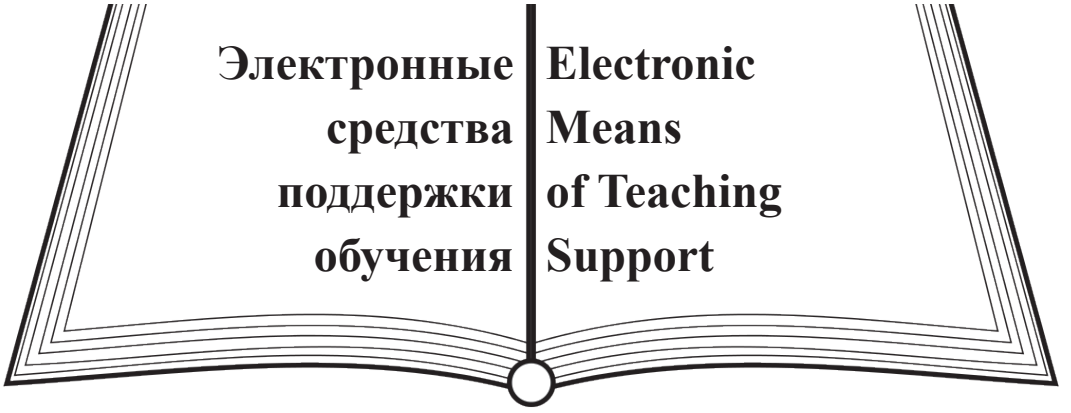

Научная статья
УДК 378.4:004
DOI: 10.24412/2072-9014-2025-171-80-89

ИНТЕГРАЦИЯ МУЛЬТИМОДАЛЬНЫХ ТЕХНОЛОГИЙ В ОБРАЗОВАТЕЛЬНЫЙ ПРОЦЕСС НА ПРИМЕРЕ РАЗРАБОТКИ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ НА ПЛАТФОРМЕ APACHE SPARK

Тимур Муртазович Босенко

Московский городской педагогический университет,
Москва, Россия
bosenkotm@mgpu.ru, https://orcid.org/0000-0002-5375-096X

Аннотация. Современные подходы к обучению в области анализа больших данных и машинного обучения требуют интеграции мультимодальных технологий, что позволяет обучающимся развивать навыки работы с разнородными источниками информации, такими как текстовые, числовые, мультимедийные и потоковые данные. Целью исследования является изучение методов машинного обучения, включая потоковую обработку данных с использованием Apache Spark, для решения реальной бизнес-задачи в рамках проектного обучения. Оценка эффективности различных методов командной работы (проектное обучение, кооперативное обучение, кейс-обучение) показала, что наилучшие результаты в точности модели (AUC = 0.91) достигаются при применении комплексного подхода с использованием всех мультимодальных технологий в рамках проектной работы. В статье подробно анализируются результаты, полученные с использованием таких технологий, как обработка текстовых данных (NLP), потоковых, числовых и мультимедийных данных. Использование этих технологий позволило значительно повысить точность предсказаний оттока клиентов, улучшить процесс принятия бизнес-решений

и развить у обучающихся практические навыки в области анализа больших данных и искусственного интеллекта.

Ключевые слова: мультимодальные технологии; машинное обучение; потоковые данные; Apache Spark; проектное обучение; прогнозирование оттока клиентов.

Original article
UDC 378.4:004
DOI: 10.24412/2072-9014-2025-171-80-89

INTEGRATION OF MULTIMODAL TECHNOLOGIES INTO THE EDUCATIONAL PROCESS ON THE EXAMPLE OF DEVELOPING MACHINE LEARNING MODEL ON THE APACHE SPARK PLATFORM

Timur M. Bosenko

Moscow City University,
Moscow, Russia
bosenkotm@mgpu.ru, https://orcid.org/0000-0002-5375-096X

