УДК 004.942
DOI

ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ БУДУЩЕЙ УСПЕВАЕМОСТИ АБИТУРИЕНТОВ В ВУЗЕ

Ромащенко А.И., Щербин С.И., Харитонов И.М., Панфилов А.Э., Жаравин Н.С.

Камышинский технологический институт (филиал) ФГБОУ ВО «Волгоградский государственный технический университет», Камышин, e-mail: wisdom_monk@mail.ru

Традиционные методы оценки абитуриентов, основанные на анализе школьных достижений и экспертных заключений, часто демонстрируют низкую точность, что приводит к ошибкам в зачислении и снижению академической успеваемости студентов. В контексте усиления конкуренции между высшими учебными заведениями актуализируется задача повышения точности прогнозирования образовательных результатов. Целью данного исследования является прогнозирование успеваемости студентов на основе сравнительного анализа и оптимизации теоретических моделей машинного обучения для повышения точности оценки их академических результатов. Использованы данные 1662 студентов, включающих школьные аттестационные оценки, результаты Основного государственного экзамена и Единого государственного экзамена, баллы за выпускную квалификационную работу и экспертные оценки, проведено кросс-валидационное тестирование. Предобработка данных осуществлена методом StandardScaler, оценка моделей выполнена по метрикам MSE, R^2, MAE, и MAPE. Результаты демонстрируют статистически значимое превосходство модели «случайный лес» (MSE = 0,18, R^2 = 0,89) над линейной (R^2 = 0,65) и полиномиальной регрессией, что объясняется ее способностью к аппроксимации нелинейных зависимостей и минимизации переобучения за счет бэггинга. Исследование подтверждает эффективность современных алгоритмов для решения задач высшей школы.

Ключевые слова: успеваемость, модель, линейная регрессия, полиномиальная регрессия, случайный лес

THE USE OF STATISTICAL MACHINE LEARNING METHODS TO PREDICT THE FUTURE ACADEMIC PERFORMANCE OF UNIVERSITY APPLICANTS

Romaschenko A.I., Scherbin S.I., Kharitonov I.M., Panfilov A.E., Zharavin N.S.

Kamyshin Technological Institute of the Volgograd State Technical University, Kamyshin, e-mail: wisdom_monk@mail.ru

Traditional methods of assessing applicants, based on the analysis of academic achievements and expert evaluations, often demonstrate low accuracy, leading to errors in admissions and reduced academic performance among students. In the context of growing competition among higher education institutions, the task of improving the accuracy of predicting academic outcomes becomes critical. The purpose of this study is to predict students' academic performance based on comparative analysis and optimization of theoretical machine learning models to improve the accuracy of their academic performance assessment. Data from 1,662 students, including school grades, results of the Main State Exam (OGE) and Unified State Exam (EGE), scores for final qualification works, and expert evaluations, were analyzed using cross-validation testing. Data preprocessing was performed using the StandardScaler method, and model evaluation was conducted using MSE, R^2, MAE, and MAPE metrics. The results demonstrate statistically significant superiority of the "random forest" model (MSE = 0.18, R^2 = 0.89) over linear (R^2 = 0.65) and polynomial regression, attributed to its ability to approximate nonlinear dependencies and minimize overfitting through bagging. The study confirms the effectiveness of modern algorithms in addressing challenges in higher education.

Keywords: academic performance, model, linear regression, polynomial regression, random forest

Введение

Методы оценки абитуриентов зачастую характеризуются недостаточной точностью и объективностью, что ведет к ошибочному отбору студентов и снижению их академической успеваемости [1]. На фоне усиления конкуренции между вузами и сокращения числа поступающих актуализируется необходимость оптимизации системы отбора [2]. Неточности в оценках повышают для абитуриентов риск выбора образовательной программы или вуза, не соответствующей их потребностям, что ухудшает успеваемость и снижает удовлетворенность учебным процессом [3, 4]. В связи с этим абитуриенты стремятся получить уверенность в том, что выбранное направление поможет достичь как академических, так и профессиональных целей.

