УДК 004.8
DOI: 10.54835/18102883_2024_35_5

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПРИМЕНЕНИЯ МОДЕЛЕЙ ЛИНЕЙНОЙ РЕГРЕССИИ И СЛУЧАЙНОГО ЛЕСА В ПРОГНОЗИРОВАНИИ УСПЕВАЕМОСТИ СТУДЕНТОВ

Левин Семен Михайлович,
кандидат юридических наук, PhD, профессор кафедры автоматизированных систем управления, факультет систем управления,
semen.m.levin@tusur.ru

Томский государственный университет систем управления и радиоэлектроники,
Россия, 634050, г. Томск, пр. Ленина, 40

Аннотация. В статье описан сравнительный анализ двух популярных моделей машинного обучения – линейной регрессии и случайного леса, применённых для прогнозирования академической успеваемости студентов. Актуальность исследования в области прогнозирования успеваемости студентов с использованием различных методов машинного обучения, таких как линейная регрессия и случайный лес, обусловлена несколькими ключевыми факторами. В современном образовательном пространстве наличие точных и эффективных инструментов для оценки и предсказания академических достижений студентов может оказать значительное влияние на качество образовательного процесса, удовлетворенность студентов обучением и в итоге на их будущее профессиональное развитие. Сравнительный анализ применения двух вышеуказанных методов в контексте прогнозирования успеваемости студентов не только расширяет понимание о возможностях каждого из подходов, но и способствует определению условий, при которых один метод может оказаться предпочтительнее другого. Учитывая разнообразие факторов, влияющих на академические достижения – от индивидуальных когнитивных способностей и мотивации до социально-экономических условий и качества образовательной среды, – выбор наиболее подходящего метода анализа данных является ключевым для разработки эффективных образовательных стратегий и инструментов. Целью исследования является сравнительная оценка эффективности двух моделей на основе данных, полученных из системы управления обучением. Основными методами исследования являются разработка и тестирование моделей машинного обучения, сравнительный и статистический анализ. В статье описано применение обеих моделей на реальных данных, анализ результатов с использованием кросс-валидации для оценки устойчивости моделей к переобучению и точности прогнозирования. Для оценки моделей используются критерии точности предсказаний, устойчивости к переобучению и интерпретируемости результатов. В заключение предлагаются рекомендации относительно выбора модели для конкретных задач в области образовательной аналитики.

Ключевые слова: машинное обучение, адаптивное обучение, линейная регрессия, случайный лес, система управления обучением, предиктивная аналитика

Введение

В эпоху цифровых технологий и быстро меняющегося мира машинное обучение воспринимается не просто как модный тренд, а как фундаментальное направление, кардинально преобразующее различные сферы жизни. В области высшего образования внедрение современных технологий открывает новые возможности, несмотря на инертность образовательной системы, – тема, которая уже давно обсуждается в контексте способности образовательных систем адаптироваться к изменениям в обществе, экономике и технологиях. Этот вопрос особенно актуален в свете бурного цифрового развития, глобализации и изменения требований рынка труда. Рассмотрим несколько ключевых аспектов, которые могут охарактеризовать степень инертности в высшем образовании. Прежде всего, университетское образование часто опирается на долгие традиции и устоявшиеся подходы к обучению, что замедляет внедрение новых технологий и новых методик обучения [1]. Помимо этого, во многих учебных заведениях процесс обновления учебных программ – длительная бюрократическая процедура, затрудняющая быструю адаптацию к новым требованиям внешней среды [2]. Также существует проблема несоответствия между навыками, которыми обладают выпускники, и требованиями рынка труда [3]. Это указывает на потребность в более гибком и актуализированном подходе к обучению вновь испечённых кадров. Да и сами учебные заведения зачастую страдают от недостатка квалифицированных кадров в таких новых и быстро развивающихся областях, как искусственный интеллект, биотехнологии и др. Многие университеты все еще опираются на традиционные лекционные методы обучения, в то время как современные образовательные

практики предполагают большую интерактивность, практическую работу и использование вспомогательных цифровых технологий.

