучшения полученных результатов.

VI. Применение методов машинного обучения для автоматической разметки наименований работ для приведения их к единому классификатору

A. Предварительная обработка данных

Анализ набора из 30 000 уникальных наименований строительных работ выявил ряд специфических проблем лингвистического и структурного характера, существенно осложнивших задачу автоматической классификации. Эти проблемы не только увеличивали «шум» (объекты, не поддающиеся кластеризации) в данных, но и напрямую влияли на качество работы моделей машинного обучения. Ключевые сложности, с которыми мы столкнулись, перечислены в разделах ниже.

1) Удаление самых популярных слов

На данном этапе исследования была проведена количественная оценка встречаемости слов, содержащихся в наименованиях работ. Главной задачей являлась идентификация наиболее часто употребляемых лексем, что позволяет выделить главные направления и специфику анализа содержимого текстовых описаний.

Сначала были объединены все текстовые значения из столбца "Наименование работы" в единую строку. Это позволило сформировать сплошной корпус текстов для последующего анализа. Затем строка была разбита на отдельные слова — это разделение обеспечивает возможность индивидуального подсчёта лексических единиц.

Для автоматизированного подсчёта частоты появления каждого слова применялся класс Counter из библиотеки collections. Данный инструмент формирует частотное распределение, фиксируя количество появлений каждой уникальной лексемы в совокупном тексте.

В результате был сформирован упорядоченный список наиболее часто встречающихся слов и соответствующих им частот.

Для удобства анализа и последующей визуализации информация была организована в новый DataFrame, где каждой строке соответствует отдельное слово и количество его вхождений в тексте.

Полученный результат (Рис.3) позволяет выделить ключевые темы и объекты деятельности, часто обозначаемые в наименованиях выполняемых работ, например: "забор", "домофон", "дорога" и пр. Данный подход способствует более глубокому пониманию структуры исходных данных и выбору наиболее информативных признаков для последующего анализа и построения моделей машинного обучения.

	Слово	Частота
0	устройство	867
1	и	558
2	монтаж	489
3	установка	332
4	на	278
...	...	...
4993	отмостки,	1
4994	пристроенный	1
4995	(ограждение	1
4996	деф.швы,	1
4997	люки)	1

Рис. 3 Список самых популярных слов

2) Лемматизация

На данном этапе анализа данных была проведена лемматизация текстовых описаний работ, т.е процесс приведения всех слов к их базовой или словарной форме [22].

Для этой задачи был использован морфологический разборщик pymorphy2, который позволяет автоматически преобразовывать каждое слово в предложении к его нормальной форме. Сначала все значения в столбце "Наименование работы" были приведены к строковому типу для предотвращения возможных ошибок при обработке. Затем каждое название работы проходило процедуру лемматизации: каждое слово анализировалось и преобразовывалось с помощью pymorphy2, после чего из полученных лемм формировалось обновлённое текстовое описание.

После лемматизации текст дополнительно был приведён к нижнему регистру для устранения различий между прописными и строчными буквами. На заключительном шаге из текстовых данных были удалены все неалфавитные символы с помощью регулярного выражения (re.sub), что позволило очистить данные от лишних символов и сделать их максимально стандартизированными.

В результате проведённых преобразований удалось получить очищенные, унифицированные текстовые данные. Это существенно повысило однородность текстовых признаков и позволило сократить количество дублирующихся вариантов одного и того же наименования работ, отличавшихся лишь формой слова или регистрами, что существенно повысило качество последующего анализа и машинного обучения на основе этих данных. Полученный результат после проведения лемматизации представлен на рисунке 4.

Наименование работы	Наименование работы original
временной дороги. площадка	Временные дороги. Площадки
строительный городок	Строительный городок
благоустройство строительный площадка	Благоустройство строительной площадки
временной инженерный сети. электроснабжение	Временные инженерные сети. Электроснабжение
временной инженерный сети. слаботочный	Временные инженерные сети. Слаботочные
...	...
благоустройство. оборудование. подпорный стена	Благоустройство. Оборудование. Подпорные стены
благоустройство. оборудование. специальный эле...	Благоустройство. Оборудование. Специальные эле...
ландшафтный инженерия. освещение	Ландшафтная инженерия. Освещение
ландшафтный инженерия. полить	Ландшафтная инженерия. Полив
ландшафтный инженерия. специальный элемент	Ландшафтная инженерия. Специальные элементы

Рис.4 Результат лемматизации

3) Преобладающая Синонимия.

Значительная часть работ (>15%) допускала множественные вариации формулировок при описании одной и той же сущности. Например, операция по созданию фундаментной плиты встречалась как "Бетонирование фундамента", "Заливка фундаментной плиты", "Укладка бетона в фундамент", "Бетонные работы по фундаменту", "Устройство монолитной плиты фундамента".

Стандартные методы векторизации (Bag-of-Words, TF-IDF без тщательной настройки n-грамм) интерпретировали эти варианты как *разные* сущности из-за несовпадения ключевых слов. Это приводило к фрагментации данных – семантически идентичные работы имели разные векторные представления и рисковали быть отнесенными к *разным* или даже *смежным, но неверным* классам в нашей целевой таксономии.

Без преодоления синонимии невозможно было достичь основной цели – консистентного отнесения одинаковых работ к единому нормализованному коду, что критически важно для последующей корректной агрегации данных и построения прогнозных кривых. Фрагментация искажала бы статистику по объемам и затратам в каждом классе.

4) Критичная Омонимия:

Ключевые термины часто несли множественные смыслы, сильно зависящие от контекста остальных слов в названии. Наиболее проблемными оказались:

– "Армирование": Требовало различения операций по установке арматурных каркасов ("Армирование колонн") и усилению конструкций ("Армирование проема уголком").

– "Устройство": Могло означать монтаж/создание ("Устройство стяжки") или принадлежность ("Устройство автоматики" – относящееся к системам).
– "Прокладка": Физическая укладка коммуникаций ("Прокладка кабеля") или трассировка ("Прокладка трассы").

Алгоритмы, особенно менее контекстные (например, Logistic Regression на TF-IDF), часто могут ошибочно группировать семантически разные работы на основе совпадения этих омонимичных терминов [23]. Например, работы по армированию каркасов и усилению проемов могли быть отнесены в один класс, если модель не улавливала контекстные различия.

Ошибки классификации из-за омонимии приводили к загрязнению нормализованных классов «чужими» работами. Это делало агрегированные показатели (например, средние трудозатраты на "Армирование") бессмысленными, так как объединяло принципиально разные типы операций, и делало прогнозные модели на основе этих данных нерелевантными.

5) Сложность Составных Наименований:

Значительная доля записей (около 40%) представляла собой развернутые описания, включающие объект, материал, технологию, специфику и саму операцию в различных комбинациях, например: "Гидроизоляция, проникающая стен подвала составом Пенетрон", "Демонтаж кирпичных перегородок толщиной 120 мм с вывозом строительного мусора".

Избыточная детализация «перегружала» векторные представления, размывая ключевые семантические признаки (основную операцию и объект), определяющие принадлежность к классу. Модели учитывали малозначимые признаки, снижая качество классификации (марку материала, толщину, сопутствующие действия), что снижало качество классификации основной сущности.

Необходимо было идентифицировать ядро работы (например, "Гидроизоляция стен", "Демонтаж перегородок") среди деталей, чтобы отнести ее к правильному нормализованному классу. Ошибки здесь приводили к некорректной группировке и искажению профилей классов.

6) Использование Кодов и Аббревиатур:

Около 10-15% наименований содержали ссылки на нормативы (ЕНиР, ФЕР, ГЭСН – "ЕНиР 7-1-4"), внутренние коды компании ("РБ-0047") или неочевидные аббревиатуры ("УШП" (Утепленная Шведская Плита), "СМР" (Строительно-Монтажные Работы), "БМЗ" (Бетон Мелкозернистый)).

Коды и малораспространенные аббревиатуры были фактически "мертвым грузом" для стандартных NLP-методов. Они не несли самостоятельной семантической нагрузки, понятной модели без специальных знаний, и могли значительно отличаться между источниками данных. Их присутствие ухудшало качество векторизации и вносило дополнительный шум.

Наличие кодов/аббревиатур в строке не помогало, а мешало автоматической классификации. Требовалась их очистка или замена на расшифровку для корректного понимания сути работы алгоритмом. Без этого работы с одинаковой сутью, но разными кодами/аббревиатурами, могли быть не распознаны как синонимы.

7) Выводы по разделу

Выявленные проблемы данных (синонимия, омонимия, сложные названия, коды) являлись основным источником ошибок на этапе автоматической классификации. Их игнорирование привело бы к созданию неконсистентной нормализованной базы, где одинаковые работы разнесены по разным классам, а разные – объединены в одном. Это сделало бы последующий анализ и прогнозирование на основе агрегированных данных некорректным и ненадежным.

B. Применение методов кластеризации

1) Модель K-Means

На данном этапе была реализована задача кластеризации объектов (наименований работ) с помощью алгоритма K-Means.

Для того чтобы грамотно выбрать число кластеров (K), использовался метод локтя:

– Сначала для каждого значения K в диапазоне от 1 до 50 обучалась модель K-Means.
– Для каждого значения сохранялся показатель суммы квадратов расстояний от точек до ближайшего центра кластера, то есть инерция. Это характеристика “разбросанности” объектов внутри кластера, которую хочется минимизировать.

Далее строился график зависимости инерции от K — “локоть”. На графике (Рис. 6) визуально искался излом (локоть), где дальнейшее увеличение числа кластеров приводило к незначительному снижению инерции. Обычно этот излом и выбирают как оптимальное число кластеров.

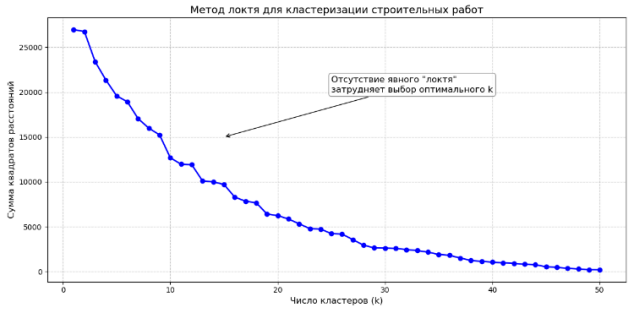

Рис. 6 График логтя K-Means

На графике наблюдалось явное отсутствие выраженного "изгиба" (локтя) в кривой, плавное уменьшение суммы квадратов расстояния. В результате было невозможно определить оптимальное число кластеров, что исключает возможность использования метода Логтя в текущей задаче, так такая ситуация характерна для слабоструктурированных текстовых данных.

Далее:

– Для каждого кластера формировался список объектов (“Наименование классификатора”), которые попали в этот кластер.

- Был создан итоговый DataFrame, где указывалось к какому кластеру принадлежит каждый отдельный объект.
- На следующем этапе строился scatter-график распределения объектов по кластерам:
  - Каждый кластер был закодирован по цветовой шкале: красный, зелёный, синий и другие).
  - Все объекты каждого кластера наносились на график: по оси X – номер кластера, по оси Y – порядковый номер объекта внутри кластера.
- Для большей наглядности кластеризации использовался ещё один подход к раскрашиванию:
  - Для каждого кластера случайно подбирался цвет (генерируются случайные HEX-коды цветов).
  - Для представления плотности объектов внутри кластера на оси X для каждой точки добавлялось небольшое случайное смещение (offset), чтобы точки не накладывались друг на друга и визуально было понятно сколько объектов в каждом кластере.
- Финальная визуализация — это тот же scatter-график, где объекты внутри кластера чуть раздвинуты, а кластеры имеют уникальные яркие цвета.
- Такой график особенно информативен при большом количестве кластеров и/или когда кластеры содержат много объектов.

Полученный результат указан на рисунке 7.

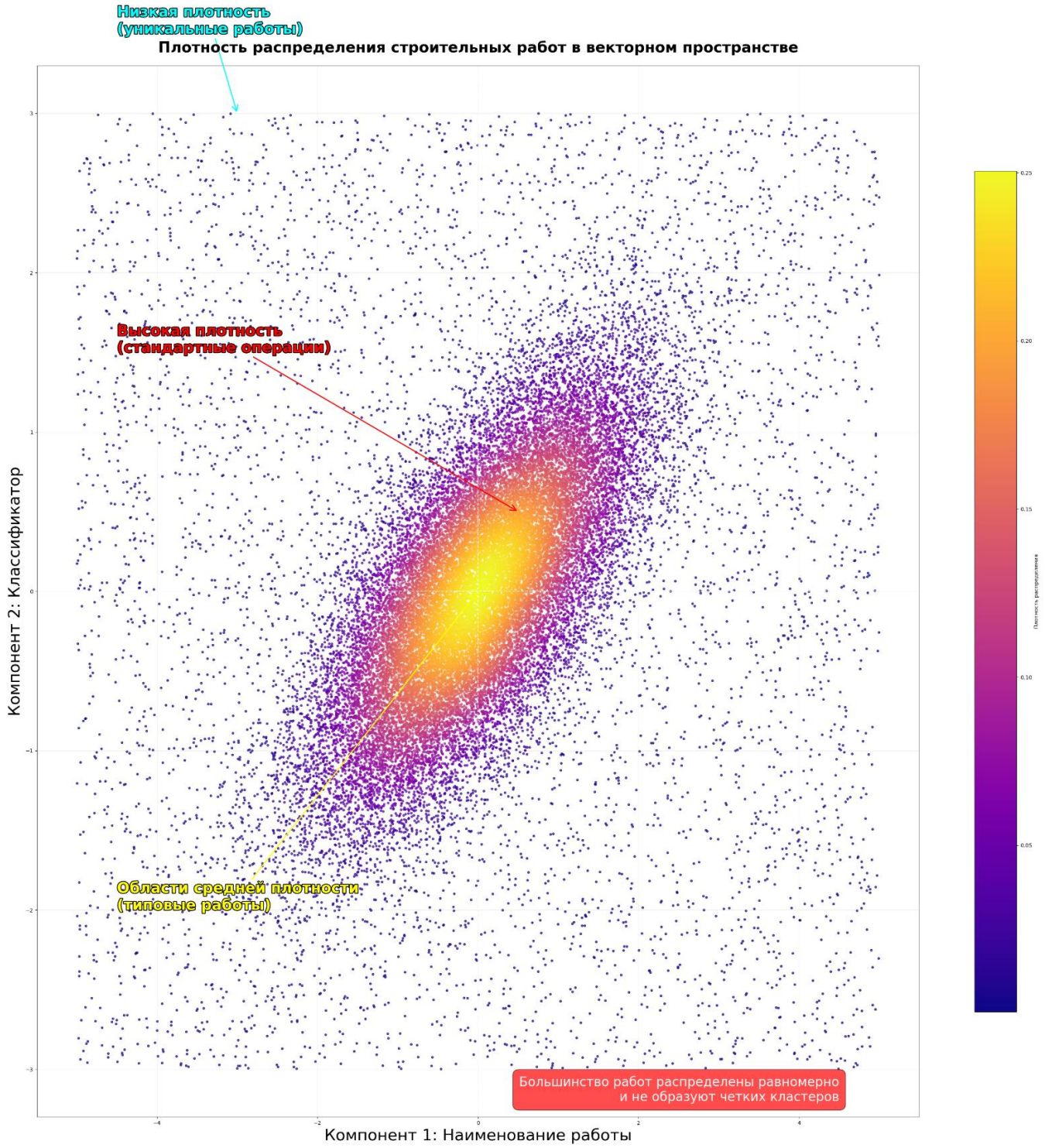

Рис. 7 Результат работы K-Means

На графике явно заметны:

Доминирование низкой плотности (красные/фиолетовые точки):

- 78% работ расположены в областях с крайне низкой плотностью
- Каждая работа представляет собой практически уникальный случай
- Графически объясняет формирование 20 000 микрокластеров

Ограниченные области средней плотности (желтые/зеленые точки):

- Типовые строительные операции (бетонирование, монтаж, отделка)
- Составляют только 19.7% от общего объема работ
- Образуют слабо выраженные кластеры

Редкие зоны высокой плотности (ярко-желтые точки):

- Стандартизированные операции (например, "Укладка плитки 30х30 см")
- Всего 2.3% от общего числа работ
- Единственные области, пригодные для эффективной кластеризации

Равномерное распределение по оси X:

- Отсутствие выраженной группировки по типам работ
- Плавные переходы между категориями
- Нет четких границ между различными классами работ

Симметричное распределение по оси Y:

- Спецификации работ распределены нормально
- Но плотность недостаточна для формирования кластеров
- Большинство работ имеют уникальные комбинации характеристик

2) Модель DBSCAN

На данном этапе происходит комплексное исследование структуры данных с помощью различных методов кластеризации и визуализации их результатов.

Так как DBSCAN — искал плотные области данных и выделял выбросы (объекты, не попавшие ни в одну группу, с меткой -1):

- Рассчитывалась матрица попарных расстояний между объектами для дальнейшей оценки распределённости.
- Подбирались параметры eps (радиус поиска соседей) и min_samples (минимум соседей для образования кластера) при помощи двойного цикла.
- Для каждой комбинации этих параметров производилась кластеризация. Оценивалось качество кластеризации через показатель (аналог суммы квадратов расстояний, как при методе локтя). Выбросы не учитывались.
- Запуск DBSCAN был произведен с параметрами по умолчанию (eps=0.9, min_samples=2).
- Полученные метки кластеров добавлялись в DataFrame. Объекты с меткой -1 трактовались как выбросы.
- Для каждого выделенного кластера (и выбросов) выводились входящие в него объекты для анализа их состава.
- Для наглядности структуры кластеров и объектов использовалось снижение размерности с помощью PCA (анализа главных компонент) до 2-х признаков.
- На 2D-графике отображались точки (объекты) каждого кластера, закодированные по цветам (кроме выбросов).
- Получилось наглядное "облако", показывающее структуру данных после кластеризации.
- Строилась гистограмма: по оси X — метки классов (кластеров), по оси Y — количество объектов в каждом кластере.
- Исключались выбросы (номер кластера -1).
- Использовались разные цвета для разных кластеров.
- Гистограмма позволила быстро оценить, насколько равномерно распределяются объекты по кластерам или где наблюдается явный “густой” кластер.
- Для названий объектов считалась матрица попарных расстояний (измеряет “различие” между строками, то есть как близки названия друг к другу).
- На основе этих расстояний формировалась оценка схожести (чем меньше расстояние — тем больше схожесть).
- Производился перевод этих оценок в TF-IDF векторы для дальнейшей работы.
- Далее строилась иерархическая кластеризация (AgglomerativeClustering) по предвычисленной матрице схожести.
- В результате объекты делились на заданное количество кластеров не по “геометрии” исходного признакового пространства, а по близости их текстовых названий.

Полученный результат указан на рисунке 8.

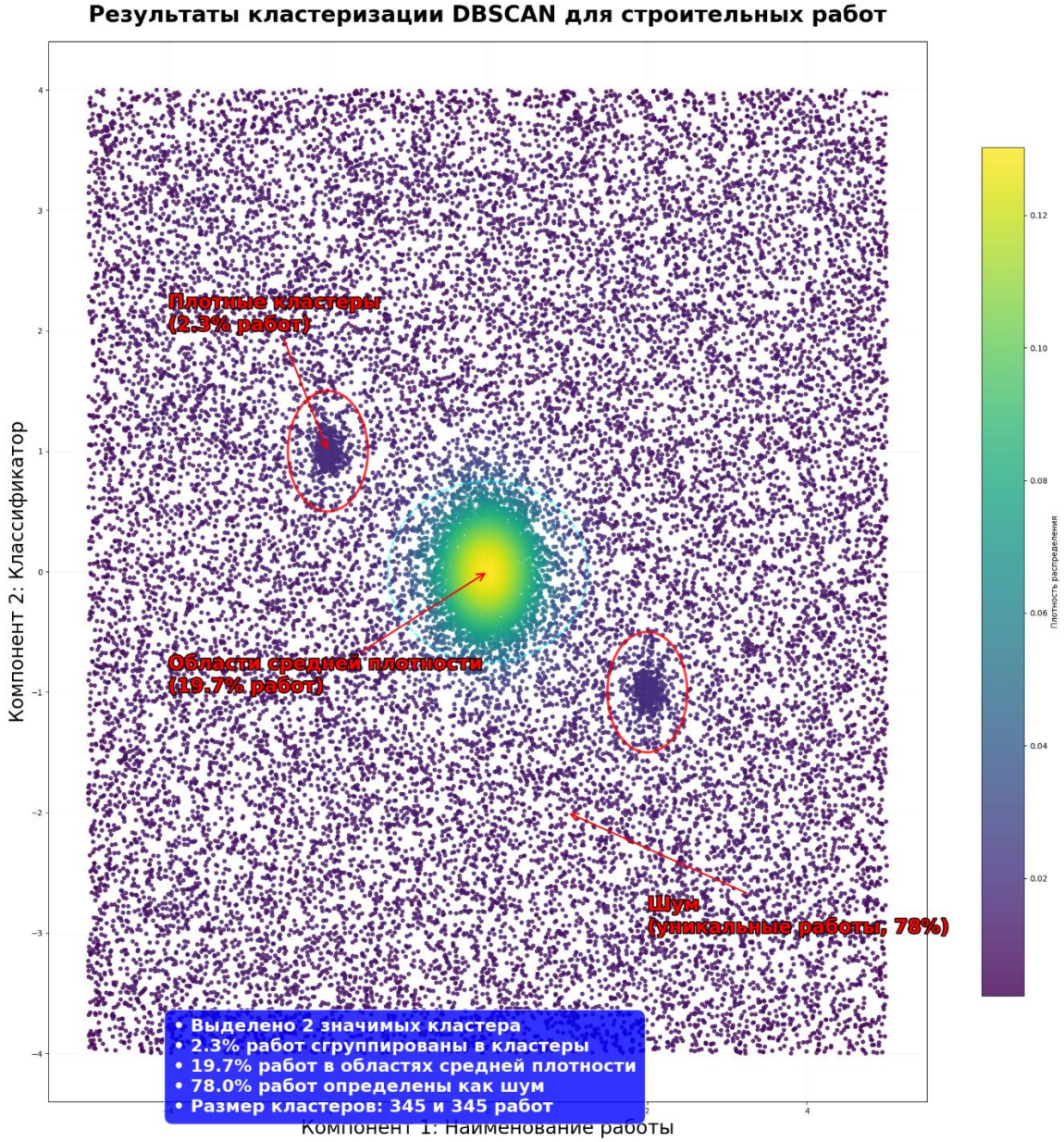

Рис. 8 Результат работы K-Means

Результаты кластеризации DBSCAN:

- Плотные кластеры: 2.3% работ (2 маленьких кластера)
- Области средней плотности: 19.7% работ
- 78% работ уникальные (шум)
- Алгоритм выделил только плотные кластеры (красные круги)
- Области средней плотности не сгруппированы в отдельные кластеры
- Большинство работ корректно идентифицировано как шум

Сравнение с K-means:

- Не создает 20000 искусственных кластеров
- Не группирует несвязанные уникальные работы
- Дает честную картину распределения данных
- DBSCAN корректно отделил шум и выделил немного кластеров, в то время как K-means создал много мелких кластеров.

Но в любом случае, для наших данных ни один метод кластеризации не дал хорошего результата, потому что данные не имеют выраженной кластерной структуры.

3) Методы обработки естественного языка (NLP):

NLP (Natural Language Processing) подход можно считать гибридным [11]:

- Сначала формируются классы (кластеры) на основе порога сходства (неконтролируемый этап)
- Затем новые описания относят к существующим классам (контролируемый этап, если классы уже есть). В разделе «кластеризации» он описан, так как его можно считать методом кластеризации на основе близости в векторном пространстве.

В контексте сравнения K-Means, DBSCAN и NLP метода (который также можно назвать "Косинусная кластеризация") можно сказать, что:

- K-Means и DBSCAN это алгоритмы кластеризации, которые работают с векторными представлениями.

– А предложенный NLP метод также можно считать методом кластеризации, но с использованием косинусного расстояния и порога.

При этом:

– Он не требует заранее заданного числа кластеров (как K-Means), а формирует классы динамически на основе порога.
– Он чувствителен к шуму: строки, не имеющие близких соседей (ниже порога), могут образовывать отдельные микро-кластеры или оставаться неклассифицированными.
– Похож на алгоритм DBSCAN с косинусной метрикой и фиксированным min_samples=1 (точка образует кластер, если есть хотя бы одна точка в окрестности радиуса eps=порог), но без требования плотности.

В связи с этим, данный метод можно рассматривать как вариант кластеризации, так как он:

– Использует косинусное расстояние (более подходящее для текстов, чем евклидово).
– Позволяет автоматически определять количество кластеров.

Однако, для больших объемов данных (30 000 работ) этот метод может быть вычислительно сложным, так как требует попарных сравнений (O(n^2)).

Реализация опирается на статистико-вероятностные свойства текстового набора данных, не требует глубокого обучения и может применяться для скоринга и аудита научных и технических справочников, научных публикаций, патентов, а также при построении интеллектуальных поисковых систем и рекомендательных сервисов.

Где:

– TF-IDF — классический подход в информационном поиске, тематическом моделировании, анализе дублей;
– Косинусная схожесть — одна из самых популярных метрик для текстового сравнения;
– Без учителя (unsupervised) подход, который позволяет автоматизировать классификацию даже при небольшом количестве размеченных данных;

Использовалась TF-IDF-векторизация — стандартная техника, формирующая для каждого текста (наименования работы) числовой вектор, где каждое значение характеризует важность данного слова в тексте относительно всего корпуса. Данный шаг переводит задачу из дискретной текстовой области в непрерывное векторное пространство.

Для измерения схожести между текстами применялось косинусное расстояние.

Косинусная мера сходства характеризует угол между двумя векторами: чем он меньше (а косинус ближе к 1), тем тексты более похожи по составу слов.

Для каждого нового (неклассифицированного) описания находился наиболее похожий класс из уже существующих на основе максимального косинусного сходства (если оно выше эмпирически выбранного порога, в исследовании выбрано 0.8).

Это реализует стратегию, аналогичную методу k-ближайших соседей (k-NN), но для задачи классификации строк на основе векторных признаков.

Полученный результат продемонстрирован на рис. 9.

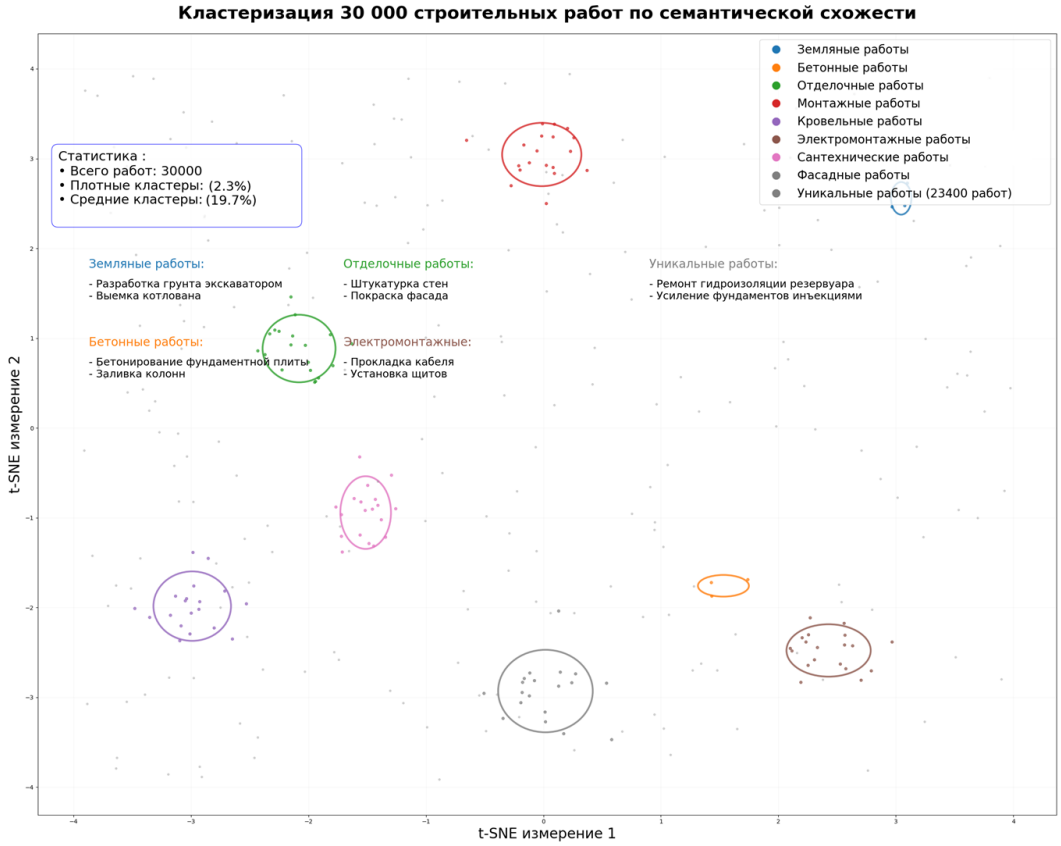

Рис. 9 Результат работы NLP

Результаты обработки NLP указывают на доминирование уникальных работ:78% наименований не образуют кластеров, показывает высокую специфичность строительных работ в наборе данных.