Abstract. Modern approaches to teaching big data analytics and machine learning require the integration of multimodal technologies, which allows students to develop skills in working with heterogeneous information sources, such as text, numeric, multimedia, and streaming data. The objective of the study is to investigate machine learning methods, including streaming data processing using Apache Spark, to solve a real-world business problem within the framework of project-based learning. Evaluation of the effectiveness of various teamwork methods (project-based learning, cooperative learning, case learning) showed that the best results in model accuracy (AUC = 0.91) are achieved when applying an integrated approach using all multimodal technologies within the framework of project work. The article analyzes in detail the results obtained using technologies such as text data processing (NLP), streaming data, numeric and multimedia data. The use of these technologies has significantly increased the accuracy of customer churn predictions, improved business decision-making, and developed students' practical skills in big data analytics and artificial intelligence.

Keywords: multimodal technologies; machine learning; streaming data; Apache Spark; project-based learning; customer churn prediction.

Для цитирования: Босенко Т. М. Интеграция мультимодальных технологий в образовательный процесс на примере разработки модели машинного обучения на платформе Apache Spark / Т. М. Босенко // Вестник МГПУ. Серия «Информатика и информатизация образования». 2025. № 1 (71). С. 80–89. DOI: 10.24412/2072-9014-2025-171-80-89

For citation: Bosenko T. M. Integration of multimodal technologies into the educational process on the example of developing machine learning model on the Apache Spark platform / T. M. Bosenko // MCU Journal of Informatics and Informatization of Education. 2025. № 1 (71). P. 80–89. DOI: 10.24412/2072-9014-2025-171-80-89

Введение

Современные тенденции в области анализа данных диктуют необходимость обработки больших объемов информации, поступающей в режиме реального времени из различных источников [1]. Платформа Apache Spark и ее библиотека машинного обучения MLlib предоставляют эффективные инструменты для решения этой задачи [2; 3]. В частности, Spark Structured Streaming позволяет работать с потоковыми данными, а MLlib содержит алгоритмы для построения моделей машинного обучения на распределенных данных [4].

Другим важным аспектом является работа с мультимодальными данными, то есть с данными различной природы — текстовыми, числовыми, мультимедийными [5]. Интеграция таких разнородных данных в единый массив данных позволяет получать более точные и релевантные результаты предсказаний и рекомендаций.

С точки зрения методологии и технологии профессионального образования внедрение в учебный процесс бизнес-задач и кейс-методов приобретает особую актуальность, поскольку это способствует формированию у обучающихся практических компетенций в работе с современными инструментами анализа данных [6; 7]. Такой подход, ориентированный на интеграцию теоретических знаний и практических навыков, реализуется в данном исследовании на примере задачи прогнозирования оттока клиентов телекоммуникационной компании с использованием потоковых данных.

Цель — изучение и анализ построения моделей машинного обучения с использованием методологии организации проектной деятельности обучающихся, ориентированной на применение мультимодальных методов обучения на основе индустриального бизнес-кейса, основой для которого являются потоковые данные. При этом первоочередное внимание уделяется интеграции различных источников данных и аналитических подходов, что способствует формированию у студентов навыков комплексного решения такого рода бизнес-задач. Это позволяет интегрировать различные виды данных и подходы к их анализу, развивая компетенции в области разработки интеллектуальных систем и применения современных методов анализа данных.

Методы исследования

Исследование проведено на базе Института цифрового образования Московского педагогического университета. Группа обучающихся разбивалась на несколько команд, каждая из которых выполняла определенную задачу. Методологической основой исследования выступают компетентностный, практико-ориентированный и проектный подходы в профессиональном образовании [8]. В соответствии с этими подходами обучение технологиям анализа больших данных строится вокруг решения реальной бизнес-задачи, которая имеет прикладное значение. В качестве эмпирической базы исследования

использованы данные телекоммуникационной компании за период 2023–2024 гг. Выбор задачи прогнозирования клиентского оттока обоснован ее комплексным характером, что позволяет применить разносторонний подход к анализу исследуемых данных. Проектная деятельность организована по методу Scrum, с разбиением на спринты и распределением ролей в команде [9].