Современное высшее образование в России сталкивается с задачей подготовки специалистов, способных работать в быстро меняющемся мире [4]. Адаптивное обучение [5], построенное с использованием моделей машинного обучения, предоставляет инструменты для индивидуального подхода к педагогическому процессу, позволяя приспосабливать учебные программы к особенностям каждого учащегося. Используя аналитику данных, вузы могут оптимизировать учебные планы, обеспечивая высокую актуальность и практическую применимость знаний [6]. В современном мире, где объемы данных растут в геометрической прогрессии, машинное обучение может стать одним из ключевых инструментов в обработке и анализе этой информации в части предиктивного прогнозирования и принятия решений. Среди множества подходов в машинном обучении особого внимания заслуживают две модели: линейная регрессия и случайный лес [7, 8]. Эти методы широко используются в различных областях, от финансового анализа до медицинских исследований, и представляют собой два различных подхода. Напомним, что машинное обучение – это подраздел искусственного интеллекта, фокусирующийся на разработке алгоритмов, которые позволяют компьютерам учиться и делать предсказания или решения на основе данных. Эта область объединяет элементы информатики, статистики и математического моделирования для анализа и интерпретации различных типов данных.

Обе упомянутые выше модели относятся к типу машинного обучения с учителем (supervised learning) [9]. Алгоритмы обучаются на помеченных данных, где каждому входному примеру соответствует известный ответ. Линейная регрессия – одна из старейших и наиболее понятных моделей [10]. Она предполагает линейную зависимость между входными переменными и целевой переменной. Простота, интерпретируемость и вычислительная эффективность делают её идеальным выбором для многих задач, особенно там, где важно понимать взаимосвязь между переменными. С другой стороны, случайный лес – это гораздо более современный ансамблевый метод, использующий множество деревьев решений для получения более точных и устойчивых предсказаний [11]. Такая модель эффективно справляется с нелинейными зависимостями и может автоматически учитывать взаимодействия между переменными.

В то время как каждая из этих моделей имеет свои преимущества, важно понимать их ограничения и условия, при которых одна модель может превосходить другую [12]. В рамках исследований в области адаптивного обучения с применением моделей машинного обучения автором проведен теоретический сравнительный анализ обеих моделей в части их практического применения для предиктивного прогнозирования успешного окончания курса студентами.

Теоретические основы моделей

Линейная регрессия является статистическим методом в машинном обучении, используемым для моделирования и анализа отношений между зависимой переменной и одной или несколькими независимыми переменными. Цель линейной регрессии состоит в нахождении линейной взаимосвязи между переменными, что позволяет предсказывать значения зависимой переменной на основе значений независимых [13].

Модель линейной регрессии обычно представляется уравнением вида:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n+\epsilon,$$

где y – зависимая (предсказываемая) переменная; β_0 – свободный член (интерцепт, то есть значение y x=0); $\beta_1,\beta_2,\ldots,\beta_n$ – коэффициенты модели (наклоны), показывающие величину влияния соответствующих независимых переменных $x_1,x_2,\ldots,x_n$; ϵ – ошибка модели, отражающая неучтенные факторы.

Для определения наилучших значений коэффициентов β используется метод наименьших квадратов [14]. Цель – минимизировать сумму квадратов разностей между наблюдаемыми значениями и значениями, предсказанными моделью.

На рис. 1 представлена графическая иллюстрация модели линейной регрессии, демонстрирующая взаимосвязь между независимой переменной (по оси абсцисс) и зависимой переменной (по оси ординат). Диаграмма включает в себя точечный график с множеством данных, расположенных в двумерной координатной системе. Через эти точки проведена линия наилучшего соответствия, демонстрирующая предполагаемую линейную зависимость между переменными.

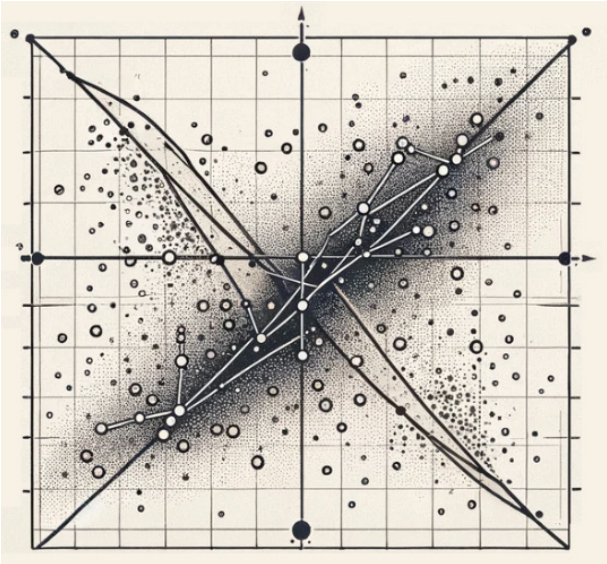

Рис. 1. Графическая иллюстрация модели линейной регрессии
Fig. 1. Graphic illustration of the linear regression model

