Методика оценки влияния качества данных на результативность моделей машинного обучения для определения опозданий исполнения контрольных точек проектов

Е. В. Никульчев, Д. Ю. Ильин, С. Е. Духовенский, Н. Ш. Газанова, А. А. Червяков

Аннотация **— В работе рассматривается процесс построения оценок возможности опоздания в сроках выполнения контрольных точек национальных и федеральных проектов на основе технологий машинного обучения. В федеральных информационных системах собраны большие объемы данных, в том числе о ходе выполнения национальных проектов, что позволяет их использовать для машинного обучения различных моделей. Одной из главных задач является контроль хода выполнения проектов и отслеживание хода проектов по заданным контрольным точкам. Несмотря на организационно-технические меры, наблюдаются опоздания в выполнении контрольных точек. Классическое применение подходов машинного обучения для решения задачи классификации на данных о контрольных точках, не позволяет получить результат пригодный для практического применения. Это объясняется наличием в исходных данных неоднозначностей. Как правило, методы машинного обучения демонстрируются и совершенствуются на типовых наборах данных, а в реальных системах качеству данных необходимо уделить существенное внимание. Исследование посвящено разработке методики оценки влияния качества данных на результативность моделей машинного обучения в задаче оценки выполнения или опоздания исполнения контрольных точек национальных и федеральных проектов. Результаты получены на реальных, обезличенных, пронормированных, закодированных данных федеральной системы мониторинга. Выявлены неоднозначности в данных и построена зависимость качества машинного обучения в зависимости от объема неоднозначных данных. Полученные результаты демонстрируют эффективность разработанной методики.**

Ключевые слова **— качество данных, контрольные точки проектов, машинное обучение, национальные проекты.**

Статья получена 7 апреля 2025.

Е. В. Никульчев – МИРЭА – Российский технологический университет (e-mail: nikulchev@mail.ru), Москва, Россия
Д. Ю. Ильин – МИРЭА – Российский технологический университет (e-mail: i@dmitryilin.com), Москва, Россия
С. Е. Духовенский – МИРЭА – Российский технологический университет, (e-mail: dukhovenskiy.s.e@edu.mirea.ru), Москва, Россия
Н. Ш. Газанова – МИРЭА – Российский технологический университет, Москва, Россия
А. А. Червяков – Федеральное Казначейство Российской Федерации (e-mail: achervyakov@roskazna.ru), Москва, Россия

I. Введение

Машинное обучение становится важным инструментом моделирования в области управления проектами и широко применяются при автоматизации повторяющихся задач, оптимизации, прогнозировании, ситуационном моделировании, оценке рисков [1]. В контексте управления проектами [2] ключевым объектом является содержание и структура плана проекта, представляющего собой схему реализации. Он включает этапы, задачи, сроки и ресурсы [3]. Для контроля реализации этапов проекта в план проекта добавляются контрольные точки (КТ) как инструмент проведения объективной оценки конкретных этапов [4], что позволяет получать измеримые показатели. Исследования в области инструментов искусственного интеллекта могут быть направлены на обучение моделей для определения фактического статуса проекта и потенциальных задержек в четкой и объективной форме [4].

В настоящем исследовании рассматриваются национальные проекты (НП) и федеральные проекты. Национальный проект — комплекс мероприятий, реализуемых для достижения национальных приоритетов и их показателей, установленных Указом Президента Российской Федерации от 7 мая 2024 г. № 309 «О национальных целях развития Российской Федерации на период до 2030 года и на перспективу до 2036 года» и иных, указанных в поручениях и (или) указаниях Президента Российской Федерации, Правительства Российской Федерации, Председателя Правительства Российской Федерации, решениях Совета или президиума Совета.

В процессе реализации НП выявлен ряд ключевых рисков [6–9], в том числе пробелы в методиках расчета показателей НП и накопленных данных, что затрудняет прогнозирование достижения целей и оценку достаточности финансирования. В рамках государственной автоматизированной информационной системы «Управление» сформирована система мониторинга, осуществляющая мониторинг эффективности реализации НП с точки зрения мероприятий проекта, созданы специализированные проектные офисы, отвечающие за соответствующие

проекты. Информационная система отслеживает фактические значения параметров на контрольных точках проекта, отклонения фактических значений от плановых. В условиях развития цифровой экономики [10] использование традиционных инструментов анализа и прогнозирования для оценки эффективности проектов может быть дополнено средствами и методами искусственного интеллекта, использующих машинное обучение.