Методы командной работы были следующими:

– Проектное обучение — обучающиеся выполняют задание в командах. Такой вид обучения включает в себя решение комплексной задачи с реальными данными и бизнес-ценностью, что позволяет обучающимся получить практический опыт в решении актуальных проблем с использованием технологий больших данных и искусственного интеллекта.

– Кооперативное обучение — акцент на совместной работе обучающихся в группе, где каждая команда или пара отвечает за выполнение части общей задачи. Обучающиеся работали в группе, выполняя взаимосвязанные подзадачи для достижения общей цели — подготовки данных для дальнейшего анализа и построения модели машинного обучения.

– Кейс-обучения — обучающиеся, разделенные на команды, решают разные аспекты одного бизнес-кейса или различные кейсы, связанные с одной темой. На основе анализа результатов работы модели в контексте реального бизнес-кейса (прогнозирование оттока клиентов), участники предоставляли выводы о ее эффективности.

Преподаватель выступал в роли консультанта, тогда как обучающиеся самостоятельно планировали работу, выбирали инструменты, подходы к решению и распределяли между собой задачи. Для технической реализации проекта использовалась платформа Apache Spark как наиболее эффективная и популярная среда для распределенной обработки больших данных [10]. Потоковые данные о действиях клиентов поступают из симулятора событий через брокера сообщений Apache Kafka и обрабатывались с помощью Spark Structured Streaming. Пакетные данные о клиентах из корпоративного хранилища в виде СУБД Postgre SQL также загружались в Spark.

Предварительная подготовка данных включала в себя типичные шаги для таких задач: кодирование категориальных признаков, векторизацию числовых признаков, масштабирование и нормализацию значений. Эти операции выполняются с помощью методов библиотеки Spark MLlib.

В качестве модели машинного обучения выбрана логистическая регрессия как одна из наиболее эффективных алгоритмов бинарной классификации [11]. Модель обучалась на исторических данных о клиентах, которые уже покинули или остались в компании. Полученные весовые коэффициенты затем применялись в режиме реального времени к входящим потоковым данным для предсказания вероятности оттока. Для оценки качества модели вычислялись метрики accuracy, precision, recall, F1-мера, а также строились ROC-кривые и рассчитывалась метрика оценки качества AUC [12]. С целью повышения обобщающей способности модели применялся алгоритм кросс-валидации с настройкой гиперпараметров.

Источником больших данных для задачи прогноза оттока клиентов выступал кластер МГПУ — «Аренадата» [13]; реализация доступа к данным, их обработка и визуализация выполнялась с помощью Python-библиотек: PySpark, NumPy, Pandas, Matplotlib и Seaborn. Обучающиеся выполняли задания в облачном сервисе Colab Google в Jupyter Notebook для обеспечения интерактивности и воспроизводимости результатов.

Результаты исследования

Исследование выявило дифференцированное влияние различных компонентов мультимодальных технологий на качество прогнозирования (табл. 1).

Таблица 1

Сравнительный анализ мультимодальных технологий

Технология	Виды мультимодальности	Значимость при использовании
Обработка текстовых данных (NLP)	Извлечение информации из текстовых данных (например, отзывы клиентов, чаты, комментарии)	Повышает точность модели за счет учета мнений и настроений клиентов
Обработка числовых данных	Использование числовых данных, таких как количество обращений, продолжительность использования	Ключевая информация для прогнозирования оттока, высокая корреляция с поведением клиентов
Обработка мультимедийных данных (изображения, графики)	Анализ изображений профилей клиентов, графиков активности и визуализаций данных	Дополняет модель визуальными признаками, которые могут указывать на поведение клиентов
Обработка потоковых данных (Streaming data)	Анализ данных в реальном времени (например, транзакции, клики, время на сайте) с помощью Spark Structured Streaming	Обеспечивает оперативность анализа и адаптацию модели к изменениям в поведении клиентов
Интеграция различных типов данных (мультимодальный анализ)	Комбинирование текстовых, числовых и мультимедийных данных для более точного прогнозирования	Улучшает точность модели за счет комплексного подхода к анализу данных