Основные предположения модели:

- линейность отношений – отношения между независимыми и зависимой переменными должны быть линейными;
- независимость ошибок – ошибки (ϵ) должны быть независимыми и идентично распределенными;
- дисперсия ошибок должна быть постоянной для всех уровней независимых переменных;
- отсутствие мультиколлинеарности – независимые переменные не должны быть сильно коррелированными между собой.

Случайный лес является одним из наиболее популярных и мощных ансамблевых методов машинного обучения [15]. Он основан на комбинации множества деревьев решений и включает в себя следующие ключевые математические и статистические принципы:

- ансамблевое обучение – модель строится путем объединения большого количества деревьев решений. Каждое из них дает свой прогноз, результаты всех деревьев агрегируются для получения окончательного прогноза. Это снижает риск переобучения и повышает точность предсказаний по сравнению с одиночным деревом решений.
- бэггинг или бутстрапинг – каждое дерево модели обучается на случайно выбранной подвыборке исходного набора данных, созданной с помощью бутстрапа (выбор с возвращением). Это означает, что одни и те же наблюдения могут появляться несколько раз в одной подвыборке, в то время как другие могут вообще не появиться. То есть, если исходный набор данных содержит N наблюдений, каждая бутстрап-выборка также будет содержать N наблюдений, но некоторые наблюдения могут повторяться, в то время как другие могут отсутствовать.
- случайный выбор признаков – при построении каждого узла дерева выбирается случайное подмножество признаков из всего доступного набора, что увеличивает разнообразие в ансамбле и уменьшает корреляцию между отдельными деревьями, повышая общую устойчивость модели к переобучению. Количество признаков в подмножестве обычно меньше общего числа признаков.
- усреднение или голосование – для задач классификации итоговый класс определяется по принципу «голосования большинства» среди всех деревьев;
- нелинейность – деревья решений могут естественным образом моделировать нелинейные взаимодействия между переменными, что даёт высокую потенциальную эффективность для анализа сложных наборов данных.

Благодаря своей структуре и методу обучения случайный лес относительно устойчив к шуму в данных и присутствию выбросов [16].

Каждое дерево строится независимо от остальных [17]. Процесс разбиения узлов в дереве продолжается до достижения максимальной глубины или до тех пор, пока в узле не останется минимально допустимое количество наблюдений. Поскольку каждое дерево обучается на разных выборках и с разными подмножествами признаков, ансамбль деревьев, как правило, демонстрирует меньшую дисперсию и лучшую обобщающую способность, чем отдельные деревья [18].

На рис. 2 представлена схематическая визуализация алгоритма, объединяющего множество деревьев решений, каждое из которых обучается на случайно выбранном подмножестве данных, для создания ансамбля, способного к высокоточной классификации или регрессии. Входные данные, представленные на диаграмме, демонстрируют начальный набор данных, который разделяется на подмножества для обучения отдельных деревьев решений. Эти деревья строятся по методу, при котором каждое дерево обучаемо на случайно сгенерированном подмножестве изначальных данных с возвращением, что обеспечивает разнообразие в обучающем процессе и помо-

гает уменьшить переобучение, характерное для отдельных деревьев решений.

Основные предположения модели:

- в отличие от многих традиционных статистических методов, случайный лес не требует предположений о распределении данных. Это делает его подходящим для работы с различными типами данных, включая непрерывные и категориальные переменные;
- нелинейность – хорошо справляется с нелинейными взаимодействиями между переменными;
- предположение о предикторах в этой модели подразумевает, что среди набора предикторов (признаков) имеются переменные, значимо влияющие на зависимую переменную. Модель может не работать наилучшим образом, если все предикторы слабо связаны с целевой переменной;
- большое количество деревьев улучшает производительность и точность модели за счет снижения вариабельности;
- устойчивость к переобучению, по сравнению с отдельными деревьями решений.

Тем не менее модель может страдать от переобучения в случаях с очень шумными данными либо когда число наблюдений значительно меньше числа признаков.