Решение задач прогнозирования своевременности завершения контрольных точек проектов рассматривается в различных областях: проекты по разработке программного обеспечения [11, 12], строительные проекты [13], образовательные проекты [14] и другие. В некоторых исследованиях рассматривается прогнозирование возможного факта задержки [11, 14], в то время как в других к этому подходят с позиции определения возможных сроков выполнения проектов [15]. Существование задержек в выполнении, может быть связано, например, с человеческим фактором [16] и носить устоявшийся характер. Интеллектуальные модели [11, 14, 15] часто используются для прогнозирования возможных задержек и сроков выполнения проектов.

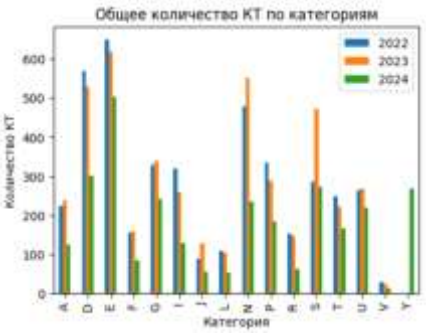
В работе рассматривалась задача разработки инструментов для прогнозирования успешности реализации проекта на основе исторических данных мониторинга контрольных точек национальных и федеральных проектов с использованием машинного обучения. В качестве исходных данных использовались нормализованные и анонимизированные данные из федеральных систем [17]. Имеющиеся в системе мониторинга данные о причинах опозданий, позволили сформировать признаковое пространство и оценить степень их влияния на результат, однако классическое применение подходов машинного обучения для решения задачи классификации на данных о контрольных точках, не позволяет получить результат пригодный для практического применения. Как правило, методы машинного обучения демонстрируются и совершенствуются на типовых наборах данных, а в реальных системах качеству данных необходимо уделить существенное внимание [18, 19].

Исследование посвящено разработке методики оценки влияния качества данных на результативность моделей машинного обучения в задаче оценки выполнения или опоздания исполнения контрольных точек национальных и федеральных проектов. Результаты получены на реальных, обезличенных, пронормированных, закодированных данных федеральной системы мониторинга. Выявлены неоднозначности в данных и построена зависимость качества машинного обучения от объема неоднозначных данных. Полученные результаты демонстрируют эффективность разработанной методики.

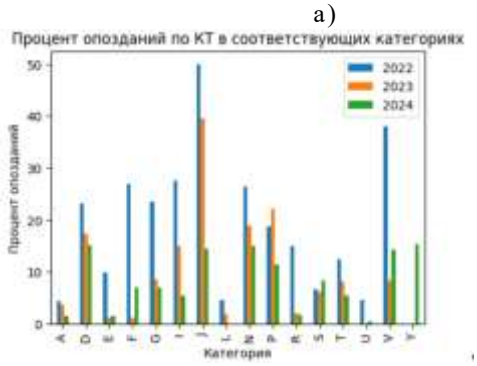
II. Описание данных

Рассматривается набор данных по контрольным точкам национальных и федеральных проектов. Объем набора составляет 11496 записей. В нем представлены 14 признаков из которых 11 номинальных и 3 бинарных. Целевая переменная отражает наличие или отсутствие опоздания выполнения контрольной точки, в связи с чем решаемая задача – бинарная классификация на основе машинного обучения. В данных присутствуют записи за 2022, 2023 и часть 2024 года.

На рис. 1 показано распределение контрольных точек по категориям в рамках признака «код национального проекта» и процент опозданий в соответствующих категориях. Меньшее количество записей за 2024 год обусловлено тем, что используются данные за неполный год. В целом наблюдается снижение доли опозданий по контрольным точкам, однако общая закономерность отсутствует. Так, например, процент опозданий для национального проекта F вырос в 2024 году.