Установлено, что интеграция разнородных источников данных существенно повышает прогностическую способность модели машинного обучения. В частности,

– обработка текстовых данных методами NLP продемонстрировала значительную эффективность в анализе неструктурированной информации, содержащейся в отзывах и коммуникациях клиентов;

– анализ числовых показателей выявил высокую корреляцию между количественными метриками пользовательской активности и вероятностью оттока;

– интеграция мультимедийных данных обеспечила дополнительный информационный слой, который повышает точность прогнозирования;

– обработка потоковых данных в режиме реального времени способствовала повышению адаптивности модели к динамическим изменениям в поведении клиентов.

Исследование позволило установить статистически значимые различия в эффективности применения методов командной работы при применении мультимодальных технологий.

В основные метрики оценки были включены следующие параметры:

– точность прогностической модели (AUC);

– совокупная результативность команды в условиях реализации бизнес-задачи.

В ходе проведенного анализа (табл. 2) выявлены такие особенности:

– проектное обучение с полной интеграцией мультимодальных технологий продемонстрировало максимальную эффективность (AUC = 0.91, прирост +14 %);

– кооперативное обучение показало умеренное повышение точности (AUC = 0.82, прирост +8 %), что свидетельствует о необходимости оптимизации методологии интеграции мультимодальных данных;

– кейс-обучение продемонстрировало существенное улучшение прогностической способности модели (AUC = 0.88, прирост +12 %), подтверждая эффективность применения мультимодальных технологий в контексте решения реальных бизнес-задач.

Таблица 2

Сравнительный анализ результатов применения мультимодальных технологий в задаче прогноза оттока клиентов

Метод командной работы	Применение мультимодальных технологий	Точность модели (AUC)	Эффективность решения бизнес-задачи
Проектное обучение	Применение всех технологий (NLP, потоковые данные, мультимедийные, числовые)	+14 % AUC: 0.91	Значительное улучшение точности прогнозирования оттока
Кооперативное обучение	Применение мультимодальных данных для совместного анализа текстов, чисел и изображений	+8 % AUC: 0.82	Хорошая интеграция, но меньшая точность в сравнении с другими подходами
Кейс-обучение	Использование текстов, числовых данных, потоковых и мультимедийных данных для решения реального кейса	+12 % AUC: 0.88	Высокая релевантность выводов, улучшенные прогнозы и рекомендации для бизнеса

В рамках исследования была проведена сравнительная оценка эффективности различных методов командной работы при решении задачи прогнозирования оттока клиентов телекоммуникационной компании. Ключевым аспектом исследования стало сравнение результативности моделей с применением мультимодальных технологий (МТ+) и без их использования (МТ–). Результаты представлены на рисунке в виде ROC-кривых в логарифмических координатах, что позволяет детально визуализировать различия в производительности моделей.