Рис. 2. Схематическая визуализация алгоритма случайного леса
Fig. 2. Schematic visualization of the Random Forest algorithm

Критерии сравнения моделей

В качестве критериев сравнения рассматриваемых моделей машинного обучения были выбраны следующие:

1. Точность предсказаний. Методы оценки: для задач регрессии – средняя квадратичная ошибка (mean square error – MSE).
2. Устойчивость к переобучению – способность модели обобщать, а не запоминать данные. Метод оценки: перекрестная проверка (cross-validation).
3. Интерпретируемость результатов. Субъективная оценка с позиции легкости объяснения результатов работы модели пользователям.

Выбор набора данных

При проведении сравнительного анализа моделей линейной регрессии и случайного леса в машинном обучении был выбран малый набор данных системы управления обучением Moodle [19]. В контексте анализа данных в системах управления обучением (Learning Management Systems – LMS) понятия «малых» и «больших» наборов данных также могут варьироваться [20], но для классификации набора были приняты следующие численные значения: количество объектов – до нескольких сотен; количество признаков – до двух десятков.

При составлении набора данных для анализа включены следующие типы данных:

- оценки по заданиям и зачётам/экзаменам – прямые индикаторы академической успеваемости;
- участие в обсуждениях и форумах – активность на форумах и в обсуждениях принята как показатель вовлеченности и понимания материала;
- данные о времени входа в систему, продолжительности сессий и частоте посещений курса как показатель уровня вовлеченности студента в учебный процесс;
- завершение учебных элементов курсов – прогресс в прохождении различных элементов курса, включая просмотренные видеолекции, прочитанные лекции, иные материалы, а также выполненные обучающие тесты;

Выбранный набор данных включал в себя указанные типы в отношении четырёх групп студентов общей численностью 148 человек, завершивших четырехлетнее обучение, за его двухлетний период (первые два курса).

Подготовка данных

Данные, необходимые для использования, получены из базы данных LMS Moodle и объединены в единый набор. Устранены дубликаты, проведена проверка целостности данных. Далее была выполнена их очистка – удалены ошибки и пропуски, выбросы, заполнены пропущенные значения, исправлены ошибки. После выполнения указанных действий данные преобразованы в формат, пригодный для анализа, – осуществлены нормализация или стандартизация числовых данных, кодирование категориальных переменных, создание производных признаков. Полученные в результате данные разбиты на обучающую и тестовую выборки случайным образом, с распределением 80 % на обучение и 20 % на тестирование.

Импорт данных, очистка и предварительная обработка, разделение на обучающую и тестовую выборки выполнены в Phyton, так же как последующее создание моделей линейной регрессии и случайного леса.

Результаты

Полученные результаты применения каждой из моделей сравнены между собой по указанным ранее критериям.

Точность прогнозирования. Модель случайного леса показала более высокую точность по сравнению с линейной регрессией. Совпадение прогноза, выполненного первой моделью, с фактическими результатами обучения студентов составило 92,3 % в отличии 83,7 % второй. Оценка точности произведена с использованием MSE, с использованием пятикратной кросс-валидации. Среднее значение MSE для линейной регрессии MSE=3,92, для случайного леса MSE=2,6.

Случайный лес показал более низкое среднее значение MSE по сравнению с линейной регрессией, что указывает на меньшие ошибки в предсказаниях модели и, следовательно, на более высокую точность.

Устойчивость к переобучению. Для оценки применен метод K-кратной кросс-валидации, где набор данных был разделён на пять подмножеств, оценка модели произведена на основе коэффициента детерминации R^2. Среднее значение для линейной регрессии R^2=0,76, стандартное отклонение σ=0,020, для случайного леса R^2=0,81 и σ=0,015 соответственно.

Таким образом, случайный лес показал более высокое среднее значение R^2 и меньшее

стандартное отклонение σ, что указывает на лучшую устойчивость модели по сравнению с линейной регрессией.

Интерпретируемость результатов. Линейная регрессия предоставляет прямые количественные выводы о взаимосвязи между каждой переменной и итоговыми оценками. Например, коэффициенты модели +0,5 для времени входа в систему и –0,3 для частоты посещений форумов означают, что увеличение времени входа в систему на одну единицу (например, на один час) связано со средним увеличением оценки на 0,5 пункта, при условии, что другие факторы остаются неизменными. Аналогично увеличение частоты посещений форумов на одну единицу (например, на одно посещение) связано со средним уменьшением оценки на 0,3 пункта.