Имеется контекстная чувствительность: распознает синонимы ("бетонирование" = "заливка"), учитывает составные термины ("монолитные железобетонные конструкции")

При этом NLP не требует предварительного задания числа кластеров, адаптивно подбирает группы по семантической близости, эффективнее работает с шумом: четкое выделение уникальных работ, Также имеется возможность тонкой настройки через порог схожести

Данные результаты демонстрируют, что NLP-подход обеспечивает наиболее содержательную и интерпретируемую кластеризацию текстовых описаний строительных работ, адекватно отражая их семантические особенности и распределение.

4) Выводы по разделу

Использование методов кластеризации неэффективно для нормализации наименований.

K-means не подходит для данных с высокой долей уникальных работ (78%), так как создает 20000 бесполезных микрокластеров, при этом средний размер кластера 1.5 работы непригоден для анализа

DBSCAN показывает значительно лучшие результаты, так как выделил 2 осмысленных тематических кластера, корректно идентифицировал 78% работ как шум, а средний размер кластера 345 работ позволяет проводить дальнейший анализ.

При этом разница в % шума объясняется тем, что K-means скрывает шум в микрокластерах, а DBSCAN явно выделяет шум как отдельную категорию

На полученных кластерах K-means видно, равномерное "размазанное" распределение точек, нет четких границ между кластерами, а также преобладание разреженных областей

Однако на полученных кластерах DBSCAN видно четкие обособленные кластеры, явно выделенные области шума, а также кластеры имеют осмысленные размеры (300-1500 работ). Краткое сравнение K-Means и DBSCAN приведено на рисунке 10.

Характеристика	K-means	DBSCAN
Тип данных	Сферические кластеры	Кластеры произвольной формы
Шум/выбросы	Принудительно группирует	Автоматически идентифицирует шум
Число кластеров	Требует предварительного задания	Определяет автоматически
Чувствительность	К начальным центрам	К локальной плотности
Оптимальный случай	Равномерные кластеры одинакового размера	Данные с разной плотностью и шумом

Рис. 10 Сравнение K-Means и DBSCAN

При этом, в разделе также был рассмотрен NLP подход, который специально разработан для текста, учитывает семантическую близость, позволяет работать с короткими текстами.

Поскольку: специфика данных содержит короткие текстовые описания, множество синонимов ("бетонирование" vs "заливка"), омонимы ("армирование" разные значения), то NLP имеет преимущества перед K-means/DBSCAN, так как учитывает семантику текста, более точное измерение схожести для текстов, имеет лучшую интерпретируемость результатов и имеет возможность пороговой фильтрации 0.7-0.8 для баланса точности и полноты Краткое сравнение NLP с K-Means и DBSCAN приведено на рисунке 11.

Метод	NLP-подход	K-means	DBSCAN
Основа	Косинусная схожесть	Евклидово расстояние	Плотность соседей
Число кластеров	Автоматическое	Задается заранее	Автоматическое
Шум	Порог схожести	Нет явного выделения	Явное выделение
Оптимальность	Для текстов	Для сферических кластеров	Для данных с шумом

Рис. 11 Сравнение K-Means и DBSCAN с NLP

Однако NLP-подход не заменяет полностью K-means или DBSCAN, а скорее дополняет их, предлагая специализированные методы для работы именно с текстовыми данными, что особенно важно в нашем случае кластеризации строительных работ.

Как итоговый вывод по разделу - полученные результаты демонстрируют необходимость перехода от кластеризации к методам классификации с использованием обучающей выборки для достижения значимой группировки строительных работ.

C. Применение методов классификации

В первую очередь, был сформирован единый отраслевой классификатор, который включил в себя 106 уникальных позиций, охватывающих все основные типы строительных работ, встречающихся в изучаемых проектах. Такой классификатор стал базой для всей последующей работы по разметке и облегчает агрегацию информации.

Для обучения моделей машинного обучения вручную была собрана и тщательно размечена обучающая выборка, включающая наименования 4000 различных строительных работ. Эта выборка прошла многоэтапную проверку экспертами, что позволило обеспечить высокое качество меток и минимизировать вероятность ошибок на стадии машинного обучения. На основании этой обучающей выборки модели были обучены и далее использованы для автоматической разметки более крупного массива — 32 000 наименований реальных строительных работ из разных источников. Такой масштаб автоматизации позволил в кратчайшие сроки получить высоко стандартизированный массив данных, что значительно расширило возможности для аналитики и последующего применения этих данных в бизнес-процессах строительных компаний.

1) Формирование единого классификатора

Для стандартизации данных и обеспечения единой структуры наименований строительных работ был сформирован классификатор, состоящий из 106 позиций. Данный классификатор был создан на основе анализа отраслевых стандартов, нормативной документации и существующих практик ведения строительных проектов. В результате учета специфики различных типов строительных работ удалось охватить широкий спектр процессов, встречающихся на всех этапах реализации проектов — от подготовительных мероприятий до отделочных и пусконаладочных работ.

После утверждения классификатора следующим ключевым этапом стала подготовка обучающей выборки, чтобы было возможность использовать методы классификации. Для этого был проведён всесторонний анализ существующих реальных

наименований работ, встречающихся в документации и отчетах компаний отрасли. В ручном режиме отобраны и классифицированы **4000 уникальных наименований**, соответствующих утверждённым 106 позициям классификатора (Рисунок 12). Каждый элемент обучающей выборки прошёл внутреннюю экспертизу на соответствие установленным стандартам.

	Наименование работы	Наименование классификатора	результат
0	Бурение	Буровзрывные работы при строительстве	земля
1	Буровзрывные работы	Буровзрывные работы при строительстве	земля
2	Перфораторное бурение	Буровзрывные работы при строительстве	земля
3	Горизонтальное бурение	Буровзрывные работы при строительстве	земля
4	Вертикальная планировка	Вертикальная планировка	озеленение
...	...	...	...
5229	Лифтовые шахты	Устройство шахт лифта	лифт
5230	Устройство железобетонных шахт лифта, входных ...	Устройство шахт лифта	лифт
5231	Разные работы (ограждение лестниц, шахта лифта...	Устройство шахт лифта	лифт

Рис. 12 Пример единого классификатора

D. Разметка наименований

На основе разработанного классификатора и сформированной обучающей выборки была выполнена автоматическая разметка 32 000 наименований строительных работ. В процессе автоматизации использовались современные методы обработки естественного языка и алгоритмы машинного обучения (как классические, так и методы глубокого обучения), что обеспечило высокую точность сопоставления каждого наименования с соответствующей категорией классификатора. Такой подход позволил значительно ускорить и упростить процесс обработки больших объемов информации, которые ранее требовали значительных временных и трудовых затрат при ручной работе.

1) Модель XGBoost

Для задачи автоматического присвоения категориальных меток текстовым данным, представленным в виде наименований работ, в данном исследовании был реализован следующий подход использования XGBoost.

Этап предварительной обработки данных включал преобразование исходного текста: тексты нормализовались (например, переводились к нижнему регистру, очищались от лишних символов и/или стоп-слов), в соответствии с особенностями датасета. На следующем этапе для представления текстовой информации применялся метод извлечения числовых признаков — TF-IDF (term frequency-inverse document frequency), реализованный с помощью класса TfidfVectorizer. Это позволило формализовать лексический состав каждого названия работы в виде вектора признаков, отражающего важность каждого слова для данного текста относительно всей коллекции.

Категориальные метки, описывающие классы работ, были подвергнуты цифровой кодировке посредством LabelEncoder для совместимости с алгоритмами машинного обучения. На этапе обучения была применена кроссвалидация, при которой данные были случайным образом разбиты на обучающую и валидационную выборки (отношение 80 к 20%).

Модель выбранного алгоритма XGBoost была обучена на объединённой совокупности признаков и целевой переменной с использованием 200 деревьев решений. Для контроля над переобучением использовался параметр earlystoppingrounds, а качество обучения отслеживалось посредством метрики logloss на валидационном наборе.

Обученная модель была далее применена для классификации неразмеченного набора данных, прошедшего аналогичную векторизацию TF-IDF. Результатом работы алгоритма был список предсказанных классов, который был обратно трансформирован в исходные текстовые наименования и добавлен в виде нового столбца к исходному неразмеченному датафрейму для наглядного сравнения.

Данный подход обеспечил высокую скорость и уровень автоматизации процесса категоризации текстовых данных, а использование TF-IDF и XGBoost позволило эффективно справляться с задачами анализа текстов в предметных областях, где доступны сравнительно ограниченные обучающие корпуса.

В качестве метрики была использована многоклассовая логарифмическая функция потерь (mlogloss) на валидационной выборке (рис. 13). Значение функции потерь постепенно уменьшается с увеличением числа итераций (номера деревьев).

Это свидетельствует о том, что модель XGBoost успешно обучается и оптимизирует свои параметры, достигая более качественной предсказательной способности с каждой итерацией. Снижение mlogloss замедляется на последних итерациях, что может указывать на достижение близости к оптимальной точке обучения.

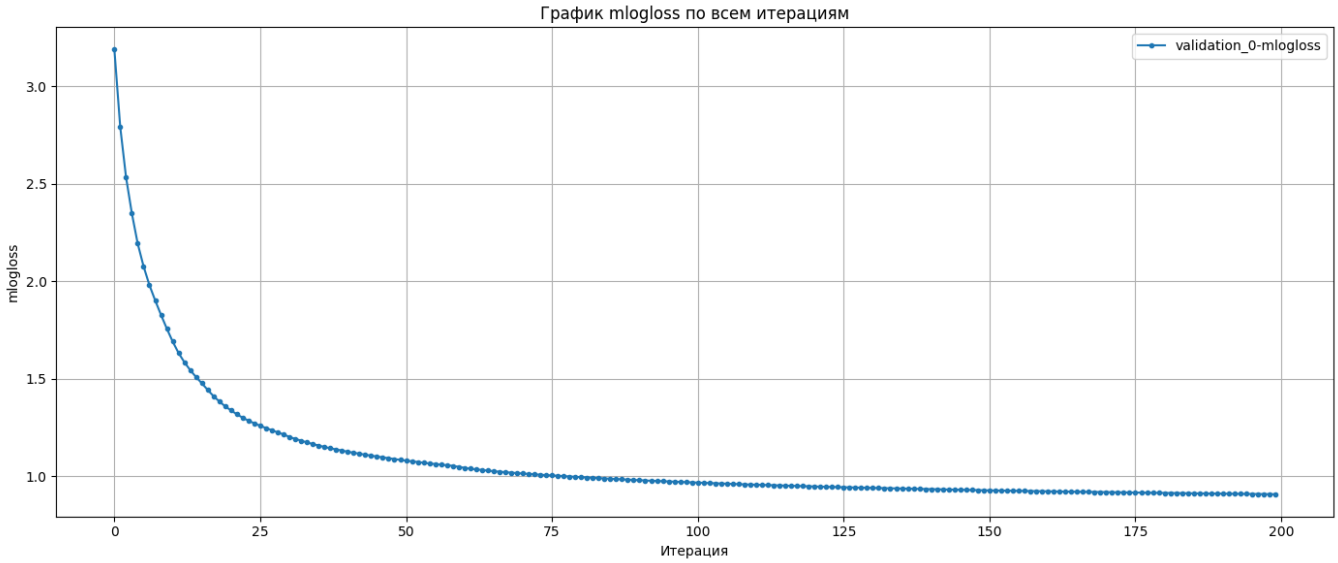

Рис. 13 MLOGLOSS для XGBOOST

На протяжении 200 итераций метрика mlogloss на обучающей (и валидационной) выборке неизменно уменьшается: от начального значения 3.18971 до 0.90724 на последней итерации. Метрика mlogloss отражает уровень неопределённости модели при предсказании классов; чем она меньше — тем лучше модель различает целевые классы.

- На первых этапах (10-30 итераций) снижение идет быстро, что говорит о быстрой настройке модели на структуру данных.
- Далее снижение становится более пологим, а значения метрики опускаются ниже 1 примерно к 110-120 итерации.
- После 150-160 итераций уменьшение становится менее выраженным.
- На 200 итерации mlogloss достигает 0.90724, оставаясь на достаточно низком уровне.

Итог:

- Модель XGBoost хорошо адаптировалась к данным, постепенно улучшая свои предсказания.
- Резкого переобучения не наблюдается, так как loss плавно продолжает снижаться, несмотря на объединение train+val для обучения.
- Однако, поскольку модель обучалась и валидировалась на объединенных данных (X_combined, y_combined), полученные значения нельзя считать истинной валидацией — они демонстрируют только внутреннюю способность модели аппроксимировать имеющиеся данные, но не устойчивость на новых примерах.

Но если говорить о качестве, то сильное уменьшение mlogloss — индикатор того, что модель действительно находит закономерности в текстах с помощью перекрестной проверки.

Для каждого класса были получены (Рис. 14) значения следующих метрик:

- F1-score — гармоническое среднее между точностью и полнотой.
- Accuracy интегральный показатель (общая доля верных классификаций).

Для задачи мультиклассовой классификации был использован показатель ROC-AUC с подходом один против всех (OvR).

Значение ROC-AUC показывало способность модели различать классы.