Цель исследования – прогнозирование успеваемости студентов на основе сравнительного анализа и оптимизации теоретических моделей машинного обучения для повышения точности оценки их академических результатов.

Основное внимание в исследовании уделяется следующим задачам:

1. Описание данных и их нормализация.
2. Описание оценочных метрик.
3. Описание каждой созданной модели.
4. Сводная таблица с оценками моделей.

В результате проведения исследования авторы ставили гипотезу о возможности прогнозирования будущей успеваемости абитуриента в вузе по выбранной специальности, на этапе окончания им школы и наличия только оценочных «объективных» результатов школьной успеваемости, без учета личностных качеств абитуриента и школьной жизни.

Материалы и методы исследования

В исследовании используются данные по 1662 студентам направления 09.03.01 «Информатика и вычислительная техника». Исходные данные включают в себя оценки по 17 школьным предметам, средний балл за школьный аттестат, результаты ОГЭ и ЕГЭ по математике, результат ЕГЭ по физике [5], а также вузовскую оценку за выпускную квалификационную работу (ВКР) и вузовскую среднюю экспертную оценку всех преподавателей выпускающей кафедры.

Данные распределены следующим образом. Входными данными являются оценки по школьным предметам, которые включают алгебру, геометрию, физику, информатику, русский язык, литературу, историю, биологию, химию, обществознание, географию, физическую культуру, иностранный язык, основы безопасности жизнедеятельности (ОБЖ), музыку, изобразительное искусство (ИЗО), технологию, средний балл за школьный аттестат по всем предметам, результаты ОГЭ по математике (все эти баллы в диапазоне оценок от 3 до 5). Также учитываются входные данные в виде результатов ЕГЭ по математике и по физике [6] (диапазон баллов от 40 до 100). Выходными данными являются оценки за ВКР и средняя экспертная оценка преподавателей (подробнее формирование оценки описано в статье [4]) с диапазоном баллов от 61 до 100. Фрагмент имеющихся данных изображен на рис. 1.

Для улучшения качества моделей и повышения точности прогнозирования все данные были нормализованы. Нормализация данных была выполнена с использованием библиотеки scikit-learn и ее класса StandardScaler, который стандартизирует данные, приводя их к нормальному распределению с нулевым средним и единичным стандартным отклонением [7].

Процесс стандартизации включает следующие шаги:

1. Вычисление среднего значения и стандартного отклонения для каждого признака. Среднее значение (mean) и стандартное отклонение (standard deviation) вычисляются по каждому признаку на обучающей выборке [8].

2. Трансформация данных. Для каждого признака выполняется операция, описываемая формулой

$$z = (x - \mu) / \sigma, \quad (1)$$

где x – значение признака, μ – среднее значение признака, σ – стандартное отклонение признака.

Фрагмент нормализованных данных изображен на рис. 2.

Certificate	Algebra	Geometry	Physics	Computer Science		Physics(EGE)	Math(EGE)	Math(OGE)	Prepod_mark	VKR
4.88	5	5	5	5		46	59	5	73.6	92
4.13	4	5	4	4		43	27	4	37.1	74
4.82	3	3	4	5		45	30	3	21.6	68
4.83	3	4	4	5		53	30	3	25.7	78
4.33	4	4	3	4		53	30	4	25.7	90
4.35	4	4	4	4	...	44	62	3	71.4	86
5.0	5	5	5	5		48	45	4	97.9	100
3.94	4	3	3	4		42	33	4	45.0	83