В модели случайного леса мы можем обнаружить, что, например, оценки по заданиям и продолжительность сессий являются наиболее важными факторами для прогнозирования итоговой оценки. Однако, в отличие от линейной регрессии, лес не дает прямого понимания о влиянии изменений этих переменных на итоговую оценку. Мы знаем, что они важны, но не знаем точно, как и почему.

Выводы

Линейная регрессия предоставляет прямые количественные выводы о взаимосвязи между каждой переменной и итоговыми оценками. Это делает ее полезной для понимания и обьяснения, какие факторы и как влияют на успеваемость студента. Случайный лес, хотя и обеспечивает информацию о том, какие переменные в целом значимы, не предоставляет чёткого понимания зависимости итогового результата от изменения этих переменных. При этом точность прогнозирования и степень устойчивости к переобучению у случайного леса выше, чем у модели линейной регрессии. Таким образом, в рамках проведённого эксперимента применение модели случайного леса более результативно, чем регрессии. Однако при более широком использовании выбор в пользу одной их двух моделей должен учитывать цели анализа. Если нужно четко понимать влияние каждой переменной, то более оптимальным станет применение модели линейной регрессии. В случае отсутствия предубеждений против «черного ящика» и стремлении к более высокой точности предиктивного прогнозирования стоит выбрать случайный лес.

СПИСОК ЛИТЕРАТУРЫ

1. Современное университетское образование: тенденции развития и проблемы трансформации / под ред. Т.А. Костюковой, Л.Г. Смышляевой. – Томск: Издательство Томского государственного университета, 2023. – 342 с.
2. Балацкий Е.В. Переформатирование российского университета в условиях гибридной войны: практико-ориентированная модель // Journal of Economic Regulation. – 2022. – Vol. 13. – № 4. – P. 24–38. DOI: 10.17835/2078-5429.2022.13.4.024-038
3. Ильин И.В., Багаева И.В. Требования к компетентностной модели выпускника университета в условиях цифровой экономики // Наука и бизнес: пути развития. – 2020. – №. 4. – С. 71–75.
4. Афанасьева Д.О., Казаева Е.А. Цифровые инновации в образовании: перспективы и вызовы для университетов // Вестник Шадринского государственного педагогического университета. – 2023. – № 1 (57). – С. 104–109. DOI: 10.52772/25420291_2023_1_104
5. Кречетов И.А., Романенко В.В. Реализация методов адаптивного обучения // Вопросы образования. – 2020. – № 2. – С. 252–277. DOI: 10.17323/1814-9545-2020-2-252-277
6. Levin S.M. Personality-oriented Learning with the Use of Electronic Technologies Based on the Analysis of LMS Data // Современное образование: интеграция образования, науки, бизнеса и власти: Материалы международной научно-методической конференции. В 2-х частях. – Томск, 27–28 января 2022. Ч. 1. – Томск: Томский государственный университет систем управления и радиоэлектроники, 2022. – P. 21–28. EDN ZTNONT.
7. Maulud D., Abdulazeez A.M. A review on linear regression comprehensive in machine learning // Journal of Applied Science and Technology Trends. – 2020. – Vol. 1. – № 4. – P. 140–147.
8. House price prediction using random forest machine learning technique / Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala, Ololade Oyewo, Yetunde Faith Akande, Gbenle Oluwadara // Procedia Computer Science. – 2022. – Vol. 199. – P. 806–813. DOI: https://doi.org/10.1016/j.procs.2022.01.100
9. Khe Foon Hew. What predicts student satisfaction with MOOCs: a gradient boosting trees supervised machine learning and sentiment analysis approach // Computers & Education. – 2020. – Vol. 145. – P. 103724. DOI: https://doi.org/10.1016/j.compedu.2019.103724
10. Luan H., Tsai C.C. A review of using machine learning approaches for precision education // Educational Technology & Society. – 2021. – Vol. 24. – № 1. – P. 250–266.

