DOI https://doi.org/10.46845/2541-8254-2026-3(55)-75-75
УДК: 004.85

ПРОГНОЗИРОВАНИЕ АКАДЕМИЧЕСКИХ РИСКОВ СТУДЕНТОВ НА ОСНОВЕ ЦИФРОВОГО СЛЕДА С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

В.Д. Черпаков, студент,
e-mail: cherpakov_vlad@mail.ru
ФГБОУ ВО «Калининградский государственный технический университет»

Ю.В. Кашпоров, доц.,
e-mail: yurij.kashporov@klgtu.ru
ФГБОУ ВО «Калининградский государственный технический университет»

Одной из наиболее актуальных задач для современных высших учебных заведений является выявление студентов, находящихся в зоне академического риска. В данной работе была исследована возможность оперативно выявлять студентов, находящихся в группе риска на основе данных их цифрового следа в информационной системе ВУЗа. Исследование проведено с использованием методов логистической регрессии, случайного леса и хист градиентного бустинга. На основе перечисленных методов были построены классификационные модели, позволяющие предсказать успешное завершение обучения студентом. Анализ был проведен на обезличенных данных Калининградского государственного технического университета (КГТУ) на N = 4422 записях (после очистки датасетов). В результате проведенного исследования было установлено, что разработанные модели классификации способны оперативно выявлять проблемных студентов ВУЗа с высокой точностью ROC-AUC 0,94.

Ключевые слова: академические риски, цифровой след, машинное обучение, классификация, прогнозирование успеваемости.

ВВЕДЕНИЕ

В современной системе образования всё чаще возникает потребность в раннем выявлении студентов, находящихся в зоне академического риска. Традиционные системы оценивания характеризуются субъективностью, поздним выявлением проблемных студентов, сложностью обработки большого количества данных и т.д. [3]. В связи с этим возникает объективная потребность в разработке классификатора, который позволит на основе неявных паттернов цифрового следа оперативно выявить студентов, которые могут оказаться в группе риска [4].

Актуальность исследования обусловлена стремлением преодолеть существующие ограничения подходов к мониторингу успеваемости, поздним реагированием и сложностью обработки значительных массивов информации. Полученная классификация поможет преодолеть трудности адаптации новоприбывших студентов [2], что поможет им быстрее вливаться в учебную жизнь университета.

Дополнительным фактором исследования выступает высокая специфичность признаков академического риска. Использование большого объема обезличенных данных из информационной среды университета, дает возможность увидеть неявные паттерны учебного поведения студента и риски академической неуспеваемости в условиях конкретного университета [4].

ОБЪЕКТ ИССЛЕДОВАНИЯ

Объектом данного исследования является цифровой след обучающихся в информационной системе высшего учебного заведения. При получении образования студенты показывают различные результаты, в том числе и неудовлетворительные. В процессе их обучения у них накапливается большое количество разнородной информации, например: оценки, посещаемость, приказы (приказы об отчислении, выпуске, назначения на практику), документы об имеющемся образовании и другие данные, которые и формируют собой цифровой след обучающихся.

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Целью анализа является разработка классификатора прогнозирования итогов обучения на основе комплексного цифрового следа для выявления студентов с высоким риском академической неуспеваемости на основе анализа их цифрового следа в электронной образовательной среде вуза. Для достижения этой цели была проведена предварительная обработка и интеграция разнородных данных цифрового следа студентов Калининградского государственного технического университета, после чего выполнен разведочный анализ. С помощью корреляционного анализа были выделены те показатели, которые в наибольшей степени влияют на успеваемость. Затем реализованы и сопоставлены по качеству несколько вариантов методов бинарной классификации, а также оценена значимость признаков и описаны полученные результаты.

РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ ЦИФРОВОГО СЛЕДА

Для обоснованного построения прогностической модели и корректной интерпретации его результатов был выполнен разведочный анализ (EDA) доступных данных. Он позволил изучить структуру доступных данных, оценить разброс ключевых переменных, выявить потенциальные взаимосвязи между предикторами, а также определить направления дальнейшей предобработки и моделирования. Особое внимание уделялось признакам, которые согласно результатам предыдущих исследований наиболее тесно связаны с академической успешностью: демографическим характеристикам, уровню предшествующей подготовки и текущим учебным результатам [1, 2]. Разведочный анализ выполнен на полном массиве доступных записей (N = 14356) до фильтрации по целевой переменной.