Рис. 1. Фрагмент исходных данных

Certificate	Algebra	Geometry	Physics	Computer Science		Physics(EGE)	Math(EGE)	Math(OGE)
-2.256	-0.044	-0.122	-0.073	-2.644		-1.116	-1.378	-1.956
-0.710	-0.044	1.333	1.330	-0.663		-1.116	-1.378	-0.272
0.484	1.519	1.333	1.330	1.317		-0.890	-1.378	-0.272
0.659	1.519	1.333	1.330	1.317		-1.116	-1.378	-0.272
0.484	1.519	1.333	1.330	-0.663		-1.116	-1.378	-0.272
0.484	1.519	1.333	1.330	1.317	...	-1.116	-1.378	-0.272
0.834	1.519	1.333	1.330	1.317		-0.663	-1.147	-0.272

Рис. 2. Фрагмент нормализованных данных

Для оценки производительности моделей машинного обучения в рамках данного исследования используются следующие основные метрики: среднеквадратичная ошибка (Mean Squared Error, MSE), коэффициент детерминации (R^2 или R-squared), средняя абсолютная ошибка (Mean Absolute Error, MAE) и средняя абсолютная процентная ошибка (Mean Absolute Percentage Error, MAPE). Эти метрики позволяют объективно оценить точность и качество моделей, выявить наиболее подходящую из них и сделать обоснованные выводы о применимости различных методов машинного обучения [9].

Среднеквадратичная ошибка (MSE) измеряет среднее квадратичное отклонение предсказанных значений от фактических. Чем меньше значение MSE, тем точнее модель, так как это означает меньшую ошибку предсказания в среднем.

Коэффициент детерминации (R^2) показывает, какая доля вариации зависимой переменной объясняется независимыми переменными модели. Значение R^2 варьируется от 0 до 1, где 1 означает, что модель идеально объясняет все вариации данных.

Средняя абсолютная ошибка (MAE) измеряет среднее абсолютное отклонение предсказанных значений от фактических. Это более интерпретируемая метрика, так как она выражается в тех же единицах, что и исходные данные, и позволяет понять, на сколько в среднем модель ошибается.

Средняя абсолютная процентная ошибка (MAPE) измеряет среднюю абсолютную процентную ошибку предсказаний и часто используется для интерпретации точности моделей в процентах, что удобно для сравнения моделей в разных контекстах.

Результаты исследования и их обсуждение

В рамках исследования для прогнозирования академической успеваемости студентов последовательно применялись три метода машинного обучения: линейная регрессия – как базовая модель для выявления линейных зависимостей, полиномиальная регрессия – для анализа нелинейных взаимосвязей и случайный лес – как ансамблевый алгоритм, способный эффективно обрабатывать сложные данные с высокой размерностью. Ниже представлены детализированные результаты, их интерпретация и обоснование выбора методов, подтверждающие ключевую роль современных алгоритмов в решении задач образовательной аналитики. Сравнительный анализ эффективности линейной, полиномиальной регрессии и случайного леса в прогнозировании академической успеваемости студентов проведен на основе ключевых метрик (MSE, R^2, MAE, MAPE). Основная задача – определить, насколько разные модели способны выявлять взаимосвязи между школьными оценками, экзаменационными результатами и экспертными оценками. Далее представлены детализированные результаты, интерпретация точности алгоритмов и обоснование выбора оптимального подхода.

Линейная регрессия – это базовая модель машинного обучения, используемая для задач регрессии. Она оценивает линейные зависимости между входными признаками и целевой переменной. Основная идея заключается в нахождении линейной зависимости, которая минимизирует сумму квадратов ошибок между предсказанными и фактическими значениями.

Математическое обоснование регрессии заключается в использовании метода наименьших квадратов для нахождения оптимальных коэффициентов линейной модели, минимизирующих сумму квадратов ошибок. Модель описывается уравнением

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n, \qquad (2)$$

где $\hat{y}$ – предсказанное значение, β_0 – свободный член, $\beta_1, \beta_2, \ldots, \beta_n$ – коэффициенты модели, $x_1, x_2, \ldots, x_n$ – входные признаки.

Для достаточного качества модели были подобраны оптимальные гиперпараметры с использованием базового метода обучения, предоставляемого библиотекой scikit-learn [10]. Данная модель не требует значительной настройки параметров, так как использует стандартный метод наименьших квадратов. Оценки качества модели представлены в табл. 1.