11. Tzenios N. Examining the impact of EdTech integration on academic performance using random forest regression // Researchberg Review of Science and Technology. – 2020. – Vol. 3. – № 1. – P. 94–106.
12. A comparative analysis of logistic regression, random forest and KNN models for the text classification / Kanish Shah, Henil Patel, Devanshi Sanghvi, Manan Shah // Augmented Human Research. – 2020. – Vol. 5. DOI: https://doi.org/10.1007/s41133-020-00032-0
13. Ali P., Younas A. Understanding and interpreting regression analysis // Evidence-Based Nursing. – 2021. – Vol. 24. – Iss. 4. – P. 116–118. DOI: https://doi.org/10.1136/ebnurs-2021-103425
14. Ji Y., Jiang X., Wan L. Hierarchical least squares parameter estimation algorithm for two-input Hammerstein finite impulse response systems // Journal of the Franklin Institute. – 2020. – Vol. 357. – № 8. – P. 5019–5032.
15. A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives and opportunities / S. González, S. García, J. del Ser, L. Rokach, F. Herrera // Information Fusion. – 2020. – Vol. 64. – P. 205–237. DOI: https://doi.org/10.1016/j.inffus.2020.07.007
16. Data cleaning method for the process of acid production with flue gas based on improved random forest / Xiaoli Li, Minghua Liu, Kang Wang, Zhiqiang Liu, Guihai Li // Chinese Journal of Chemical Engineering. – 2023. – Vol. 59. – P. 72–84. DOI: https://doi.org/10.1016/j.cjche.2022.12.013
17. Hehn T.M., Kooij J.F.P., Hamprecht F.A. End-to-end learning of decision trees and forests // International Journal of Computer Vision. – 2020. – Vol. 128. – № 4. – C. 997–1011.
18. Comparative performance of eight ensemble learning approaches for the development of models of slope stability prediction / Shan Lin, Hong Zheng, Bei Han, Yanyan Li, Chao Han, Wei Li // Acta Geotechnica. – 2022. – Vol. 17. – P. 1477–1502. DOI: https://doi.org/10.1007/s11440-021-01440-1
19. Moodle. URL: https://moodle.com (дата обращения 07.02.2024).
20. Oh Y., Park S., Ye J.C. Deep learning COVID-19 features on CXR using limited training data sets // IEEE transactions on medical imaging. – 2020. – Vol. 39. – № 8. – P. 2688–2700. DOI: 10.1109/TMI.2020.2993291

Поступила: 15.02.2024
Принята: 28.05.2024
Опубликована: 30.06.2024

UDC 004.8
DOI: 10.54835/18102883_2024_35_5

COMPARATIVE ANALYSIS OF THE APPLICATION OF LINEAR REGRESSION AND RANDOM FOREST MODELS IN PREDICTING STUDENT PERFORMANCE

Semen M. Levin,
Cand. Sc., PhD, Professor,
semen.m.levin@tusur.ru

Tomsk State University of Control Systems and Radioelectronics,
40, Lenin avenue, Tomsk, 634050, Russian Federation

Abstract. The article presents a comparative analysis of two popular machine learning models – linear regression and random forest that are applied to predict students' academic performance. The relevance of research in predicting student performance using various machine learning methods, such as linear regression and random forest, is driven by several key factors. In the modern educational landscape, the presence of accurate and effective tools for assessing and predicting students' academic achievements can significantly impact the quality of the educational process, student satisfaction with learning, and, ultimately, their future professional development. The comparative analysis of the two methods mentioned above in the context of predicting student performance not only broadens the understanding of the capabilities of each approach but also aids in determining the conditions under which one method may be preferable over the other. Considering the variety of factors affecting academic achievements – from individual cognitive abilities and motivation to socio-economic conditions and the quality of the educational environment – choosing the most suitable data analysis method is key to develop effective educational strategies and tools. The study aims to comparatively evaluate the effectiveness of the two models based on data obtained from the learning management system. The main research methods include developing and testing machine learning models and comparative and statistical analysis. The article describes the application of both models on real data and the analysis of the results using cross-validation to assess the models resistance to overfitting and prediction accuracy. Criteria for evaluating the models include accuracy of predictions, resistance to overfitting, and interpretability of results. In conclusion, recommendations are offered regarding the choice of model for specific tasks in educational analytics.