а)



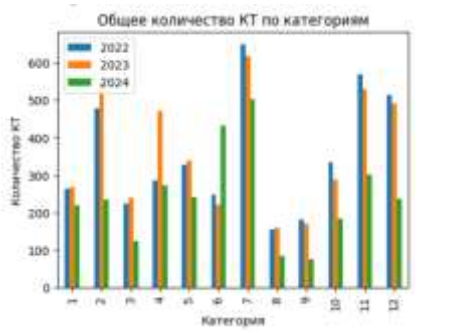
б)

Рис. 1. Распределение контрольных точек по категориям (а) и процент опозданий по контрольным точкам в соответствующих категориях (б) в рамках признака «код национального проекта»

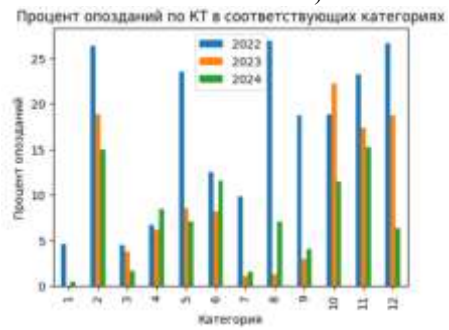
На рис. 2 рассматривается признак «федеральный орган исполнительной власти, ответственный за национальный проект». Тогда как для 1 и 3 категорий характерна небольшая доля опозданий, 2, 10 и 11 показывают высокий процент. При этом есть категории, доля опозданий в которых значительно изменялась, например 5, 7 и 9.

На рис. 3 показано распределение по признаку «тип результата контрольной точки». Тогда как общее количество контрольных точек между годами пропорционально и практически не меняется, доля опозданий изменяется. Крайние значения для 5 и 14 категории обусловлены малым количеством контрольных точек.

Как видно из представленных данных на рис. 1-3, распределение опозданий различается от года к году. В связи с этим было решено отложить данные за 2024 год как валидационную выборку, а данные за 2022 и 2023 год использовать как обучающую и тестовую выборки с разделением данных в соотношении 4 к 1, т.е. 20% данных отложено в качестве тестовой выборки.

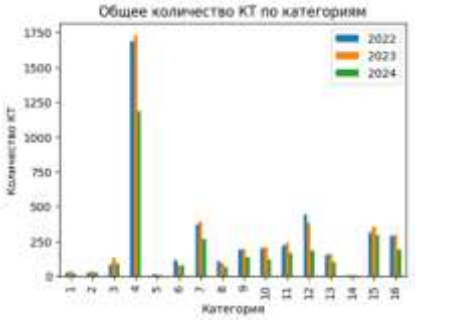


а)

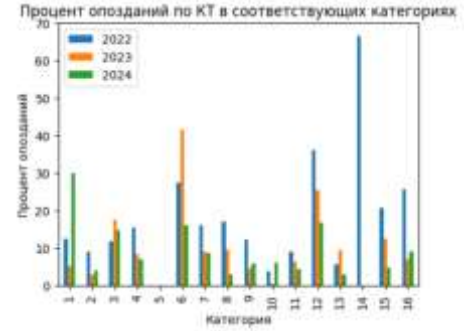


б)

Рис. 2. Распределение контрольных точек по категориям (а) и процент опозданий по контрольным точкам в соответствующих категориях (б) в рамках признака «федеральный орган исполнительной власти, ответственный за национальный проект»



а)



б)

Рис. 3. Распределение контрольных точек по категориям (а) и процент опозданий по контрольным точкам в соответствующих категориях (б) в рамках признака «тип результата контрольной точки»

Исходные попытки обучить модели на полных данных обучающей выборки приводили к невысоким оценкам на тестовой выборке и к более низким на валидационной. Так были рассмотрены классификаторы на основе дерева решений, опорных векторов, случайного леса, k-ближайших соседей, логистической регрессии, адаптивного бустинга и линейный классификатор с обучением методом стохастического градиентного спуска. Ансамбль классификаторов с использованием мягкого голосования, включающий все вышеперечисленные методы также не позволил получить приемлемых результатов. Кросс-валидация и оверсемплинг с помощью метода SMOTEN позволили улучшить результаты, но незначительно.

На рис. 4 показана матрица спутанности при обучении ансамбля моделей с оверсемплингом и проверке результата на валидационной выборке. При рассмотренных методах обучения показатель полноты равен 0.29, точности – 0.17 и F1 – 0.22. Оценка полноты показывает, что значительная часть опозданий не была выявлена.