Таблица 1

Оценки качества модели линейной регрессии

Метрика	Prepod_mark	VKR
MSE	178,2677	400,9954
R^2	0,6873	0,7280
MAE	9,8820	15,0149
MAPE	25,7830	–

Основываясь на полученных значениях, можно сделать следующий вывод о качестве модели. Данная регрессия показала недостаточную точность предсказаний для обеих целевых переменных, что говорит о слабой применимости данной модели для решения такой задачи.

Полиномиальная регрессия – это расширение линейной регрессии, которое позволяет моделировать нелинейные зависимости путем добавления полиномиальных признаков. Основная идея заключается в трансформации исходных признаков в по-

линомиальные и применении линейной регрессии на новых признаках.

Математическое обоснование полиномиальной регрессии заключается в использовании метода наименьших квадратов для нахождения оптимальных коэффициентов полиномиальной модели, минимизирующих сумму квадратов ошибок. Модель описывается уравнением

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n, \qquad (3)$$

где $\hat{y}$ – предсказанное значение, β_0 – свободный член, β_1, β_2, …, β_n – коэффициенты модели, x_1, x_2, …, x_n – исходные признаки, n – степень полинома.

Оценки качества модели представлены в таблице 2.

Таблица 2

Оценки качества модели полиномиальной регрессии

Метрика	Prepod_mark	VKR
MSE	75,5864	181,9152
R^2	0,8674	0,8766
MAE	4,8756	7,6984
MAPE	13,8591	–

Основываясь на полученных значениях, можно сделать следующий вывод о качестве модели. Полиномиальная регрессия показала значительно лучшие результаты по сравнению с линейной регрессией для прогнозирования оценок преподавателей и ВКР студентов. Значительно более низкие значения MSE указывают на то, что модель делает меньшие ошибки в предсказаниях. Высокие значения R^2 подтверждают, что модель хорошо объясняет вариации данных. Более низкие значения MAE и MAPE для оценки преподавателей указывают на более точные предсказания.

Таким образом, полиномиальная регрессия лучше справляется с задачей прогнозирования академической успеваемости студентов, обеспечивая более точные и надежные результаты [11].

Случайный лес (Random Forest)– это ансамблевый метод машинного обучения, который использует множество деревьев решений для улучшения точности и устойчивости модели [12]. Метод случайного леса основан на идее создания множества деревьев решений с использованием случайных подвыборок данных и случайных подмножеств признаков. Прогноз конечной модели получается путем усреднения прогнозов всех деревьев.

Математическое обоснование случайного леса заключается в построении N деревьев решений, где каждое дерево обучается на случайной подвыборке данных с использованием случайного подмножества признаков. Итоговый прогноз для регрессии получается путем усреднения прогнозов всех деревьев:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i, \qquad (4)$$

где $\hat{y}_i$ – прогноз i-го дерева.

Случайный лес реализован с помощью библиотеки scikit-learn. Инструменты, использованные для реализации, включают Random Forest Regressor для построения модели случайного леса [13].

Оценки качества модели представлены в таблице 3.

Таблица 3

Оценки качества модели случайный лес

Метрика	Prepod_mark	VKR
MSE	22,7476	9,1656
R^2	0,9601	0,9938
MAE	2,3277	1,1600
MAPE	5,1938	–

Основываясь на полученных значениях, можно сделать следующий вывод о качестве модели. Случайный лес показал лучший результат среди выбранных моделей для прогнозирования академической успеваемости студентов [14, 15]. Значительно более низкие значения MSE и высокие значения R^2 указывают на то, что модель случайного леса способна эффективно объяснять вариации данных и делать точные прогнозы. Значения MAE также подтверждают высокую точность модели, а низкий MAPE для оценки Prepod_mark указывает на малую среднюю абсолютную процентную ошибку.