Keywords: machine learning, adaptive learning, linear regression, random forest, learning management system, predictive analytics

REFERENCES

1. *Modern university education: development trends and problems of transformation.* Eds. T.A. Kostyukova, L.G. Smyshlyaeva. Tomsk, Tomsk State University Publ. House, 2023. 342 p. (In Russ.)
2. Balatsky E.V. Reformatting Russian university in hybrid war: Practice-oriented model. *Journal of Economic Regulation*, 2022, vol. 13, no. 4, pp. 24–38. (In Russ.) DOI: 10.17835/2078-5429.2022.13.4.024-038
3. Ilyin I.V., Bagaeva I.V. Requirements for the competency model of a university graduate in the digital economy. *Science and business: ways of development*, 2020, no. 4, pp. 71–75. (In Russ.)
4. Afanasyeva D.O., Kazaeva E.A. Digital innovations in education: perspectives and challenges for universities. *Journal of Shadrinsk state pedagogical university*, 2023, no. 1 (57), pp. 104–109. (In Russ.) DOI: 10.52772/25420291_2023_1_104
5. Krechetov I.A., Romanenko V.V. Implementing the adaptive learning techniques. *Educational Studies Moscow*, 2020, no. 2, pp. 252–277. (In Russ.) DOI:10.17323/1814-9545-2020-2-252-277
6. Levin S.M. Personality-oriented learning with the use of electronic technologies based on the analysis of LMS data. *Modern education: integration of education, science, business and government. Proceedings of the international scientific and methodological conference.* Tomsk, January 27–28, 2022. P. 1. Tomsk, Tomsk State University of Control Systems and Radioelectronics, 2022. pp. 21–28. (In Russ.) EDN ZTNONT.
7. Maulud D., Abdulazeez A.M. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 2020, vol. 1, no. 4, pp. 140–147.
8. Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala, Ololade Oyewo, Yetunde Faith Akande, Gbenle Oluwadara. House price prediction using random forest machine learning technique. *Procedia Computer Science*, 2022, vol. 199, pp. 806–813. DOI: https://doi.org/10.1016/j.procs.2022.01.100
9. Khe Foon Hew. What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, 2020, vol. 145, pp. 103724. DOI: https://doi.org/10.1016/j.compedu.2019.103724
10. Luan H., Tsai C.C. A review of using machine learning approaches for precision education. *Educational Technology & Society*, 2021, vol. 24, no. 1, pp. 250–266.

11. Tzenios N. Examining the impact of EdTech integration on academic performance using random forest regression. *Researchberg Review of Science and Technology*, 2020, vol. 3, no. 1, pp. 94–106.
12. Kanish Shah, Henil Patel, Devanshi Sanghvi, Manan Shah. A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, 2020, vol. 5. DOI: https://doi.org/10.1007/s41133-020-00032-0
13. Ali P., Younas A. Understanding and interpreting regression analysis. *Evidence-Based Nursing*, 2021, vol. 24, Iss. 4, pp. 116–118. DOI: https://doi.org/10.1136/ebnurs-2021-103425
14. Ji Y., Jiang X., Wan L. Hierarchical least squares parameter estimation algorithm for two-input Hammerstein finite impulse response systems. *Journal of the Franklin Institute*, 2020, vol. 357, no. 8, pp. 5019–5032.
15. González S., García S., Del Ser J., Rokach L., Herrera F. A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 2020, vol. 64, pp. 205–237. DOI: https://doi.org/10.1016/j.inffus.2020.07.007
16. Xiaoli Li, Minghua Liu, Kang Wang, Zhiqiang Liu, Guihai Li. Data cleaning method for the process of acid production with flue gas based on improved random forest. *Chinese Journal of Chemical Engineering*, 2023, vol. 59, pp. 72–84. DOI: https://doi.org/10.1016/j.cjche.2022.12.013
17. Hehn T.M., Kooij J.F.P., Hamprecht F.A. End-to-end learning of decision trees and forests. *International Journal of Computer Vision*, 2020, vol. 128, no. 4, pp. 997–1011.
18. Shan Lin, Hong Zheng, Bei Han, Yanyan Li, Chao Han, Wei Li. Comparative performance of eight ensemble learning approaches for the development of models of slope stability prediction. *Acta Geotechnica*, 2022, vol. 17, pp. 1477–1502. DOI: https://doi.org/10.1007/s11440-021-01440-1
19. Moodle. Available at: https://moodle.com (accessed: 7 February 2024).
20. Oh Y., Park S., Ye J.C. Deep learning COVID-19 features on CXR using limited training data sets. *IEEE transactions on medical imaging*, 2020, vol. 39, no. 8, pp. 2688–2700. DOI: 10.1109/TMI.2020.2993291

Received: 15.02.2024
Revised: 28.05.2024
Accepted: 30.06.2024