Рис. 4. Матрица спутанности при обучении ансамбля моделей на данных 2022 и 2023 года, с проверкой на данных 2024 года

Анализ данных показал, что в датасете присутствуют записи, имеющие большое количество одинаковых значений признаков, при этом принадлежащих противоположным классам. Таким образом, актуальным является сокращение большего класса путем удаления противоречивых данных.

III. Методика оценки влияния качества данных на результативность моделей

Для выявления неоднозначных данных и подготовки набора данных для эффективного машинного обучения разработана соответствующая методика.

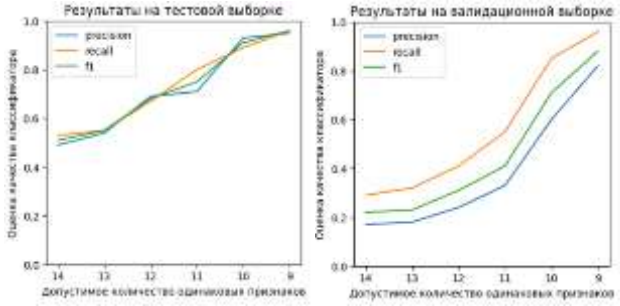
Пусть в наборе исходных данных представлено n категориальных и бинарных признаков. Часть записей, принадлежащих разным классам, может иметь одинаковые значения признаков, что приводит к низкому качеству обучения моделей. Так, запись из класса 1 и запись из класса 2 могут иметь s одинаковых признаков, при этом $0 \le s \le n$. Для устранения неопределенности в данных предлагается следующая процедура.

1. Задать пороговое значение t, отражающее допустимое количество одинаковых признаков в противоположных классах.
2. Разделить датасет на 2 части по классам.
3. Для каждой записи из класса 1 найти записи из класса 2, для которых $s > t$, т.е. количество одинаковых признаков больше порогового значения.
4. Удалить найденные записи из класса 2.
5. Объединить данные в единый датасет.

6. Провести обучение модели с использованием сокращенного датасета.

7. Применить обученную модель к проверочным данным из полной выборки.

На рис. 5 показана динамика изменения оценок классификации в зависимости от заданного порога. По мере снижения порога и, тем самым, сокращения объема противоречивых записей в датасете, качество обучения растет. В табл. 1 дополнительно приведены количественные характеристики датасета по мере его сокращения.



а) б)

Рис. 5. Графики изменения оценок качества классификации по мере удаления схожих по признакам записей с различным классом из большего класса: (а) на тестовой выборке, (б) на валидационной выборке

Таблица 1. Результаты сокращения датасета с помощью предложенной процедуры

Допустимое количество одинаковых признаков	Оценки для тестовой выборки			Оценки для валидационной выборки			Записей в датасете	Записей класса 1 (КТ с опозданием)	Записей класса 2 (КТ без опоздания)
	Точность	Полнота	F1	Точность	Полнота	F1			
14	0.49	0.53	0.51	0.17	0.29	0.22	11496	1425	10071
13	0.54	0.55	0.55	0.18	0.32	0.23	11260	1425	9835
12	0.69	0.67	0.68	0.24	0.41	0.31	10175	1425	8750
11	0.71	0.80	0.75	0.33	0.55	0.41	7851	1425	6426
10	0.93	0.89	0.91	0.60	0.85	0.71	4783	1425	3358
9	0.95	0.96	0.96	0.82	0.96	0.88	2577	1425	1152

При заданном пороге $t = 10$ допустимых одинаковых признаков получена матрица спутанности, показанная на рис. 6. Она показывает хорошую полноту, равную 0.85, и приемлемую точность классификации 0.6.

В результате применения обученной модели к исходным данным валидационной выборки получены оценки, показанные на рис. 7. Полнота сохранилась на уровне 0.85, но, ввиду более высокой чувствительности обученной модели, точность составила 0.14. При этом оценка F1 для модели равна 0.24, что выше, чем у модели, обученной на полных данных.