Итоговое сравнение моделей приведено в табл. 4.

Таблица 4

Сравнение значимых оценок исследуемых моделей

Модель	Prepod_mark (MSE)	Prepod_mark (R^2)	VKR (MSE)	VKR (R^2)
Линейная регрессия	178,27	0,6873	401,00	0,7280
Полиномиальная регрессия	75,59	0,8674	181,92	0,8766
Случайный лес	22,75	0,9601	9,17	0,9938

Заключение

Проведенное исследование подтвердило, что применение алгоритмов машинного обучения, в частности случайного леса, позволяет существенно повысить точность прогнозирования успеваемости студентов за счет учета нелинейных зависимостей в данных. Линейная регрессия, несмотря на простоту интерпретации, показала ограниченную эффективность, а полиномиальная модель может давать результаты достаточной точности, но при этом требует дополнительной настройки для минимизации риска переобучения.

Полученные результаты работы показывают, что благодаря применению современных методов машинной обработки статистических данных возможно проведение прогнозирования будущей успеваемости студента в вузе еще на этапе его поступления в вуз на выбранное направление обучения, располагая только данными школьных оценок об успеваемости. Отдельно можно отметить, что использование только «объективных» данных о школьной успеваемости, без учета «субъективных» факторов, таких как мотивация обучающегося, взаимодействие с учителем, интерес к школьному предмету и т.д., также может показывать неплохие результаты прогнозирования. Однако авторы работы считают, что внедрение некоторого количества «субъективных» факторов в качестве исходных данных прогнозирования может помочь еще больше повысить качество прогнозирования успеваемости, в связи с чем планируется продолжать дальнейшие исследования в данной области.

Практическая значимость работы видится в создании автоматизированных систем самостоятельного прогнозирования возможной будущей успеваемости абитуриентов в вузе для выбранных направлений обучения. Применение таких систем поможет абитуриентам снизить риск неудачного выбора будущего направления обучения, а следовательно, повысить его возможную компетентность на рынке труда в будущем.

Список литературы

1. Щербин С.И., Харитонов И.М., Огар Т.П., Панфилов А.Э., Кравец А.Г. Применение методов кластерного анализа в задаче прогнозирования успеваемости абитуриентов в вузе // Современные наукоемкие технологии. 2024. № 5–2. С. 321–325. URL: https://top-technologies.ru/ru/article/view?id=40046 (дата обращения: 17.07.2025). DOI: 10.17513/snt.40046. EDN: MWIKXH.

2. Исаева Е.Р., Тюсова О.В., Тишков А.В., Шапоров А.М., Павлова О.В., Ефимов Д.А., Власов Т.Д. Поиск прогностических критериев академической успеваемости // Университетское управление: практика и анализ. 2017. Т. 21. № 2. С. 163–172. URL: https://www.umj.ru/jour/article/view/86 (дата обращения: 17.07.2025).

3. Алпатов А.В. Применение машинного обучения для анализа образовательных результатов студентов вузов // Информационные и математические технологии в науке и управлении. 2023. № 4 (32). С. 67–78.

4. Харитонов И.М., Крушель Е.Г., Привалов О.О., Степанченко И.В., Степанченко О.В. Прогнозирование качества обучения в вузе с помощью методов регрессионного анализа // Известия Санкт-Петербургского государственного технологического института (технического университета). 2021. № 56. С. 72–80.

5. Ерохина Е.А., Хруслова Д.В. Влияние результатов ЕГЭ на успеваемость студентов ВУЗ // Информационные технологии в науке, образовании и управлении: материалы XLIV международной конференции и XIV международной конференции молодых учёных IT + S&E`16 / под редакцией Е.Л. Глориозова (Гурзуф, 22 мая – 01 июня 2016 года). М.: ООО «Институт новых информационных технологий», 2016. С. 265–272. EDN: WFFQKL.