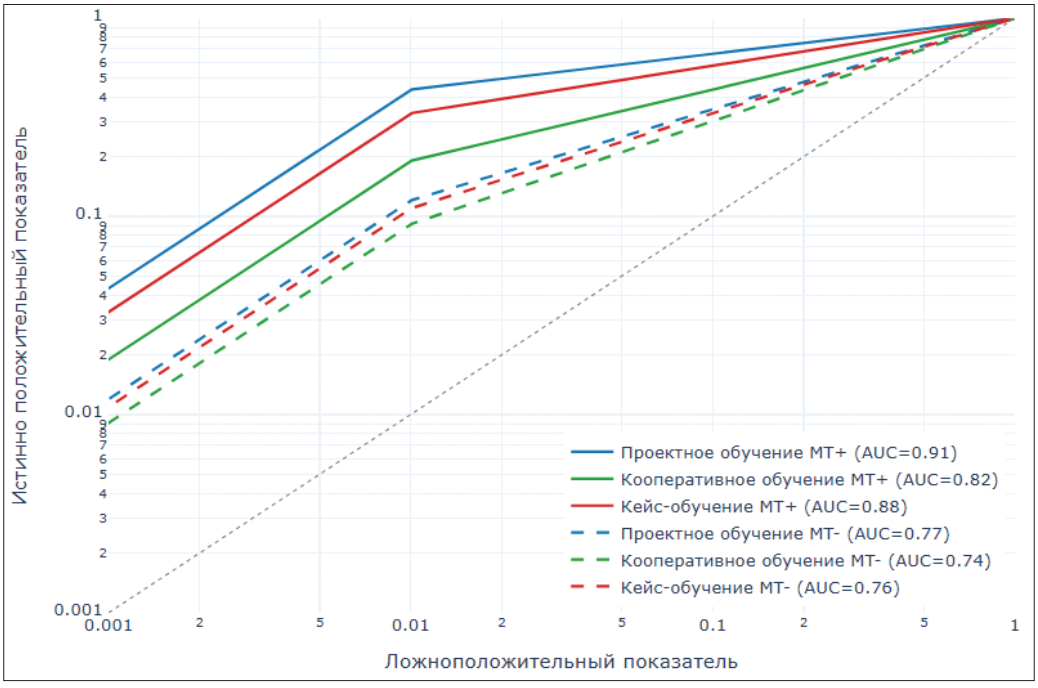

Рис. ROC-кривые для разных подходов к прогнозированию оттока клиентов

Проектное обучение. Метод проектного обучения продемонстрировал лучшую эффективность среди всех исследуемых подходов. При использовании мультимодальных технологий (МТ+) достигнут показатель AUC = 0.91, что на 14 % превышает результат базовой модели (МТ–) с AUC = 0.77. Графический анализ ROC-кривых показывает значительное превосходство мультимодального подхода, особенно в области небольших значений *ложноположительного показателя,* что критически важно для практического применения в бизнес-задачах.

Кооперативное обучение. Применение кооперативного обучения привело к наименьшему, но статистически значимому приросту эффективности AUC. Использование мультимодальных технологий позволило достичь AUC = 0.82, что на 8 % превышает показатели базового подхода (AUC = 0.74). ROC-кривые демонстрируют более плавный характер улучшения производительности модели, что может свидетельствовать о более равномерном распределении ошибок классификации.

Кейс-обучение. Метод кейс-обучения показал промежуточные результаты с существенным улучшением с учетом мультимодальных технологий. Достигнутый показатель AUC = 0.88 при использовании MT+ на 12 % превышает результаты базового подхода (AUC = 0.76). Анализ ROC-кривых указывает на заметное улучшение в области средних значений *ложноположительного показателя,* что свидетельствует о более сбалансированной работе модели.

Заключение

Результаты исследования подтверждают высокую эффективность применения мультимодальных технологий в образовательном процессе, особенно при решении задач, связанных с большими данными и искусственным интеллектом. Графический анализ ROC-кривых и количественные показатели AUC подтверждают значительное улучшение качества прогнозирования при использовании комплексного подхода к обработке данных. Метод проектного обучения с применением всех доступных мультимодальных технологий показал наилучшие результаты, что указывает на перспективность данного подхода для решения сложных бизнес-задач. Каждый из методов командной работы — проектное обучение, кооперативное обучение и кейс-обучение — вносит свой вклад в развитие компетенций обучающихся. Использование мультимодальных данных позволяет не только освоить современные технологии, но и развить критическое мышление, необходимые для принятия обоснованных решений в реальных условиях.

Список источников

1. *Sahoo C.* Sentiment analysis using deep learning techniques: a comprehensive review / C. Sahoo, M. Wankhade, B. K. Singh // International Journal of Multimedia Information Retrieval. 2023. Vol. 12. No. 2. P. 41. DOI: 10.1007/s13735-023-00308-2

2. *Симонов В. С.* Метод преобразования императивного кода для платформ параллельной обработки данных / В. С. Симонов, М. С. Хайретдинов // Проблемы информатики. 2023. № 3 (60). С. 68–80. DOI: 10.24412/2073-0667-2023-3-68-80

3. *Босенко Т. М.* Инструменты для сбора, хранения и передачи больших данных в инфокоммуникационных системах: учебно-методическое пособие / Т. М. Босенко. М.: Эдитус, 2023. 68 с.