Анализ возрастного распределения студентов рассматривается как фактор, отражающий неоднородность контингента по жизненному опыту, форме обучения и уровню мотивации. Распределение возраста в выборке варьируется от 18 до 77 лет при медиане 27 лет. Основная категория обучающихся (7000 человек) сосредоточена в интервале 22 – 30 лет (Рис. 1). Выявлена правосторонняя асимметрия, что обусловлено присутствием студентов заочной формы обучения и лиц, получающих второе высшее образование. Данная неоднородность указывает на необходимость включить возраст в набор числовых значений предикторов модели.

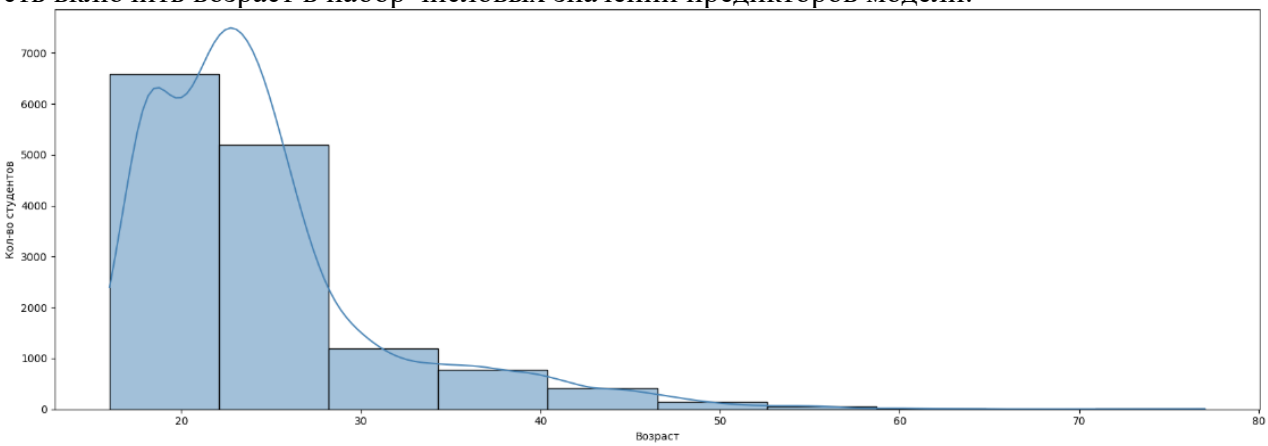

Рисунок 1 – Распределение возраста среди студентов

Анализ распределения студентов по уровню получаемого образования показало, что наиболее массовым видом является бакалавриат (8122 человек), за которым с заметным отрывом следует специалитет (4169 человек), а затем магистратура (1448 человек) и аспирантура (235 человек) (Рис. 2). Понимание данной структуры важно для проверки гипотез о том, что вероятность успешного завершения обучения может различаться в зависимости от ступени образования, а также для последующей стратификации выборки.

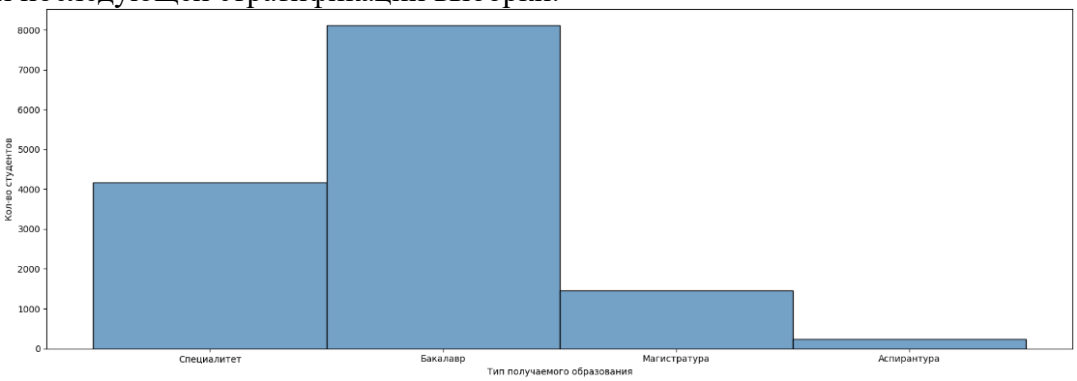

Рисунок 2 – Распределение получаемого образования среди студентов

Анализ академической успеваемости выявил, что наиболее частыми являются отметки «Зачтено» (130435 отметок) и «Неявка» (95466 отметок), тогда как отметки «Удовлетворительно» (71819 отметок), «Хорошо» (72963 отметок) и «Отлично» (68212 отметок) встречаются примерно с одинаковой частотой (Рис. 3). Данное наблюдение указывает на то, что сам факт наличия записи о неявке на экзамен может служить сигналом о потенциальных академических трудностях студента.