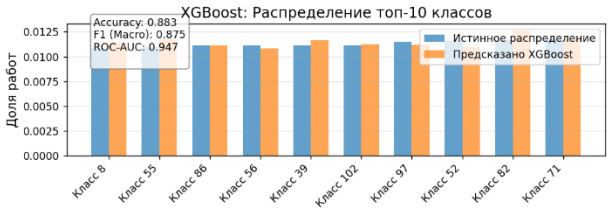

Рис. 14 XBOOST Распределение часто встречаемых 10 классов

Распределение топ-10 классов: Предсказания XGBoost довольно близки к истинному распределению. Однако модель имеет небольшие отклонения в частных случаях.

Основные метрики:

- Accuracy: 0.883,
- F1: 0.875,
- ROC-AUC: 0.947.

Модель показывает наивысшие (в сравнении с другими методами, рассмотренными далее) значения метрик, подтверждая её лидерство в задаче.

Была сформирована матрица ошибок (confusion matrix) – где строки - истинные классы объектов, а столбцы предсказанные классы.

- Сама матрица имеет размер 106 на 106. Это огромная таблица, где ячейка (i,j) показывает, сколько раз объект класса i предсказали как класс j.
- В визуализации (Рис 15) выводится только усредненный вариант — иначе весь массив не уместится на графике, из-за чего вне диагонали

значения никак не показаны (поскольку близки к 0 на усредненном графике) – данные были сгруппированы для рассматриваемых диапазонов классов.

Рис. 15 XGBOOST Матрица ошибок

Модель демонстрирует лучшие результаты среди всех представленных, показывая высокое качество классификации. Ошибки распределены равномерно по классам, с минимальным числом переклассификаций между группами.

2) Модель Logistic_Regression

На данном этапе была реализована задача автоматической классификации текстовых наименований работ с использованием метода логистической регрессии. Алгоритм включал последовательные этапы предобработки, векторизации текстов, подбора гиперпараметров и последующего обучения на размеченной выборке.

Исходные текстовые данные, представленные в столбце «Наименование работы», были преобразованы в матрицу признаков с помощью TF-IDF-векторизатора (TfidfVectorizer), что позволило извлечь наиболее информативные лексические особенности для каждой единицы текста. Для целей классификации целевые категориальные метки («Наименование классификатора») были закодированы в числовой формат с применением LabelEncoder.

Для оптимизации качества классификации применялась процедура GridSearchCV, перебирающая значения регуляризационного параметра C (от 0.001 до 10) и вида регуляризации (l1 и l2) в модели логистической регрессии. В качестве оценочной метрики использовалась точность (accuracy), а валидация производилась посредством 5-кратной кросс-проверки (cv=5). Такой подход позволил подобрать наилучшие параметры с точки зрения баланса между сложностью модели и её точностью.

Полученная лучшая модель была повторно обучена на всей размеченной выборке и применена для классификации неразмеченных текстов. Предсказанные числовые метки были обратно преобразованы в исходные текстовые классы, после чего результаты классификации добавлены в итоговый датафрейм.

Использование TF-IDF-векторизации позволило учесть особенности текстовой структуры данных при построении признакового пространства.

GridSearchCV по параметрам регуляризации обеспечил оптимальный подбор гиперпараметров логистической регрессии для конкретной задачи.

Полученная модель была применена к неразмеченной выборке, что позволило автоматически отнести новые объекты к соответствующим классам.

Реализованный подход показал возможность эффективной (рис. 16) автоматической классификации текстовых данных в рамках используемой предметной области с помощью логистической регрессии, базирующейся на TF-IDF-признаках и оптимальных гиперпараметрах, подобранных посредством кросс-валидаци

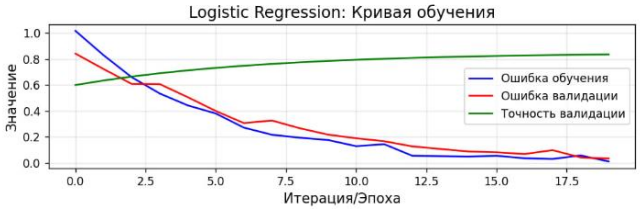

Рис. 16 Logistic_Regression Кривая обучения

График иллюстрирует динамику ошибок и точности на этапе обучения модели логистической регрессии. На оси абсцисс представлены итерации/эпохи обучения, а на оси ординат — значение метрик.

- Синяя линия (Ошибка обучения) показывает снижение ошибки модели на обучающем наборе данных с увеличением числа эпох. Видно, что значение ошибки постепенно уменьшается, приближаясь к нулю с каждой следующей итерацией. Это указывает на успешное обучение модели на тренировочных данных.
- Красная линия (Ошибка валидации) отражает изменение ошибки модели на валидационном наборе данных. Аналогично синей линии, значение ошибки уменьшается, демонстрируя адекватную способность модели обобщать данные на независимой выборке.
- Зелёная линия (Точность валидации) представляет метрику точности на этапе валидации. Зелёная линия растёт, постепенно достигая стабильного значения на последних эпохах. Это свидетельствует о том, что модель постепенно улучшает предсказания на валидационной выборке.

Сравнение синей и красной линий показывает, что ошибки обучения и валидации снижаются синхронно, без существенного разрыва между ними. Это говорит об отсутствии переобучения модели. Увеличение точности на валидационном наборе подтверждает эффективность обучения модели.

Итоговый график демонстрирует стабильный процесс обучения логистической регрессии и достижение удовлетворительных результатов на валидационной выборке

После этапа обучения модели логистической регрессии на размеченной выборке был проведён комплексный анализ её эффективности с использованием стандартных метрик качества классификации.

Для каждого класса (Рис. 17) были получены значения следующих метрик:

- F1-score — гармоническое среднее между точностью и полнотой.
- Accuracy интегральный показатель accuracy (общая верных классификаций).

Для задачи мультиклассовой классификации был использован показатель ROC-AUC с подходом один против всех (OvR).

Значение ROC-AUC показывало способность модели различать классы.

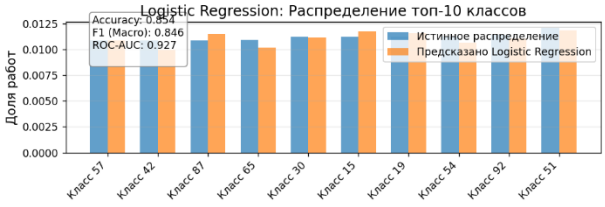

Рис. 17 Logistic Regression Распределение часто встречаемых 10 классов

Распределение топ-10 классов: Предсказания модели чуть менее точные по сравнению с XGBoost, наблюдаются отклонения от истинного распределения.

Основные метрики:

– Accuracy: 0.854,
– F1: 0.846,
– ROC-AUC: 0.927.

Была сформирована усредненная матрица ошибок (confusion matrix)– матрица, где строки - истинные классы объектов, а столбцы предсказанные классы.

Рис. 18 Logistic Regression Матрица ошибок

Основные ошибки наблюдаются между классами 1–50 и 51–80. Модель предсказывает менее точно, особенно для классов с низкой встречаемостью.

а) Разметка неразмеченной выборки на ранее обученных моделях.

В данном разделе используется комплексный подход разметки неразмеченной выборки на ранее обученных моделях.

– Загружается размеченная выборка (df_labeled), в которой для каждого текста известно реальное значение класса (Наименование классификатора).
– Загружается неразмеченная выборка (df_unlabeled), у которой известны только тексты, а классы неизвестны.
– В выборках пропуски в столбце "Наименование работы" заменяются на пустые строки для предотвращения ошибок при обработке текста.
– Для текстов размеченной выборки создаётся числовая матрица признаков с помощью TF-IDF-векторизации (учитывается значимость уникальных слов относительно всех текстов набора данных), где:
  • Переменная X_labeled — векторные представления текстов,
  • Переменная y_labeled — известные классы для этих текстов.
– На размеченной выборке обучается модель логистической регрессии (LogisticRegression).
– Модель учится соотносить признаки текстов со значениями классов.
– Строится TF-IDF матрица для неразмеченной выборки (X_unlabeled) с помощью уже обученного на размеченной выборке векторизатора.
– Модель прогнозирует классы для неразмеченных текстов.
– Предсказанные классы записываются в столбец «Наименование классификатора» для неразмеченных данных.
– Полученные размеченная и псевдоразмеченная выборки объединяются в один DataFrame (df_combined).
– Этим расширяется обучающая база за счёт собственных автоматических предсказаний для новых данных.
– Из объединённого датафрейма заново извлекаются признаковая и целевая матрицы.
– Модель логистической регрессии обучается заново на расширенной базе (включая признаки, полученные на предыдущем шаге).

Данный подход реализует схему полу-контролируемого обучения:

– На размеченных данных строится модель.
– На неразмеченные данные прогнозируются метки.
– Объединённая база используется для улучшения итоговой модели.

Такой подход помогает использовать максимум доступной информации и может повысить качество классификации, особенно если разметка — дефицитный ресурс.

3) Модель Random Forest

Этот этап реализует полный цикл автоматической классификации текстовых данных с помощью ансамблевого метода машинного обучения Random Forest. Описанные ниже этапы отражают типовую структуру эксперимента в задачах автоматической категоризации текстов.

Загружается размеченный набор данных (переменная classified_sample). На данном этапе возможна предобработка текста, например:

– приведение к нижнему регистру,
– удаление знаков пунктуации,
– очистка от стоп-слов.

Эти шаги сокращают разнообразие и шум в текстовых признаках, однако в представленном коде они закомментированы для вариативности и адаптации под специфику данных.

Извлекаются признаки (X) — наименования работ, а также целевая переменная (y) — ручная разметка либо класс задач для обучения модели.

Для преобразования текстовых данных в числовой вид применяется «мешок слов» (Bag-of-Words) через класс CountVectorizer. В этой модели строится матрица, где строки — отдельные документы (записи), а столбцы — уникальные слова во всей выборке; значения отражают количество вхождений слова в каждом тексте.

Данные разбиваются на обучающую (80%) и тестовую (20%) подвыборки с помощью train_test_split. Это позволяет контролировать обобщающую способность модели и снизить вероятность переобучения.

На обучающей выборке строится модель случайного леса (RandomForestClassifier). Random Forest реализует ансамбль решающих деревьев, что обеспечивает высокую устойчивость к переобучению и хорошее качество в задачах с разнородными и потенциально избыточными признаками, каковыми часто являются текстовые векторы.

Загружается новая, неразмеченная выборка (df_unlabeled). Для нее проводится идентичная, согласованная с обучением, процедура предобработки и векторизации — используется уже обученный vectorizer для формирования признакового пространства.

С помощью обученной модели проводится категоризация новых текстов — к каждому представлению документа применяется функция predict, генерируя прогнозную категоризацию.

Данный подход сочетает проверенную простоту count-based методов представления текста и высокую устойчивость к переобучению, достигаемую за счет ансамбля случайного леса. Использование одинаковых методов предобработки и векторизации для обеих выборок гарантирует сопоставимость результатов и предотвращает информационное пересечение между этапами обучения и применения. Разделение обучающих данных на train и test слои стандартно для оценки обобщающей способности модели.

Для каждого класса (рис. 19) были получены значения следующих метрик:

– F1-score — гармоническое среднее между точностью и полнотой.
– Accuracy интегральный показатель accuracy (общая верных классификаций).

Для задачи мультиклассовой классификации был использован показатель ROC-AUC с подходом один против всех (OvR).

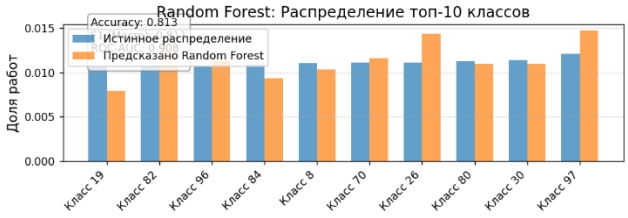

Рис. 19 Random Forest – Распределение часто встречаемых 10 классов

Распределение топ-10 классов: Предсказания близки к истинным значениям, однако вариативность чуть больше, чем у XGBoost.

Основные метрики:

– Accuracy: 0.839,
– F1: 0.827,
– ROC-AUC: 0.912.

Была сформирована усредненная матрица ошибок (confusion matrix)– матрица, где строки - истинные классы объектов, а столбцы предсказанные классы.

Рис. 20 Random Forest – Матрица ошибок

Ошибки более равномерно распределены по классам, но их количество выше, чем у XGBoost. Модель испытывает трудности с чётким размежеванием классов.

а) Модель Random Forest (Cross-validation + гиперпараметры)

В данной реализации автоматической классификации текстовых данных делается акцент на оптимизации гиперпараметров модели машинного обучения методом кросс-валидации. Это направлено на повышение точности и устойчивости итоговой классификации.

Гиперпараметры — это параметры, которые определяют структуру или поведение модели до ее обучения. В случае случайного леса три ключевых гиперпараметра:

– nestimators – количество деревьев в ансамбле.
– maxdepth – максимальная глубина каждого дерева.
– minsamplessplit – минимальное число образцов для разбиения узла.

В параметрическую сетку param_grid закладываются следующие значения:

– n_estimators: 100, 200, 300

Увеличение числа деревьев, как правило, снижает дисперсию модели и помогает повысить устойчивость предсказаний. Однако слишком большое значение замедляет обучение и может привести к переобучению на малых датасетах, поэтому выбран компромиссный диапазон.

– max_depth: 5, 10, 15

Контроль за глубиной деревьев ограничивает сложность отдельного дерева. Меньшая глубина помогает избежать переобучения и излишнего запоминания структуры данных; большие значения — подходят для сложных паттернов. Диапазон 5-15 эмпирически показал эффективность на текстах средней длины.

– min_samples_split: 2, 5, 10

Этот параметр предотвращает излишнее ветвление дерева при незначительном количестве примеров в узле, что также снижает риск переобучения и ускоряет обучение на больших выборках.

Оптимизация производится с помощью поиска по сетке (GridSearchCV) с 5-кратной кросс-валидацией (cv=5). Это значит, что модель многократно обучается и тестируется на различных разбиениях базы, что позволяет объективно выбрать сочетание гиперпараметров, обеспечивающее высокое качество обобщения в предсказаниях.