4. *Benymol J.* Enhanced query performance for stored streaming data through structured streaming within spark SQL / J. Benymol, R. Nithin, J. Lumy // Indonesian Journal of Electrical Engineering and Computer Science. 2024. Vol. 35. No. 3. P. 1744–1750. DOI: 10.11591/ijeecs.v35.i3.pp1744-1750

5. *Садыкова Р. Х.* Сторителлинг как нейролингводидактический прием в обучении (практико-ориентированный подход) / Р. Х. Садыкова, М. Т. Кордзадзе // Педагогический журнал. 2023. Т. 13. № 12-1. С. 176–190. DOI 10.34670/AR.2024.20.99.020.

6. *Шунина Л. А.* Использование облачных технологий при организации производственной практики студентов педагогического вуза / Л. А. Шунина // Вестник МГПУ. Серия «Информатика и информатизация образования». 2022. № 2 (60). С. 18–29. DOI 10.25688/2072-9014.2022.60.2.02.

7. *Босенко Т. М.* Использование OLAP-технологии в процессе обучения студентов специальности 38.03.05 — бизнес-информатика / Т. М. Босенко, П. К. Григорьев // Актуальные проблемы теории и практики обучения физико-математическим и техническим дисциплинам в современном образовательном пространстве: сборник избранных статей VI Всероссийской (с международным участием) научно-практической конференции (Курск, 15–16 декабря 2022 г.). Курск: Курский государственный университет, 2022. С. 146–150.

8. *Горбачева Е. А.* Эффективность развития информационной компетентности при применении проектного подхода в изучении гуманитарных дисциплин в высшей школе / Е. А. Горбачева // Вестник МГПУ. Серия «Информатика и информатизация образования». 2024. № 3 (69). С. 91–100. DOI: 10.25688/2072-9014.2024.69.3.8

9. *Тропникова В. В.* Использование метода eduscrum: проблемы и пути решения / В. В. Тропникова // Среднее профессиональное образование. 2021. № 12 (316). С. 3–5.

10. *Jordan M. I.* Machine learning: Trends, perspectives, and prospects / M. I. Jordan, T. M. Mitchell // Science. 2015. Vol. 349. No. 6245. P. 255–260. DOI: 10.1126/science.aaa8415

11. *Hossin M.* A Review on Evaluation Metrics for Data Classification Evaluations / M. Hossin, M. A. Sulaiman // International Journal of Data Mining & Knowledge Management Process. 2015. Vol. 5. No. 2. P. 1–11. DOI: 10.5121/ijdkp.2015.5201

12. *Мальчевская П. А.* Прогнозирование оттока клиентов на основе модели машинного обучения / П. А. Мальчевская // Лига исследователей МГПУ: сборник материалов студенческой открытой конференции: в 2 ч. (Москва, 20–24 ноября 2023 г.). М.: МГПУ, 2024. С. 116–119.

13. *Фролов Ю. В.* Платформы данных AI/ML на основе отечественного программного обеспечения / Ю. В. Фролов, Т. М. Босенко, Д. В. Яценко // Современная {цифровая} дидактика. М.: А-Приор, 2023. С. 119–128.

References

1. *Sahoo C.* Sentiment analysis using deep learning techniques: a comprehensive review / C. Sahoo, M. Wankhade, B. K. Singh // International Journal of Multimedia Information Retrieval. 2023. Vol. 12. No. 2. P. 41. DOI: 10.1007/s13735-023-00308-2

2. *Simonov V. S.* Method of imperative code transformation for parallel data processing platforms / V. S. Simonov, M. S. Khayretdinov // Problems of computer Science. 2023. No. 3 (60). P. 68–80. DOI: 10.24412/2073-0667-2023-3-68-80

3. *Bosenko T. M.* Tools for collecting, storing and transmitting big data in infocommunication systems: a textbook / T. M. Bosenko. Moscow: Editus, 2023. 68 p.