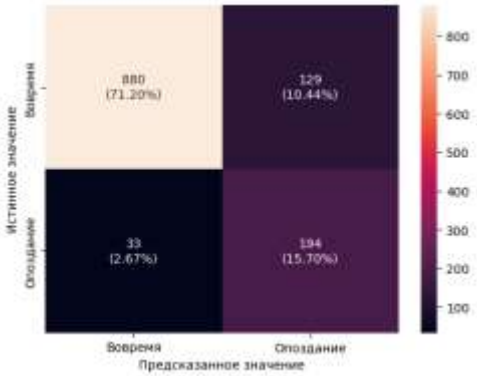


Рис. 6. Матрица спутанности при обучении и проверке ансамбля моделей на сокращенных валидационных данных

Рис. 7. Матрица спутанности при обучении ансамбля моделей на сокращенных данных (по большему классу) и проверке на полных валидационных данных

Сокращение меньшего класса с помощью предложенной процедуры также дает положительные результаты, однако ограничено из-за меньшего количества записей. В соответствии с рис. 8, точность составляет 0.28, но при меньших оценках полноты (0.06) и F1 (0.1).

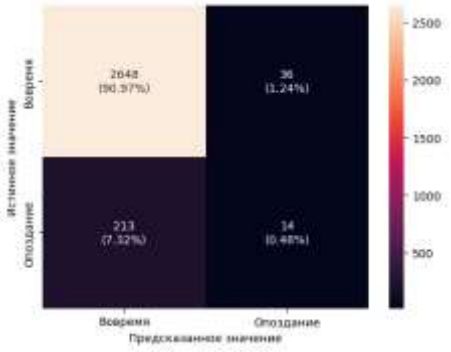


Рис. 8. Матрица спутанности при обучении ансамбля моделей на сокращенных данных (по меньшему классу) и проверке на полных валидационных данных

Таким образом, предложенная методика позволяет выявить зависимость качества обучения модели прогнозирования опозданий заявленный точек от качества данных. Исходные данные могут быть разделены на две категории – основную, которая прогнозируется на основе интеллектуальных моделей, и, вторую, требующую специализированного подхода к оценке вероятности исполнения КТ.

IV. Заключение

В рамках проведенного исследования разработана методика оценки влияния качества данных на результативность моделей машинного обучения в задаче оценки выполнения или опоздания исполнения контрольных точек национальных и федеральных проектов. Результаты исследования позволили выявить набор данных, для которого интеллектуальные модели эффективно решают задачу классификации контрольных точек по заданному факторному пространству.

С помощью предложенной методики удалось устранить неоднозначность в данных, что позволило повысить оценки качества обучения моделей. В частности, для валидационных данных оценка полноты была улучшена с 0.29 до 0.85.

Предложенную процедуру можно отнести к вариации андерсемплинга. Она применима для сокращения большего класса задач с целью устранения неоднозначности в данных. Полученные результаты демонстрируют эффективность разработанной методики.

Финансирование

Работа выполнена в рамках Государственного задания на 2024 год паспорта № 5560-24 по научно-методическому и ресурсному обеспечению системы образования на тему: «Научно-методическое обеспечение работ по анализу деятельности управления общественными финансами Российской Федерации с применением искусственного интеллекта»

Библиография

[1] Barcaui A., Monat A., Who is better in project planning? Generative artificial intelligence or project managers? // Project Leadership and Society. 2023 V. 4. P. 100101. https://doi.org/10.1016/j.plas.2023.100101

[2] Grzeszczyk T.A. Artificial intelligence and project management: an integrated approach to knowledge-based evaluation. — Taylor & Francis, 2024. https://doi.org/10.4324/9781003341611

[3] Titov S., Nikulchev E., Brikoshina I., Sueti A. Client communications and quality satisfaction in project-based company // Quality - Access to Success. 2020. N. 21(174). P. 68–71.

[4] Красильников В.М., Ильинский А.А. Роль аналитики при управлении проектами по диверсификации предприятия // Проблемы экономики и управления нефтегазовым комплексом. 2024. № 8 (236). С. 34-41.

[5] Niederman F. Project management: openings for disruption from AI and advanced analytics // Information Technology & People. 2021. V. 34, N. 6. P. 1570–1599 https://doi.org/10.1108/ITP-09-2020-0639

[6] Ильченко С.В. Национальные проекты России и риски их реализации // Бизнес и дизайн ревю. 2021. № 2(22). С. 1.