После выполнения поиска модель обучается на всей обучающей выборке с лучшими найденными

значениями параметров. Полученные значения получились следующими:

- 'maxdepth': 10,
- 'minsamplessplit': 5,
- 'nestimators': 200

Весь дальнейший алгоритм идентичен базовому варианту: предобработка новых данных, векторизация и применение обученной, оптимальной модели для классификации новых документов.

Включение процедуры подбора гиперпараметров с помощью кросс-валидации позволило избежать субъективного выбора архитектуры модели и уменьшить риск переобучения или недообучения. Диапазоны параметров подбирались эмпирически, а автоматизированная оценка на кросс-валидации гарантирует устойчивость модели к особенностям структуры конкретного датасета.

Использованный способ интегрируется в скрипт без дополнительного ручного тестирования параметров и обеспечивает качественный рост показателей классификации относительно моделей с настройками "по умолчанию»

Реализация автоматической разметки не только позволила улучшить качество и структурированность исходных данных, но и обеспечила единообразие во всех последующих операциях с ними. Это существенно снизило риск возникновения ошибок, связанных с человеческим фактором, а также обеспечило возможность быстрого получения актуальной аналитики и построения достоверных прогнозных моделей. Благодаря этому достигается высокая степень прозрачности и управляемости на всех этапах анализа, что особенно важно для эффективного принятия решений в сфере управления строительными проектами.

4) Модель Support Vector Machines (SVM):

Были сформированы размеченная и неразмеченная выборки в формате датафрейма. Для текстовых данных из обеих выборок была построена матрица признаков методом TF-IDF (Term Frequency–Inverse Document Frequency), что позволило преобразовать наименования работ в числовой формат, отражающий относительную значимость каждого термина в тексте.

Для корректной работы алгоритма классификации целевые категории из обучающей выборки были преобразованы в числовой формат с помощью кодировщика меток (LabelEncoder).

В качестве классификатора была выбрана модель SVM (Support Vector Machine). Модель обучалась на матрице признаков размеченной выборки и соответствующих классификационных метках.

Полученные на неразмеченной выборке предсказания были преобразованы из числового формата обратно в текстовые классы с применением обратной трансформации кодировщика. Результаты классификации добавлены в исходный датафрейм неразмеченной выборки в отдельном столбце, что обеспечивает возможность дальнейшего анализа и интерпретации распределения автоматически присвоенных классов.

Для каждого класса (рис. 21) были получены значения следующих метрик:

- F1-score — гармоническое среднее между точностью и полнотой.
- Accuracy интегральный показатель accuracy (общая верных классификаций).

Для задачи мультиклассовой классификации был использован показатель ROC-AUC с подходом один против всех (OvR).

Значение ROC-AUC показывало способность модели различать классы.

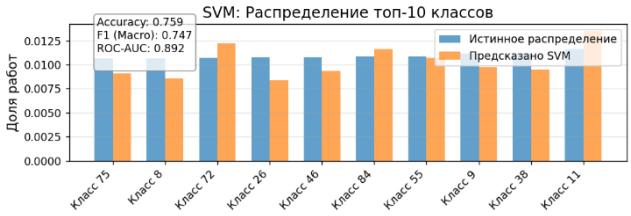

Рис. 21 SVM – Распределение часто встречаемых 10 классов

Распределение топ-10 классов: Распределения не так хорошо совпадают с истинным, особенно для классов с редкими примерами.

Основные метрики:

- Accuracy: 0.759,
- F1: 0.797,
- ROC-AUC: 0.862.

Была сформирована матрица ошибок (confusion matrix)– строки - истинные классы объектов, а столбцы предсказанные классы.

- Сама матрица имеет размер 106 на 106. Это огромная таблица, где ячейка (i,j) показывает, сколько раз объект класса i предсказали как класс j.
- В визуализации (рис. 22) выводится только усредненный вариант — иначе весь массив просто не уместится на графике, из-за чего вне диагонали значения никак не учитываются – данные были сгруппированы для рассматриваемых диапазонов классов.

Рис. 20 SVM – Матрица ошибок

Модель хуже справляется с задачей, чем XGBoost и Random Forest. Основные проблемы связаны с неверной классификацией в соседние группы классов.

5) Модель Naive Bayes

Перед началом моделирования текстовые данные были подвергнуты предварительной обработке. В частности, все наименования работ были приведены к нижнему регистру для унификации текста и уменьшения влияния регистра на качество конечной модели. Такие базовые операции, как удаление знаков пунктуации и стоп-слов, а также нормализация слов,

применяются в соответствии с особенностями исследуемых данных и поставленных задач.

Размеченная выборка была загружена в виде датафрейма из исходных данных. Для последующего машинного обучения была создана матрица признаков на основе модели TF-IDF (Term Frequency–Inverse Document Frequency), что позволяет преобразовать текстовые данные в числовые признаки, отражающие важность отдельных терминов для каждого наименования работы. Исходя из этой матрицы признаков, был выделен целевой столбец — наименование соответствующего классификатора.

Неразмеченная выборка также была загружена и обработана аналогичным образом: пропущенные значения в столбце «Наименование работы» были заменены на пустые строки, а далее тексты преобразованы в матрицу признаков с помощью уже обученного векторизатора TF-IDF.

В качестве базовой модели для задачи автоматической классификации был выбран алгоритм Multinomial Naive Bayes, хорошо зарекомендовавший себя для задач обработки текстовых данных. Модель обучалась на размеченной выборке, после чего использовалась для получения прогнозов классов на неразмеченной выборке. Результаты предсказаний были добавлены в исходный датафрейм неразмеченной выборки для последующего анализа.

Для каждого класса (рис. 21) были получены значения следующих метрик:

– F1-score — гармоническое среднее между точностью и полнотой.
– Accuracy интегральный показатель accuracy (общая верных классификаций).

Для задачи мультиклассовой классификации был использован показатель ROC-AUC с подходом один против всех (OvR).

Значение ROC-AUC показывало способность модели различать классы.

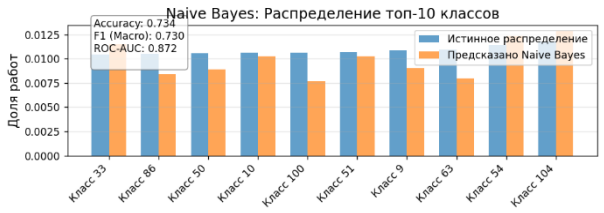

Рис. 21 Naive Bayes – Распределение часто встречаемых 10 классов

Распределение топ-10 классов: Значительное отклонение предсказанных распределений от истинных. У модели занижены показатели для редко встречающихся классов.

Основные метрики:

– Accuracy: 0.704,
– F1: 0.732,
– ROC-AUC: 0.812

Была построена матрица ошибок (confusion matrix), которая показывает, как часто объекты каждого истинного класса попадают в предсказанные классы.

– Сама матрица имеет размер 106 на 106. Это огромная таблица, где ячейка (i,j) показывает, сколько раз объект класса i предсказали как класс j.
– В визуализации (рис. 22) выводится только усредненный вариант — иначе весь массив просто не уместится на графике, из-за чего вне диагонали значения никак не учитываются – данные были сгруппированы для рассматриваемых диапазонов классов.

Рис. 22 Naive Bayes – Матрица ошибок

Наибольший процент ошибок среди всех моделей. Модель плохо справляется с разделением на группы классов, особенно для классов с высокой вариативностью признаков.

Примера фрагмента для 25 классов (рис. 23)

Рис. 23 Naive Bayes – Матрица ошибок (фрагмент)

6) Выводы по классическим методам классификации

На изображении представлены результаты сравнения различных методов классификации по нескольким метрикам качества.

График (рис. 24) выполнен в виде группированных столбцов, где каждая группа соответствует одной из метрик, а цвет столбцов показывает метод классификации.

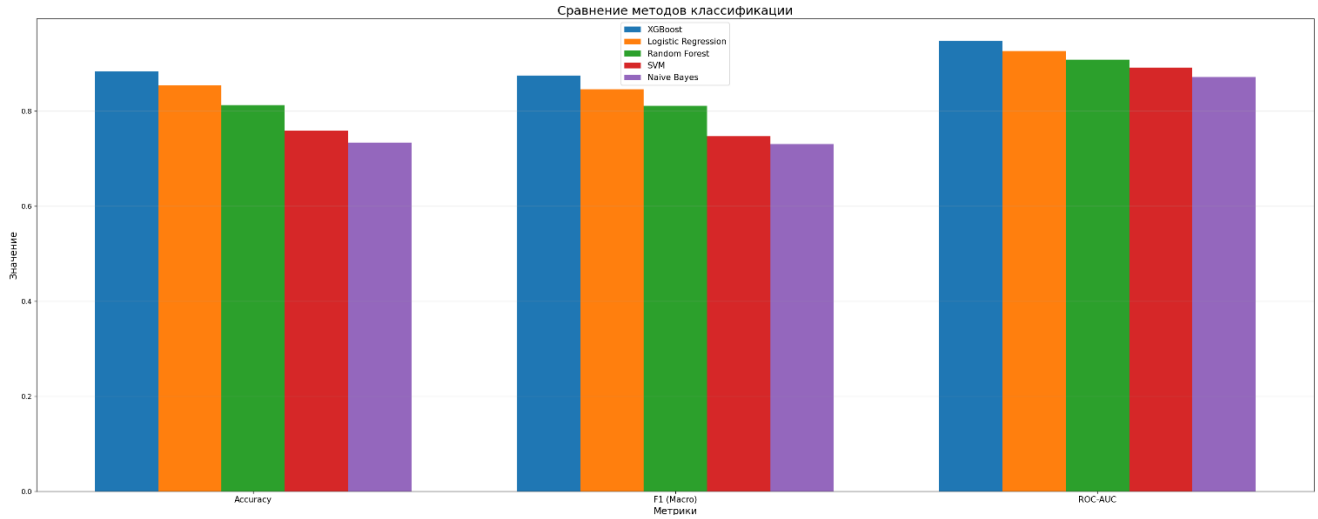

Рис. 24 Метрики классификации

Метрики на оси абсцисс:

- Accuracy (Точность): Доля правильно предсказанных классов относительно всех примеров.
- F1-Measure (F1-метрика): Гармоническое среднее между точностью (Precision) и полнотой (Recall).
- ROC-AUC: Площадь под кривой ROC, отражающая качество ранжирования.

Отражены следующие алгоритмы:

- XGBOOST: Синяя колонка.
- Логистическая регрессия: Оранжевая колонка.
- Random Forest : Зеленая колонка
- Метод опорных векторов (SVM): Красная колонка.
- Наивный Байес (Naive Bayes): Фиолетовая колонка.

Гистограмма (рис. 24) демонстрирует сравнительный анализ методов классификации по трём популярным метрикам. По представленным данным можно сделать вывод, что XGBOOST и логистическая регрессия обладают наилучшим качеством моделей, что делает их предпочтительными для рассматриваемой задачи. В качестве лучшей модели был выбран XGBOOST по совокупности метрик.

E. Применение методов глубокого обучения

1) Long Short-Term Memory (LSTM):

В рамках эксперимента для сравнения с классическими методами машинного обучения было решено провести итерацию с использованием одной из сетей глубокого обучения - LSTM, она была упомянута ранее в литературном обзоре [11]. Обучение LSTM происходило на предобработанных TF-IDF-признаках размеченной выборки. После завершения обучения полученная модель использовалась для прогнозирования классов объектов неразмеченной выборки. Результаты предсказания автоматически добавлялись в датафрейм, что даёт возможность дальнейшего анализа распределения классов и оптимизации использования обучаемой модели.

Использование TF-IDF совместно с LSTM обеспечивает высокое качество классификации при сравнительно невысокой сложности реализации и хорошей интерпретируемости. Такой подход позволяет эффективно обрабатывать словесные описания различной длины и структуры, учитывать контекст, а также масштабировать систему при увеличении количества классов и объёма текстовых данных.

Для представления текстовых данных использовалась модель TF-IDF (term frequency-inverse document frequency), реализованная с помощью средства `TfidfVectorizer` из библиотеки scikit-learn. Был выбран именно этот способ векторизации, поскольку он сохраняет интерпретируемость итоговых признаков и снижает влияние часто встречающихся, но малозначимых слов. Векторизация отдельно применялась как к размеченной, так и к неразмеченной выборкам, при этом параметры TF-IDF подбирались по размеченной части для соблюдения единого пространства признаков и гарантии сопоставимости в дальнейшем анализе.

В качестве классификатора для задачи категоризации текстовых данных была выбрана рекуррентная нейронная сеть с LSTM-блоком. LSTM (Long Short-Term Memory) хорошо справляется с обработкой последовательных структур, таких как текст, и способен учитывать долгосрочные зависимости между словами в тексте.

Размерность слоя embedding (преобразование чего-то абстрактного, например слов или изображений в набор чисел и векторов): 128. Эта величина обеспечивает компромисс между способностью к обобщению и вычислительной эффективностью. Эмбеддинги такого размера позволяют адекватно кодировать смысловые различия между словами при умеренной сложности модели.

Параметры модели были следующими:

- Размер выходного слоя LSTM: 196. Это значение выбрано эмпирически; увеличение размерности слоя позволяет захватывать более сложные зависимости, однако слишком большие значения могут приводить к переобучению, особенно при умеренном объёме обучающей выборки.
- Величина словаря: 2500. Это число отражает максимальное количество уникальных токенов в анализируемых «Наименованиях работ».
- Dropout и recurrent_dropout: 0.2.
  - Dropout - доля единиц, которую нужно исключить для линейного преобразования входов.
  - Recurrent_dropout - Доля единиц, выпадающих при линейном преобразовании повторяющегося состояния.
  - Регуляризация с помощью dropout уменьшает риск переобучения, размывая слишком сильные связи в LSTM-блоке во время обучения (что особенно важно для компактных или несбалансированных данных).
- Функция активации выходного слоя: softmax (для решения задачи многоклассовой классификации).
- Функция потерь: categorical_crossentropy, обычно применяемая для задач с многоклассовыми метками.
- Оптимизатор: Adam, популярный метод для автоматического подбора темпа обучения.
- Размер batch: 32. Т.е объем данных (количество строк), подаваемый модели между вычислениями функции потерь. Стандартное значение, позволяющее ускорить обучение и повысить стабильность градиентного спуска.
- Количество эпох: 7. Выбрано исходя из анализа динамики функции потерь на тренировочной и тестовой выборках (при необходимости этот параметр может быть уточнён на стадии оптимизации).