4. *Benymol J.* Enhanced query performance for stored streaming data through structured streaming within spark SQL / J. Benymol, R. Nithin, J. Lumy // Indonesian Journal of Electrical Engineering and Computer Science. 2024. Vol. 35. No. 3. P. 1744–1750. DOI: 10.11591/ijeecs.v35.i3.pp1744-1750

5. *Sadykova R. H.* Storytelling as a neuro-linguistic didactic method in teaching (practice-oriented approach) / R. H. Sadykova, M. T. Kordzadze // Pedagogical Journal. 2023. Vol. 13. No. 12-1. P. 176–190. DOI: 10.34670/AR.2024.20.99.020

6. *Shunina L. A.* The use of cloud technologies in the organization of industrial practice of students of a pedagogical university / L. A. Shunina // MCU of Journal of Informatics and Informatization of Education. 2022. No. 2 (60). P. 18–29. DOI: 10.25688/2072-9014.2022.60.2.02

7. *Bosenko T. M.* The use of OLAP technology in the process of teaching students of specialty 03/38/05 — business informatics / T. M. Bosenko, P. K. Grigoriev // Actual problems of theory and practice of teaching physico-mathematical and technical disciplines in the modern educational space: Collection of selected articles of the VI All-Russian (with international participation) Scientific and Practical Conference (Kursk, December 15–16, 2022). Kursk: Kursk State University, 2022. P. 146–150.

8. *Gorbacheva E. A.* The effectiveness of the development of information competence in the application of the project approach in the study of humanities in higher education / E. A. Gorbacheva // MCU of Journal of Informatics and Informatization of Education. 2024. No. 3 (69). P. 91–100. DOI: 10.25688/2072-9014.2024.69.3.8

9. *Tropnikova V. V.* Using the eduscrum method: problems and solutions / V. V. Tropnikova // Secondary vocational education. 2021. No. 12 (316). P. 3–5.

10. *Jordan M. I.* Machine learning: Trends, perspectives, and prospects / M. I. Jordan, T. M. Mitchell / Science. 2015. Vol. 349. No. 6245. P. 255–260. DOI: 10.1126/science.aaa8415

11. *Hossin M.* A Review on Evaluation Metrics for Data Classification Evaluations / M. Hossin, M. A. Sulaiman // International Journal of Data Mining & Knowledge Management Process. 2015. Vol. 5. No. 2. P. 1–11. DOI: 10.5121/ijdkp.2015.5201

12. *Malchevskaya P. A.* Forecasting customer outflow based on a machine learning model / P. A. Malchevskaya // League of Researchers of Moscow State Pedagogical University: Collection of materials of the student open conference. In 2 parts (Moscow, November 20–24, 2023). Moscow: Moscow City University, 2024. P. 116–119.

13. *Frolov Yu. V.* AI/ML data platforms based on domestic software / Yu. V. Frolov, T. M. Bosenko, D. V. Yatsenko // Modern {digital} didactics. Moscow: A-Prior, 2023. P. 119–128.

Статья поступила в редакцию: 17.12.2024;
одобрена после рецензирования: 30.01.2025;
принята к публикации: 30.01.2025.

The article was submitted: 17.12.2024;
approved after reviewing: 30.01.2025;
accepted for publication: 30.01.2025.

Информация об авторе / Information about author:

Тимур Муртазович Босенко — кандидат технических наук, доцент департамента информатики, управления и технологий Института цифрового образования, Московский городской педагогический университет, Москва, Россия.

Timur M. Bosenko — Candidate of Technical Sciences, Associate Professor of the Department of IT, Management and Technology at the Institute of Digital Education, Moscow City University, Moscow, Russia.

bosenkotm@mgpu.ru, https://orcid.org/0000-0002-5375-096X