6. Щеголева Л.В., Суровцева Т.Г. Результаты ЕГЭ и успеваемость студентов первого курса // Непрерывное образование: XXI век: научный электронный журнал. 2015. Вып. 4 (12). С. 1–9. URL: http://elibrary.petrsu.ru/books/48238 (дата обращения: 17.07.2025). DOI: 10.15393/j5.art.2015.2946.

7. Мосин В.Г., Козловский В.Н. Повышение производительности регрессионных моделей при оценке качества потребления электронного контента // Известия ТулГУ. Технические науки. 2024. № 2. URL: https://cyberleninka.ru/article/n/povyshenie-proizvoditelnosti-regressionnyh-modeley-pri-otsenke-kachestva-potrebleniya-elektronnogo-kontenta (дата обращения: 17.07.2025).

8. Шуметов В.Г., Барбашова Е.В., Слатинов В.Б. Методические аспекты преобразования показателей в оптимизационных управленческих задачах региональной экономики // Среднерусский вестник общественных наук. 2016. № 6. URL: https://cyberleninka.ru/article/n/metodicheskie-aspekty-preobrazovaniya-pokazateley-v-optimizatsionnyh-upravlencheskih-zadachah-regionalnoy-ekonomiki (дата обращения: 17.07.2025).

9. Шергин С., Усманов Р.Т. Краткосрочное прогнозирование энергопотребления малых населенных пунктов Крайнего Севера // ОмГТУ. 2024. № 3. URL: https://cyberleninka.ru/article/n/kratkosrochnoe-prognozirovanie-energopotrebleniya-malyh-naselennyh-punktov-kraynego-severa (дата обращения: 17.07.2025).

10. Фартушнов Н.С. Библиотеки языка Python для машинного обучения, их возможности и преимущества // Теория и практика современной науки. 2020. № 5 (59). URL: https://cyberleninka.ru/article/n/biblioteki-yazyka-pythondlya-mashinnogo-obucheniya-ih-vozmozhnosti-i-preimuschestva (дата обращения: 17.07.2025).

11. Пьянкова С.Г., Ергунова О.Т. Трансформация базовых компетенций человеческих ресурсов в эпоху цифровизации // Ars Administrandi. 2025. № 2. URL: https://cyberleninka.ru/article/n/transformatsiya-bazovyh-kompetentsiy-chelovecheskih-resursov-v-epohu-tsifrovizatsii (дата обращения: 17.07.2025).

12. Ахикян А.И., Данилюк С.С. Алгоритм машинного обучения Адаптивный случайный лес и его применение // Вестник науки. 2024. № 6 (75). URL: https://cyberleninka.ru/article/n/algoritm-mashinnogo-obucheniya-adaptivnyy-sluchaynyy-les-i-ego-primenenie (дата обращения: 17.07.2025).

13. Бадыкова И.Р., Биктимирова К.Р. Выявление факторов воздействия на сектор связи и телекоммуникаций с применением ансамблевых методов машинного обучения // π-Economy. 2024. № 6. URL: https://cyberleninka.ru/article/n/vyyavlenie-faktorov-vozdeystviya-na-sektor-svyazi-i-telekommunikatsiy-s-primeneniem-ansamblevyh-metodov-mashinnogo-obucheniya (дата обращения: 17.07.2025).

14. Харитонов И.М., Огар Т.П., Щербин С.И., Степанченко И.В. Применение нейронных сетей для прогнозирования будущей успеваемости абитуриента в вузе // Вестник компьютерных и информационных технологий. 2022. Т. 19. № 12 (222). С. 38–45. DOI: 10.14489/vkit.2022.12.pp.038-045. EDN: TKURRQ.

15. Bravo-Agapito J., Romero S.J., Pamplona S. Early Prediction of Undergraduate Student's Academic Performance in Completely Online Learning: A Five-Year Study // Computers in Human Behavior. 2021. Vol. 115. P. 106595. DOI: 10.1016/j.chb.2020.106595. EDN: DUBPWP.