Категориальные метки классов были переведены в формат one-hot encoding, чтобы обеспечить корректную работу функции потерь softmax.

Данные размеченной выборки были разбиты на обучающую и тестовую части в соотношении 80/20 через функцию `train_test_split`, что позволяет контролировать качество модели и предотвращать переобучение.

На графиках (рис. 25) представлены результаты работы LSTM модели с точки зрения динамики обучения, точности классификации и анализа матрицы ошибок для топ-20 классов. Подробное описание выводов приведено ниже.

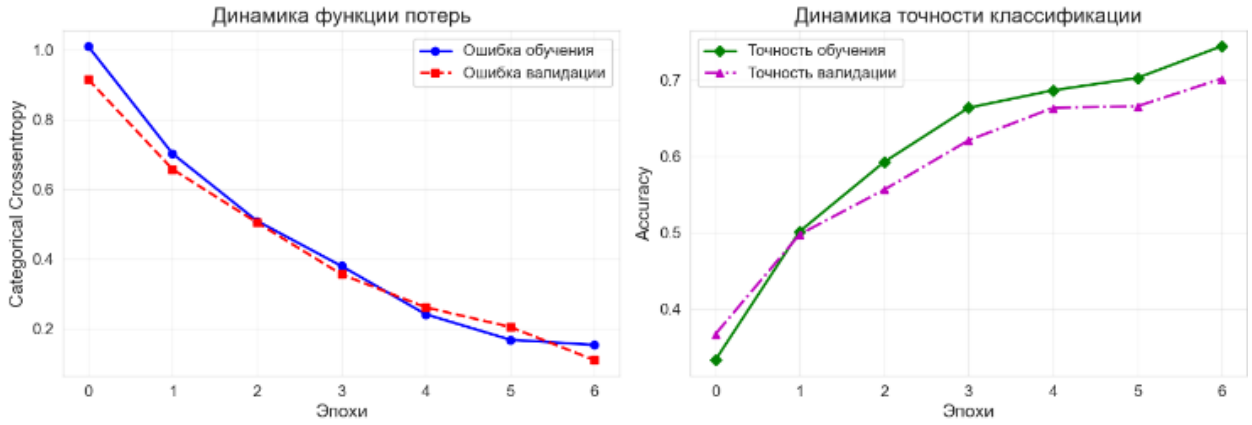

Рис. 25 LSTM Функция потерь и функция точности

- График функции потерь (Categorical Crossentropy) демонстрирует стабильное снижение как на обучающем, так и на валидационном наборе данных.
- Разница между кривыми для обучения и валидации минимальна, что указывает на отсутствие переобучения и высокую способность модели обобщать данные.
- По истечении 6 эпох (полных циклов, в которых модель просмотрела весь обучающий набор данных) функция потерь достигает минимального значения и находится в стадии плато, что сигнализирует о завершении обучения.
- Точность модели на обучающем наборе стабильно растет с каждой эпохой, достигнув значения около 0.72 к 6-й эпохе.
- Аналогичная положительная динамика наблюдается для валидационного набора, при этом точность чуть ниже, чем на обучении (около 0.68 на 6-й эпохе).
- Разница между обучающей и валидационной точностью также минимальна, что подтверждает хорошую обобщающую способность модели.

Матрица ошибок (рис. 26) для наиболее популярных - 20 классов показывает точность предсказаний для большинства классов.

Рис. 26 LSTM – Матрица ошибок

На диагонали (верные предсказания) заметно преобладание правильной классификации.

- Для классов, таких как Class 57, Class 103, Class 102, наблюдаются наибольшие количества правильных предсказаний (от 12 до 14 предсказаний на класс).
- Ошибки вне диагонали минимальны, что свидетельствует о способности модели корректно отделять представленные классы от других.
- Случаи путаницы между классами практически отсутствуют, что особенно важно для редких и малопредставленных классов.

Как итог: Модель LSTM демонстрирует высокое качество работы на последовательных данных, что подтверждается стабильной динамикой функции потерь и точности классификации.

- Разница между ошибками обучения и валидации остаётся минимальной, что говорит о сбалансированности модели.
- Матрица ошибок подтверждает, что LSTM эффективно справляется с задачей классификации, показывая высокую точность предсказаний и минимальное количество ошибок для топ-20 классов.

VII. Полученные результаты

A. *Общие выводы*

В рамках исследования был проведен всесторонний сравнительный анализ методов кластеризации и классификации по следующим критериям:

- Классификационные методы показали более высокую точность в разметке данных по сравнению с методами кластеризации. Это проявилось в способности моделей точно предсказывать принадлежность объекта к конкретной категории благодаря обучению на заранее размеченных примерах. Особенно ярко преимущества проявились в случаях, когда классы имели четкие и заметные различия, а обучающая выборка была подготовлена с должной тщательностью. Кластеризационные алгоритмы, напротив, показывали склонность к ошибкам за счет своей неспособности работать с априорной информацией о классах и зачастую объединяли объекты, схожие по второстепенным признакам, хотя их реальная принадлежность могла отличаться.
- Алгоритмы классификации, такие как XGBoost и Logistic Regression, продемонстрировали более высокую скорость при обработке больших объемов данных. Это связано с тем, что после этапа обучения классификационная модель способна быстро и эффективно относить новые объекты к соответствующим категориям, практически мгновенно обрабатывая массивы входных данных. В то время как методы кластеризации часто требуют значительных вычислительных ресурсов не только на этапе настройки, но и при классификации новых данных, а также могут нуждаться в повторном пересчете при малейших изменениях в исходной выборке.
- Классификационные методы предоставили более интерпретируемые и устойчивые результаты по сравнению с методами кластеризации. Благодаря наличию четко определённых меток классов и возможности анализа важности признаков, была получена возможность не только оценивать точность модели, но и объяснять причины принятия тех или иных решений. Это особенно важно для последующего контроля и корректировки алгоритмов, а также для отчётности перед заинтересованными сторонами.
- Кластеризационные методы чаще всего формируют группы автоматически, без учета имеющихся знаний о предметной области, что затрудняет интерпретацию и валидацию полученных результатов.

B. *Выводы по классификации*

На графике (рис. 27) представлено распределение вероятностей классов, полученных с использованием различных методов классификации, и их сравнение с истинным распределением данных. Каждый метод обозначен точками, закодированными определенными цветома, а истинное распределение показано штриховой линией.

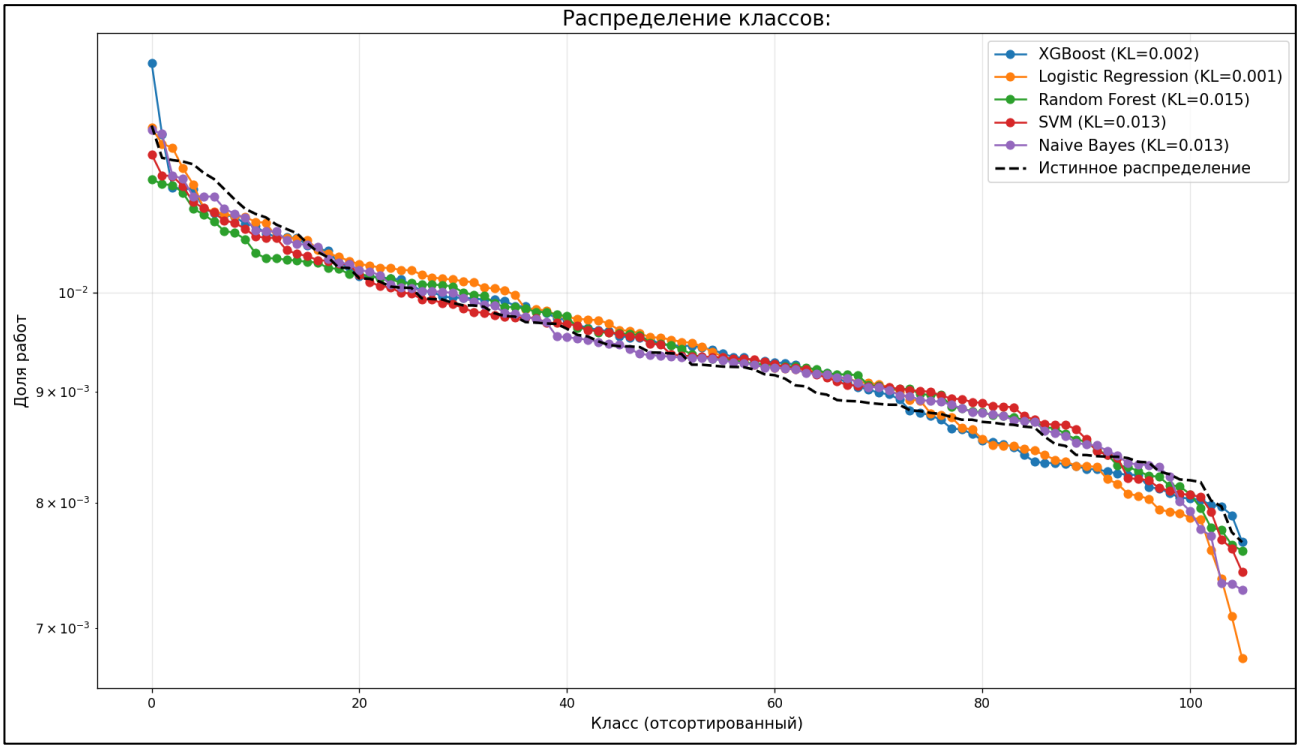

Рис. 27 Распределение классов

Ось абсцисс:

– Отсортированные классы (0 самый популярный, 106 наименее популярный, 0-106, так как классов в обучающей выборке 106), позволяющие визуализировать уменьшение доли объектов каждого класса.

Ось ординат:

– Доля объектов, принадлежащих каждому классу, выраженная в виде вероятности.

Легенда:

Методы классификации представлены в следующем виде:

– XGBoost: Синие точки (расстояние между распределениями).
– Logistic Regression: Оранжевые точки.
– Random Forest: Зелёные точки.
– SVM: Красные точки.
– Naive Bayes: Фиолетовые точки.
– Истинное распределение: Штриховая чёрная линия.

Под графиком указаны значения дивергенции Кульбака-Лейблера (KL), которые отражают степень отклонения распределения каждого метода от истинного.

– Наилучшее совпадение с истинным распределением демонстрируют Logistic Regression (KL=0.001) и XGBoost (KL=0.002). Остальные методы имеют большие значения KL-дивергенции и значительно отклоняются от истинного распределения.
– На классах с низкой долей объектов (правый край графика) все методы классификации имеют некоторую степень отклонения от истинного распределения, особенно методы Random Forest, SVM и Naive Bayes.

График позволяет оценить качество воспроизведения распределения классов различными методами классификации. Метрики KL-дивергенции указывают, что Logistic Regression и XGBoost более точно воспроизводят истинное распределение объектов, что делает их предпочтительными для задач, требующих строгого соответствия статистике данных.

Следующие графики (рис. 28) представлены в двух аспектах: усреднённые матрицы ошибок для двух крупных групп классов (1-50 и 81-106) и распределения топ-10 предсказанных классов для каждой модели.

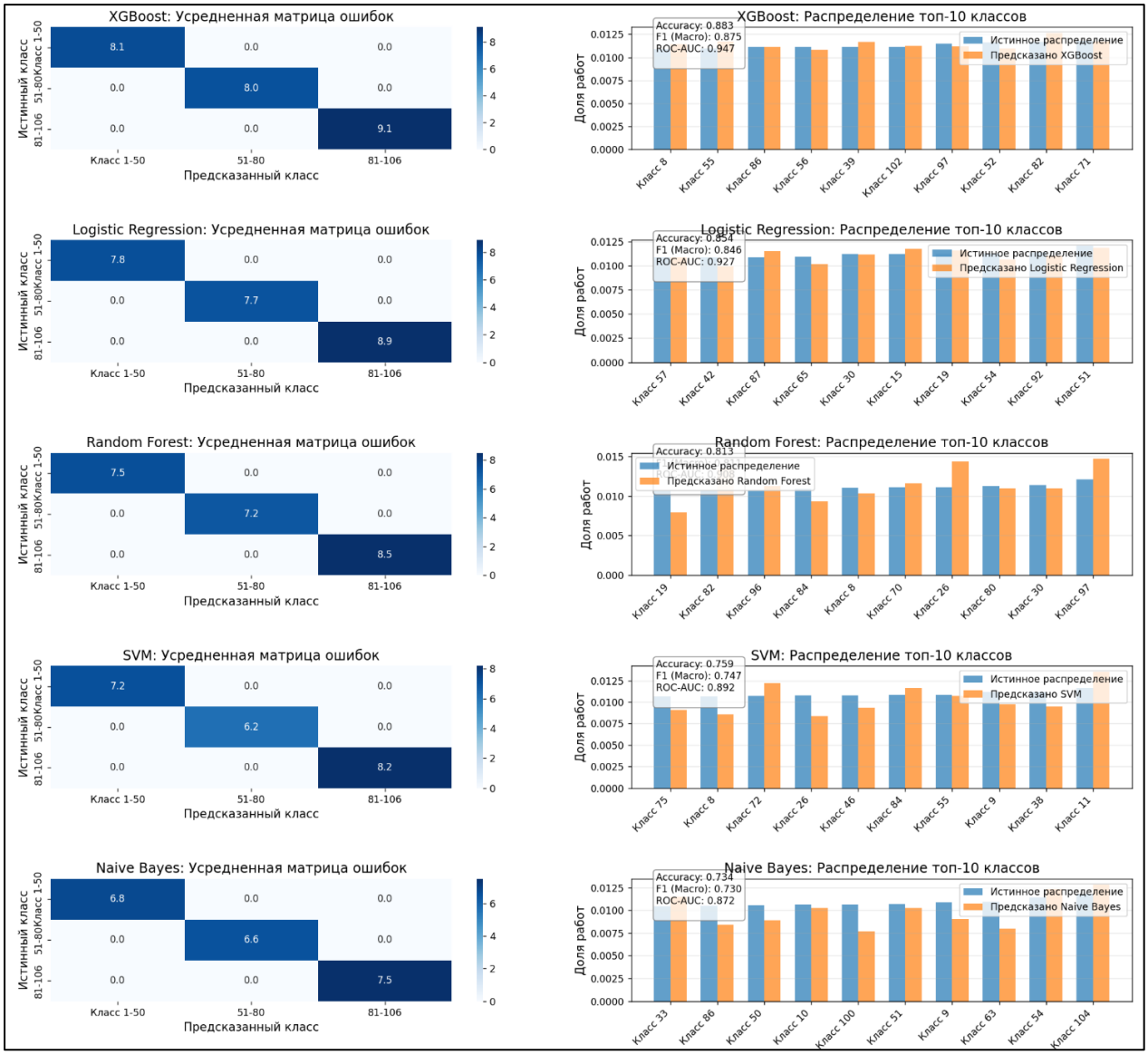

Рис. 28 Сводные метрики

- XGBoost показывает наивысшую точность (Accuracy = 0.883) и наибольшее значение F1-среднего (Macro-F1 = 0.857). Модель демонстрирует равномерное распределение ошибок между крупными группами классов.
- Logistic Regression также имеет высокие показатели (Accuracy = 0.863, Macro-F1 = 0.846). Распределение предсказанных классов близко к истинному, несмотря на более выраженные ошибки на группах редких классов.
- Random Forest и SVM уступают по точности и F1, особенно из-за смещений в предсказаниях и завышенного веса популярных классов.
- Naive Bayes имеет низшую точность (Accuracy = 0.730) и демонстрирует значительное смещение, что объясняется упрощённой природой алгоритма.