[7] Перевалова О.С., Буньковский В.И.: Национальные проекты россии для предприятий: экономические выгоды, возможности, снижение рисков. Первый экономический журнал. 2(344). 64–71 (2024) https://doi.org/10.58551/20728115_2024_2_64

[8] Бадалова А.Г., Кириченко П.С., Олейник А.В. Совершенствование системы управления национальными проектами на стадии их разработки // Экономика, предпринимательство и право. 2024. № 14(7). С. 3341–3358. https://doi.org/10.18334/epp.14.7.121315

[9] Строев В.В., Кузнецов Н.В.: Мониторинг национальных проектов в Российской Федерации и риски, связанные с их реализацией // Вестник университета. 2023. № 11. С. 14–20 https://doi.org/10.26425/1816-4277-2023-11-14-20

[10] Maron M.A. The choice of control points of projects taking into account possible change of structure of works // Business Informatics. 2016. N. 2(36). P. 57–62 https://doi.org/10.17323/1998-0663.2016.2.57.62

[11] Choetkiertikul M., Dam H.K., Tran T., Ghose A. Predicting the delay of issues with due dates in software projects // Empirical Software Engineering. 2017. V. 22. P. 1223–1263. https://doi.org/10.1007/s10664-016-9496-7

[12] Martens A., Vanhoucke M. An empirical validation of the performance of project control tolerance limits // Automation in Construction. 2018. V. 89. P. 71–85 https://doi.org/10.1016/j.autcon.2018.01.002

[13] Sepasgozar S.M., Karimi R., Shirowzhan S., Mojtahedi M., Ebrahimzadeh S., McCarthy D. Delay causes and emerging digital tools: A novel model of delay analysis, including integrated project delivery and PMBOK // Buildings. 2019. V. 9, N. 9. P. 191 https://doi.org/10.3390/buildings9090191

[14] Sheoraj Y., Sungkur R.K. Using AI to develop a framework to prevent employees from missing project deadlines in software projects-case study of a global human capital management (HCM) software company // Advances in engineering software. 2022. V. 170. 103143. https://doi.org/10.1016/j.advengsoft.2022.103143

[15] Lishner I., Shtub A. Using an artificial neural network for improving the prediction of project duration // Mathematics. 2022 V. 10, N. 22. P. 4189. https://doi.org/10.3390/math10224189

[16] Плохов Д.В., Никульчев Е.В., Титов С.А., Осипов И.В. Методика оценки влияния социальных коммуникаций на результативность инновационного проекта // Cloud of science. 2016. Т. 3, № 3. С. 444-492

[17] Албычев А.С. ИТ-приоритеты Казначейства России // Журнал Бюджет. 2023. № 6(246). С. 42–47.

[18] Духовенский С.Е. Разработка репозитория метаданных на основе объектно-ориентированной логической модели для оценки качества данных // International Journal of Open Information Technologies. 2024. Т. 12, № 4. С. 60-67.

[19] Духовенский С.Е., Пушкин, П.Ю., Никульчев Е.В. Методика оценки качества данных реестра операторов персональных данных // International Journal of Open Information Technologies. 2024. Т. 12, № 1. С. 129-136.

Methodology for assessing the impact of data quality on the effectiveness of machine learning models in the task of estimation of the implementation of project checkpoints

E. Nikulchev, D. Ilin, S. Dukhovenskiy, N. Gazanova, A. Chervyakov

***Abstract* - The paper examines the process of estimation of the implementation of checkpoints of national and federal projects based on machine learning technologies. Federal information systems contain large volumes of data, including data on the progress of national projects, which allows them to be used for machine learning using various models. One of the main tasks is to control the progress of projects and track it at specified checkpoints. Despite organizational and technical measures, there are occasional delays in completing checkpoints. The data on the reasons for delays available in the system made it possible to form a feature space and assess the degree of their influence on the result. However, the classical application of machine learning approaches to solve the classification problem does not allow obtaining a result suitable for practical application. This is explained by the presence of ambiguities in the original data. Typically, machine learning methods are demonstrated and improved on typical data sets, but in real systems, the quality of the data used for machine learning must be given significant consideration. The study is devoted to the development of a technique for assessing the impact of data quality on the effectiveness of machine learning models in the task of estimation of the implementation of national projects' checkpoints. The study was conducted on real, anonymized, standardized, coded data from the federal monitoring system. The obtained results demonstrate the effectiveness of the developed technique.**

Keywords: **data quality, project checkpoints, machine learning, national projects**

REFERENCES

[1] A. Barcaui, A. Monat, "Who is better in project planning? Generative artificial intelligence or project managers?" Project Leadership and Society, vol. 4, p. 100101, 2023.