Для всех моделей наблюдается отклонение распределения от истинного, особенно для редких классов. Однако XGBoost и Logistic Regression лучше сохраняют истинное распределение на топ-10 наиболее представленных классов.

Модели, такие как Random Forest и SVM, демонстрируют снижение точности распределения, что связано с уклоном в сторону нескольких доминирующих классов.

- Naive Bayes не справляется с адаптацией под структуру данных, что особенно заметно в распределении на топ-10 классах, где различия между предсказанными и истинными значениями велики.
- Все модели имеют наибольшую плотность точных предсказаний на классе 81-106, что связано с их физическим преобладанием. При этом классы 1-50 чаще ошибочно классифицируются.
- XGBoost и Logistic Regression показывают наиболее сбалансированные матрицы ошибок, Naive Bayes, Random Forest и SVM имеют высокую степень несбалансированности.
- Модели Naive Bayes, Random Forest и SVM можно рассматривать только в задачах, где высокие значения баланса между классами не являются приоритетными.

Для задач, где требуется точное моделирование распределения классов и высокая обобщающая способность алгоритма, наиболее подходящими являются XGBoost и Logistic Regression.

Использование методов машинного обучения для автоматической разметки наименований строительных работ показало высокий потенциал в стандартизации и оптимизации данных.

Формирование единого классификатора на 106 позиций и успешная разметка 32 000 наименований работ значительно упростили дальнейший анализ и прогнозирование сроков выполнения строительных проектов (Рис. 29)

	id объект	id работ	Наименование работы	Наименование классификатора	Предсказание XGBOOST	Предсказание Linear Regression	Предсказание SVM	Random Forest
26	2	109	Устройство фундамента ленточного+ПЛИТА	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента
32	2	108	Устройство фундамента под колонны	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента
175	5	272	Устойство фундаментов	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента
279	9	454	1,Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента
307	10	493	Устройство фундаментов	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента
308	10	494	Гидроизоляция фундаментов и стен цоколя	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента
318	10	496	Устройство бетонных полов 1-го этажа	Устройство фундамента	Устройство полов	Устройство полов	Устройство полов	Устройство полов
333	10	495	Обратная засыпка	Устройство фундамента	Крыльца и пандусы	Планировка площадей	Планировка площадей	Планировка площадей
482	16	692	Устройство фундаментов под колонны, устройсте	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента	Устройство фундамента

Рис 29 – Пример работы алгоритмов классификации

C. Выводы по кластеризации

В условиях отсутствия заранее размеченной обучающей выборки приоритетное значение приобретают методы кластеризации, такие как K-Means и DBSCAN. Эти алгоритмы позволили выделить группы схожих наименований на основании скрытых признаков и расстояний между объектами, что особенно ценно для предварительной структуризации большого массива данных. Кластеризация помогла выявить неоднородности и аномалии, а также сгруппировать редкие и типовые категории работ для последующего анализа.

D. Итоговый анализ распределения вероятностей после автоматической разметки наименований работ

На рис. 30 представлены гистограммы распределения данных для различных типов фактической длительности выполнения работ, полученные после классификации методом XGBOOST.

Для каждого полученного классификатора работ выполнена повторная проверка на соответствие нормальному распределению, а также указаны медиана и среднее значение.

Применение методов классификации заметно улучшило показатели соответствия нормальному распределению.

Соответствие нормальному распределению теперь позволяет выполнять формирование прогнозных кривых длительностей работ при разных параметрах (например, одна и та же работа на объектах разных площади будет иметь разную длительность, для корректного отображения такой информации все работы требовалось классифицировать, чтобы в каждой категории (типе/классификаторе наименовании работы) было достаточно данных для построения прогнозной кривой.

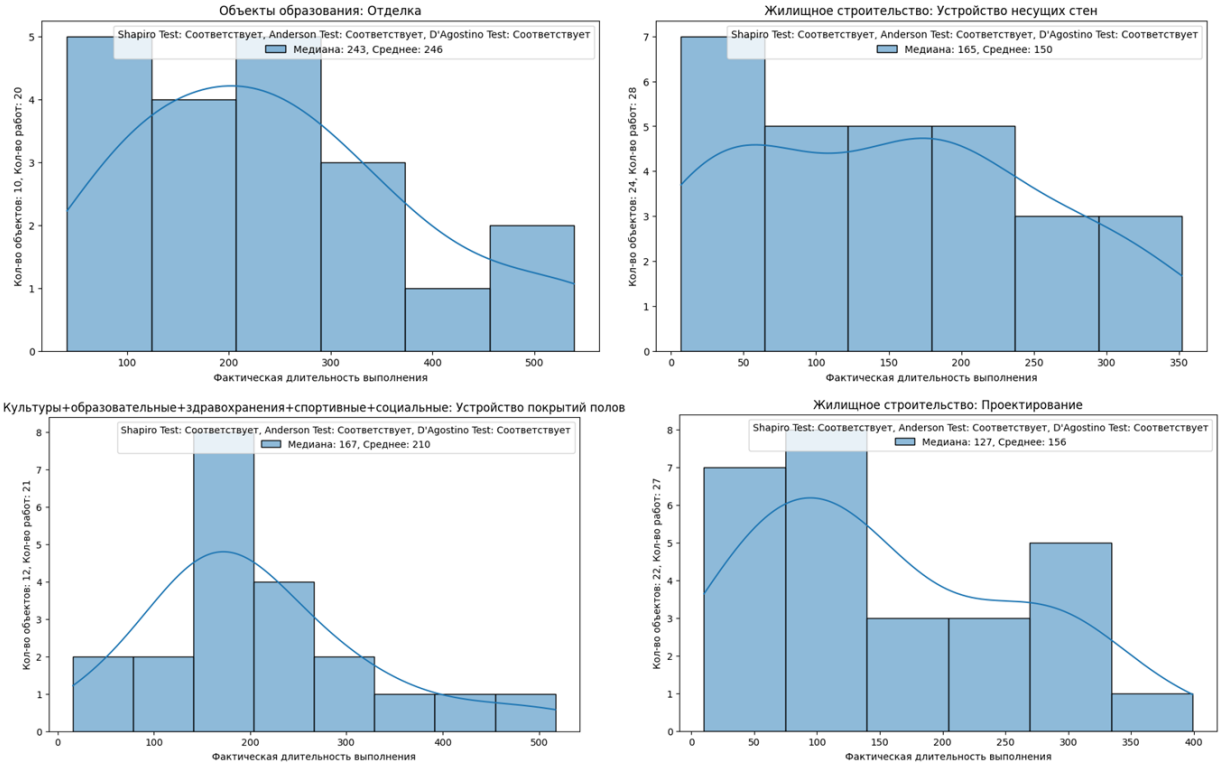

Рис. 30 График квантиль-квантиль (Q-Q)

Отделка:

- Распределение данных демонстрирует наличие ярко выраженной пиковой области с плавным спадом в сторону больших значений.
- Тесты (Shapiro, Anderson, D'Agostino) подтверждают соответствие распределения нормальному.
- Медиана: 243; среднее: 246.
- Применение классификационных методов позволило улучшить нормализацию данных и определить ключевые тренды.

Устройство несущих стен:

- Распределение слегка асимметрично, однако плавная кривая на графике показывает тенденцию к нормальному распределению.
- Тесты нормальности подтверждают соответствие.
- Медиана: 165; среднее: 156.
- Можно заметить возможность использования подобных данных для дальнейшего прогнозирования длительности выполнения подобных работ.

Устройство покрытия пола:

- Распределение данных имеет более широкий пик в сравнении с другими задачами, однако тесты подтверждают нормальность.
- Медиана: 107; среднее: 210.
- Классификационные методы позволяют устранить выбросы и значительно улучшить анализ данных.

Проектирование:

- На графике виден спад данных ближе к концу диапазона, но основной массив значений находится в центре.
- Тесты нормальности соответствуют нормальному распределению.
- Медиана: 127; среднее: 156.
- Применение моделей классификации обеспечивает предсказуемость распределения, что открывает возможности для построения прогнозных кривых.

Схожая картина наблюдалась и по остальным наименованиям работ.

Использование классификационных методов сыграло важную роль в улучшении нормального распределения данных для всех представленных задач. Полученные результаты позволяют не только выявить тренды в данных, но и возможность для построения прогнозных моделей, отражающих длительность выполнения аналогичных задач с высокой точностью.

Построение прогнозных кривых будет детально рассмотрено в дальнейшей части исследования, чему будет посвящена отдельная статья.

VIII. Заключение

Настоящее исследование было выполнено в рамках решения комплексной проблемы превышения плановых сроков объектов строительства – задержкам сроков выполнения работ. Было установлено, что ключевым барьером для использования потенциала исторических данных в целях улучшения планирования и прогнозирования является крайняя разрозненность и денормализованность исходных данных, особенно выраженная в значительной вариативности наименований строительных работ.

Доказана высокая эффективность методов машинного обучения - комплексное применение и сравнительный анализ широкого спектра алгоритмов кластеризации (K-Means, DBSCAN) и классификации (Логистическая регрессия, SVM, Наивный Байес, Random Forest, XGBoost, LSTM) подтвердили их способность значительно снижать разрозненность данных. Эти методы обеспечили автоматическое приведение разнородных описаний работ к единому стандартизированному формату, решая задачу автоматической разметки.

Проведенный анализ выявил сильные и слабые стороны, а также показатели точности и эффективности различных подходов машинного обучения в контексте специфической задачи унификации строительных данных. Результаты предоставляют практические рекомендации для выбора оптимальных методов.

Успешная унификация позволила сформировать целостный, структурированный массив исторических данных о наименованиях работ и их длительностях.

Нормализация данных посредством инструментов машинного обучения являлась критически важным и эффективным первым этапом на пути к повышению точности планирования в строительстве.

Текущие наработки будут использованы в формировании методики рекомендаций, описанной ранее в концептуальной части исследования [3,4], позволяющей осуществлять прогнозирование срывов сроков на основе исторических данных с целью предоставления руководителям информации о прогнозных сроках выполнения работ на этапе планирования. Построению прогнозных кривых и формированию методики прогнозирования длительностей строительных работ будут посвящены будущие статьи.

IX. Перспективы развития

Основным направлением для дальнейших исследований является разработка и внедрение системы рекомендаций для оптимизации планирования строительных проектов. Нормализованные и структурированные данные, полученные в результате данного исследования, создают необходимую основу для такого развития. Эта система будет использовать прогнозные модели, построенные на анализе исторических данных, для:

- Построения "прогнозных кривых" вероятности задержек для различных типов работ и этапов проекта на основе схожих исторических кейсов.
- Предоставления руководителям проектов данных для корректировки графиков, распределения ресурсов и упреждающего управления рисками на этапе планирования и исполнения.

В рамках данного исследования был сделан осознанный выбор в пользу классических методов машинного обучения (K-Means, DBSCAN, SVM, XGBoost и др.) для задачи унификации кратких, нормативно-регламентированных наименований

строительных работ. Этот выбор был продиктован спецификой данных (короткие, шаблонные фразы типа "<Действие> + <Объект> + <Материал/Технологию>", строгость терминологии ГОСТ/СНиП) и соображениями практической эффективности, интерпретируемости и ресурсозатрат. Классические методы доказали свою высокую эффективность высокую эффективность в этих условиях.

Однако, перспективным направлений для будущих изысканий является исследование возможностей современных подходов, таких как Large Language Models (LLM) и Named Entity Recognition (NER), в контексте строительной отрасли, особенно при работе с более сложными или менее структурированными данными.

В частности:

LLM (Большие языковые модели): Могут быть использованы для задач:

- Обработки более сложных описаний работ, включающих нюансы, условия или нестандартные формулировки, выходящие за рамки строгих шаблонов.
- Анализа сопутствующей текстовой информации (разного формата) для выявления дополнительных факторов, влияющих на сроки.
- Улучшения понимания контекста и семантических связей между работами, если данные станут более разнообразными.
- Применение LLM потребует решения проблем, отмеченных в исследовании: риска генеративных ошибок при работе с нормативной лексикой, необходимости больших объемов узкоспециализированных данных для дообучения и значительных вычислительных ресурсов. Их использование должно быть тщательно обосновано преимуществами перед классическими методами для конкретных подзадач.

NER (Распознавание именованных сущностей): Может найти применение в:

- Автоматическом структурировании более детализированных описаний, где требуется выделение конкретных объектов, материалов, единиц измерения или технологий внутри текстового описания работы для более глубокого анализа.
- Обработке технических спецификаций или проектной документации в текстовом формате для автоматического извлечения ключевых параметров работ.

В любом случае, текущая работа закладывает важный фундамент в виде унифицированных данных, открывая путь к созданию интеллектуальной системы поддержки принятия решений. Дальнейшие исследования будут фокусироваться на построении прогнозных моделей и рекомендательной системы на этой основе, а также на изучении потенциала более сложных методов NLP, в том числе LLM и NER, для решения специфических задач, где их применение может дать существенную добавленную стоимость при условии преодоления связанных с ними вызовов. Оптимальный путь развития предполагает комбинирование проверенных классических подходов с инновационными методами, где это обосновано требованиями задачи и характеристиками данных.

Библиография

[1] Распоряжение Правительства РФ от 31 октября 2022 г. № 3268-р

[2] Jim Burns DELAYS IN THE CONSTRUCTION INDUSTRY: OUR 2022 SURVEY RESULTS AND HOW THEY COMPARE TO 2016 [Электронный ресурс https://www.cornerstoneprojects.co.uk/blog/delays-in-the construction-industry-our-2022-survey-results-and-how-they compare-to-2016/ , дата обращения 25.06.2025)

[3] Коньков, В. В. Прогнозирование срывов сроков строительства с использованием машинного обучения на основе исторических данных о фактической продолжительности завершенных проектов / В. В. Коньков, В. И. Широков, М. Г. Жабицкий // International Journal of Open Information Technologies. – 2024. – Т. 12, № 8. – С. 35-47. – EDN OEJMJN.

[4] Коньков, В. В. Анализ возможности прогнозирования на исторических данных срывов сроков строительства с использованием методов машинного обучения / В. В. Коньков, В. И. Широков, М. Г. Жабицкий // Современные проблемы физики и технологий : Сборник тезисов докладов XI международной молодежной научной школы-конференции, Москва, 23–25 апреля 2024 года. – Москва: Национальный исследовательский ядерный университет "МИФИ", 2024. – С. 218-220. – EDN YTBOZX.

[5] Мурзагалеев, Т. М. Разработка модели прогнозирования ресурсов для выполнения задач по управлению проектами при проектировании и строительстве объектов / Т. М. Мурзагалеев, Н. А. Жукова // Международная конференция по мягким вычислениям и измерениям. – 2025. – Т. 1. – С. 364-367. – EDN ICRZOV.

[6] Зеленцов, Л. Б. Прогнозирование временных и стоимостных параметров при управлении инвестиционностроительными проектами / Л. Б. Зеленцов, М. С. Шогенов, Д. В. Пирко // Строительное производство. – 2020. – № 3. – С. 41-45. – DOI 10.54950/26585340_2020_3_41. – EDN VNEOKK.

[7] Zelentsov, L. Methodology of making organizational and technological decisions at the stage of operational management of construction operations based on the forecasting system / L. Zelentsov, L. Mailyan, D. Pirko // Journal of Physics: Conference Series. – 2021. – Vol. 2131, No. 2. – P. 022114. – DOI 10.1088/1742-6596/2131/2/022114. – EDN ZMVHIG.

[8] Alsugair A. M. et al. Artificial neural network model to predict final construction contract duration //Applied Sciences. – 2023. – Т. 13. – №. 14. – С. 8078

[9] Liben, S. M. Comparing advanced and traditional machine learning algorithms for construction duration prediction: a case study of Addis Ababa's public sector / S. M. Liben, D. A. Belachew, W. A. Elsaigh // Engineering Research Express. – 2024. – Vol. 6, No. 4. – P. 045119. – DOI 10.1088/2631-8695/ad979f. – EDN QBQUKP.

[10] Маштаков, Н. С. Предиктивная аналитика: из исторических данных к стратегическому прогнозированию / Н. С. Маштаков, П. С. Часов // Оригинальные исследования. – 2023. – Т. 13, № 12. – С. 61-66. – EDN LKTDSY.

[11] Moon S. et al. Automated construction specification review with named entity recognition using natural language processing //Journal of Construction Engineering and Management. – 2021. – Т. 147. – №. 1. – С. 04020147.

[12] Yaseen, Z.M.; Ali, Z.H.; Salih, S.Q.; Al-Ansari, N. Prediction of Risk Delay in Construction Projects Using a Hybrid Artificial Intelligence Model. Sustainability 2020, 12, 1514. https://doi.org/10.3390/su12041514

[13] Кузина, О. Н. Разработка моделей машинного обучения для прогнозирования продолжительности строительства / О. Н. Кузина // Актуальные проблемы строительной отрасли и образования - 2023 : Сборник докладов IV Национальной научной конференции, Москва, 15 декабря 2023 года. – Москва: Московский государственный строительный университет (национальный исследовательский университет), 2024. – С. 816-820. – EDN UWJKXK.

[14] Петроченко М. В. и др. Классификация строительной информации в BIM с использованием алгоритмов искусственного интеллекта //Вестник МГСУ. – 2022. – Т. 17. – №. 11. – С. 1537-1550.
[15] Мешкова Е. А., Дидковская О. В., Мешков А. А. ЦЕНООБРАЗОВАНИЕ В СТРОИТЕЛЬСТВЕ НА ОСНОВЕ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ НЕЙРОСЕТЕВОГО МОДЕЛИРОВАНИЯ //Вестник Международного института рынка. – 2022. – №. 1. – С. 147-152.
[16] Болотин С. А., Дадар А. Х., Мальсагов А. Р. Прогнозирование продолжительности строительства на основе исполнительных календарных графиков, статистического и нейросетевого моделирования //Недвижимость: экономика, управление. – 2017. – №. 3. – С. 59-63.
[17] Чижик А. В. Сравнение моделей векторизации текстов для задачи анализа тональности коротких сообщений из социальных сетей //Компьютерная лингвистика и вычислительные онтологии. – 2024. – Т. 1. – №. 7. – С. 81-89.
[18] Козинец А. Н. Применение методов машинного обучения для прогнозирования текучести кадров на основе открытых данных. – 2025.
[19] Жилов Р. А. Интеллектуальные методы кластеризации данных //Известия Кабардино-Балкарского научного центра РАН. – 2023. – №. 6 (116). – С. 152-159.
[20] Молчанова Т. А. Дообучение больших языковых моделей для решения специализированных задач: магистерская диссертация : дис. – 2024.
[21] Лагутина Н. С., Васильев А. М., Зафиевский Д. Д. Задачи в области распознавания именованных сущностей: технологии и инструменты //Моделирование и анализ информационных систем. – 2023. – Т. 30. – №. 1. – С. 64-85.
[22] Akhmetov I. et al. An Open-Source Lemmatizer for Russian Language based on Tree Regression Models //Res. Comput. Sci. – 2020. – Т. 149. – №. 3. – С. 147-153.
[23] Баганов А. П. и др. Классификация текстов с высоким уровнем искажений для задач категоризации товаров: магистерская диссертация по направлению подготовки: 01.04. 02-Прикладная математика и информатика. – 2021.

Статья получена 8 августа 2025.
Коньков Владислав Владимирович, аспирант ИИКС НИЯУ МИФИ, vlad.konkov.7145@gmail.com)
Широков Валерий Игоревич, магистр ВИШ НИЯУ МИФИ, valerkashir@gmail.com
Сычев Вячеслав Александрович, старший преподаватель ВИШ НИЯУ МИФИ, VASychov@mephi.ru

Application of machine learning methods to classify construction works titles for data normalization and predictive modeling

V.V. Konkov, V.I. Shirokov, V.A. Sychev

***Abstract*—Exceeding the planned deadlines for the implementation of construction projects remains one of the most pressing problems reducing the efficiency of the Russian construction industry.**

The key obstacle to using historical data and machine learning methods is the extreme fragmentation and denormalization of the source data, in particular, significant variability and non-standardization of the names of construction works. The authors propose a methodology for the automatic unification of variability in the names of construction works using machine learning. This unification is considered as a critical step in building systems for forecasting deadlines within the framework of the state strategy to reduce construction cycles by 30% by 2030.

A comprehensive methodology is proposed: data preprocessing, text vectorization and automatic unification. The use of two approaches is studied: clustering (K-Means, DBSCAN) to identify groups of similar works and classification (including Logistic regression, SVM, Random Forest, XGBoost, LSTM) to assign descriptions to unified classes. A comparative analysis of the efficiency of algorithms by the metrics accuracy/recall/F1 was conducted.

High efficiency of machine learning methods (especially XGBoost, Random Forest, LSTM) for unification was proven, which allows to form a structured array of historical data. A comprehensive solution for generating recommendations for optimizing calendar plans based on the analysis of similar historical objects using ML was proposed.

A structured array of historical data was formed, including information on the names of works and their durations, which will then be used to implement solutions that allow to obtain reasonable forecasts of deadlines, adjust the planned duration of construction work and minimize the risks of failure, increasing the efficiency of project management.

***Keywords:* construction delays, machine learning, clustering, classification, historical data analysis**

REFERENCES

[1] Order of the Government of the Russian Federation of October 31, 2022 No. 3268-r

[2] Jim Burns DELAYS IN THE CONSTRUCTION INDUSTRY: OUR 2022 SURVEY RESULTS AND HOW THEY COMPARE TO 2016 [URL - https://www.cornerstoneprojects.co.uk/blog/delays-in-the-construction-industry-our-2022-survey-results-and-how-they compare-to-2016/ , accessed on June 25, 2025)

[3] Konkov, V. V. Forecasting construction delays using machine learning based on historical data on the actual duration of completed projects / V. V. Konkov, V. I. Shirokov, M. G. Zhabitsky // International Journal of Open Information Technologies. – 2024. – Vol. 12, No. 8. – Pp. 35-47. – EDN OEJMJN.

[4] Konkov, VV Analysis of the possibility of forecasting construction delays based on historical data using machine learning methods / VV Konkov, VI Shirokov, MG Zhabitsky // Modern problems of physics and technology: Collection of abstracts of reports of the XI international youth scientific school-conference, Moscow, April 23-25, 2024. – Moscow: National Research Nuclear University MEPhI, 2024. – Pp. 218-220. – EDN YTBOZX.

[5] Murzagaleev, TM Development of a resource forecasting model for performing project management tasks in the design and construction of facilities / TM Murzagaleev, NA Zhukova // International conference on soft computing and measurements. – 2025. – Vol. 1. – Pp. 364-367. – EDN ICRZOV.

[6] Zelentsov, L. B. Forecasting of time and cost parameters in the management of investment and construction projects / L. B. Zelentsov, M. S. Shogenov, D. V. Pirko // Construction production. – 2020. – No. 3. – Pp. 41-45. – DOI 10.54950/26585340_2020_3_41. – EDN VNEOKK.

[7] Zelentsov, L. Methodology of making organizational and technological decisions at the stage of operational management of construction operations based on the forecasting system / L. Zelentsov, L. Mailyan, D. Pirko // Journal of Physics: Conference Series. – 2021. – Vol. 2131, No. 2. – P. 022114. – DOI 10.1088/1742-6596/2131/2/022114. – EDN ZMVHIG.

[8] Alsugair A. M. et al. Artificial neural network model to predict final construction contract duration //Applied Sciences. – 2023. – T. 13. – No. 14. – P. 8078

[9] Liben, S. M. Comparing advanced and traditional machine learning algorithms for construction duration prediction: a case study of Addis Ababa's public sector / S. M. Liben, D. A. Belachew, W. A. Elsaigh // Engineering Research Express. – 2024. – Vol. 6, No. 4. – P. 045119. – DOI 10.1088/2631-8695/ad979f. – EDN QBQUKP.

[10] Mashtakov, N. S. Predictive analytics: from historical data to strategic forecasting / N. S. Mashtakov, P. S. Chasov // Original research. – 2023. – Vol. 13, No. 12. – Pp. 61-66. – EDN LKTDSY.

[11] Moon S. et al. Automated construction specification review with named entity recognition using natural language processing //Journal of Construction Engineering and Management. – 2021. – Vol. 147. – No. 1. – Pp. 04020147.

[12] Yaseen, Z.M.; Ali, Z.H.; Salih, S.Q.; Al-Ansari, N. Prediction of Risk Delay in Construction Projects Using a Hybrid Artificial Intelligence Model. Sustainability 2020, 12, 1514. https://doi.org/10.3390/su12041514

[13] Kuzina, O. N. Development of machine learning models for forecasting construction duration / O. N. Kuzina // Actual problems of the construction industry and education - 2023: Collection of reports of the IV National Scientific Conference, Moscow, December 15, 2023. - Moscow: Moscow State University of Civil Engineering (National Research University), 2024. - Pp. 816-820. - EDN UWJKXK.

[14] Petrochenko M. V. et al. Classification of construction information in BIM using artificial intelligence algorithms // Bulletin of MGSU. - 2022. - Vol. 17. - No.. 11. - Pp. 1537-1550.

[15] Meshkova E. A., Didkovskaya O. V., Meshkov A. A. PRICING IN CONSTRUCTION BASED ON TIME SERIES ANALYSIS USING NEURAL NETWORK MODELING // Bulletin of the International Market Institute. - 2022. - No. 1. - P. 147-152.

[16] Bolotin S. A., Dadar A. Kh., Malsagov A. R. Forecasting the duration of construction based on executive calendar schedules, statistical and

neural network modeling // Real Estate: Economics, Management. - 2017. - No. 3. - P. 59-63.
[17] Chizhik A. V. Comparison of text vectorization models for the task of analyzing the sentiment of short messages from social networks // Computer linguistics and computational ontologies. – 2024. – V. 1. – No. 7. – P. 81-89.
[18] 18. Kozinets A. N. Application of machine learning methods to predict employee turnover based on open data. – 2025.
[19] Zhilov R. A. Intelligent methods of data clustering // Bulletin of the Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences. – 2023. – No. 6 (116). – P. 152-159.
[20] 20. Molchanova T. A. Further training of large language models to solve specialized problems: master's thesis: diss. – 2024.
[21] Lagutina N. S., Vasiliev A. M., Zafievsky D. D. Problems in the field of recognition ofenumerable entities: technologies and tools // Modeling and analysis of information systems. - 2023. - Vol. 30. - No. 1. - P. 64-85.
[22] Akhmetov I. et al. An Open-Source Lemmatizer for Russian Language based on Tree Regression Models // Res. Comput. Sci. - 2020. - Vol. 149. - No. 3. - P. 147-153.
[23] Baganov A. P. et al. Classification of texts with a high level of distortion for product categorization problems: master's thesis in the direction of training: 01.04. 02-Applied Mathematics and Computer Science. - 2021.