[2] T.A. Grzeszczyk, "Artificial Intelligence and Project Management: An Integrated Approach to Knowledge-based Evaluation," (Taylor & Francis, 2024)

[3] S. Titov, E. Nikulchev, I. Brikoshina, A. Suetin, "Client Communications and Quality Satisfaction in Project-based Company," Quality - Access to Success, no. 21(174), pp. 68–71, 2020.

[4] V.M. Krasilnikov, A.A. Ilyinsky, "The role of analytics in managing enterprise diversification projects," Problems of economics and management of the oil and gas complex, vol. 8, no. 236, pp. 34-41, 2024. [Rus]

[5] F. Niederman, "Project management: openings for disruption from AI and advanced analytics," Information Technology & People, vol. 34, no. 6, pp. 1570–1599, 2021.

[6] S.V. Ilchenko, "National projects of Russia and risks of their implementation," Business and design review, no. 2(22), p. 1, 2021 [Rus]

[7] O.S. Perevalova, V.I. Bunkovskiy, "National projects of Russia for enterprises: economic benefits, opportunities, risk reduction," First Economic Journal, no. 2(344), pp. 64–71, 2024. [Rus]

[8] A.G. Badalova, P.S. Kirichenko, A.V. Oleynik, "Improving the management system of national projects at the stage of their development," Economy, Entrepreneurship and Law, vol. 14, no. 7, pp. 3341–3358, 2024. [Rus]

[9] V.V. Stroev, N.V. Kuznetsov, "Monitoring of national projects in the Russian Federation and risks associated with their implementation," University Bulletin, no. 11, pp. 14–20, 2023. [Rus]

[10] M.A. Maron, "The choice of control points of projects taking into account possible change of structure of works," Business Informatics, vol. 2, no. 36, pp. 57–62, 2016.

[11] M. Choetkiertikul, H.K. Dam, T. Tran, A. Ghose, "Predicting the delay of issues with due dates in software projects," Empirical Software Engineering, vol. 22, pp. 1223–1263, 2017.

[12] A. Martens, M. Vanhoucke, "An empirical validation of the performance of project control tolerance limits," Automation in Construction, vol. 89, pp. 71–85, 2018.

[13] S.M Sepasgozar., R. Karimi, et al., "Delay causes and emerging digital tools: A novel model of delay analysis, including integrated project delivery and PMBOK," Buildings, vol. 9, no. 9, p. 191, 2019.

[14] Y. Sheoraj, R.K. Sungkur, "Using AI to develop a framework to prevent employees from missing project deadlines in software projects-case study of a global human capital management (HCM) software company," Advances in engineering software, vol. 170, p. 103143, 2022

[15] I. Lishner, A. Shtub. "Using an artificial neural network for improving the prediction of project duration," Mathematics, vol. 10, no. 22, p. 4189, 2022.

[16] D.V. Plokhov, E.V. Nikulchev, S.A. Titov, I.V. Osipov, "Methodology for assessing the impact of social communications on the effectiveness of an innovation project," Cloud of Science, vol. 3, no. 3, pp. 444-492, 2016. [Rus]

[17] A.S. Albychev, "IT priorities of the Treasury of Russia," Budget Magazine, no. 6(246), pp. 42–47, 2023. [Rus]

[18] S.E. Dukhovenskiy, "Implementation of metadata repository based on object-oriented model for data quality assessment," International Journal of Open Information Technologies, vol. 12, no. 4, pp. 60-67, 2024. [Rus]

[19] S.E. Dukhovenskiy, P. Pushkin, E. Nikulchev, "The data quality assessment technique of personal data operators register," International Journal of Open Information Technologies, vol. 12, no. 1, pp. 129-136,. 2024 [Rus]