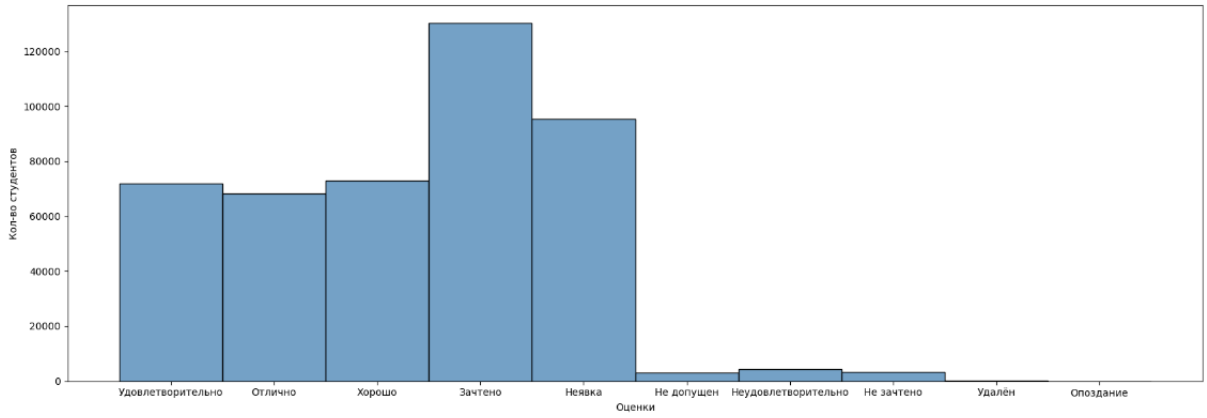

Рисунок 3 – Распределение полученных оценок среди студентов

Анализ имеющегося образования показал, что доминирует среднее общее образование (5748 человек), что отражает типичную траекторию поступления в университет после школы (Рис. 4). Несмотря на то, что последующий корреляционный анализ не выявил сильной линейной связи этого предиктора с целевой переменной, его распределение учитывается при формировании признакового пространства.

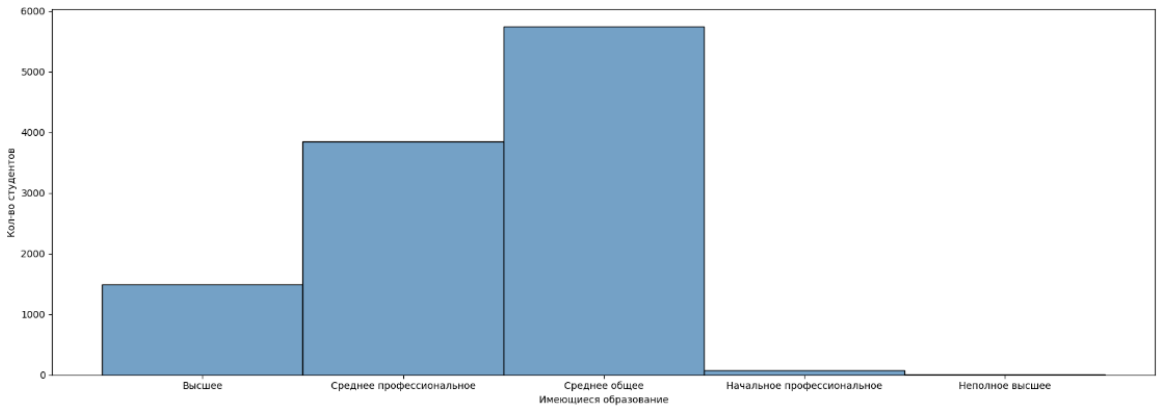

Рисунок 4 – Распределение имеющегося образования среди студентов

Ввиду особенностей данных при построении матрицы корреляций числовых признаков был использован коэффициент Спирмена, так как он превращает значения в ранги. Для непрерывных переменных будут присвоены ранги исходных чисел, для порядковых – порядковые ранги, а для бинарной переменной (целевой переменной) будут присвоены ранги 0 и 1. Результат построения представлен на рисунке номер 5. По нему можно сделать вывод, что наибольшую значимость для целевой переменной представляют возраст и средний балл. Кроме того, наибольшее влияние на целевую переменную имеет средний балл или же GPA (r = 0.756) меньшее влияние оказывает возраст (r = 0.127), что пересекается с исследованиями других авторов [3, 9].

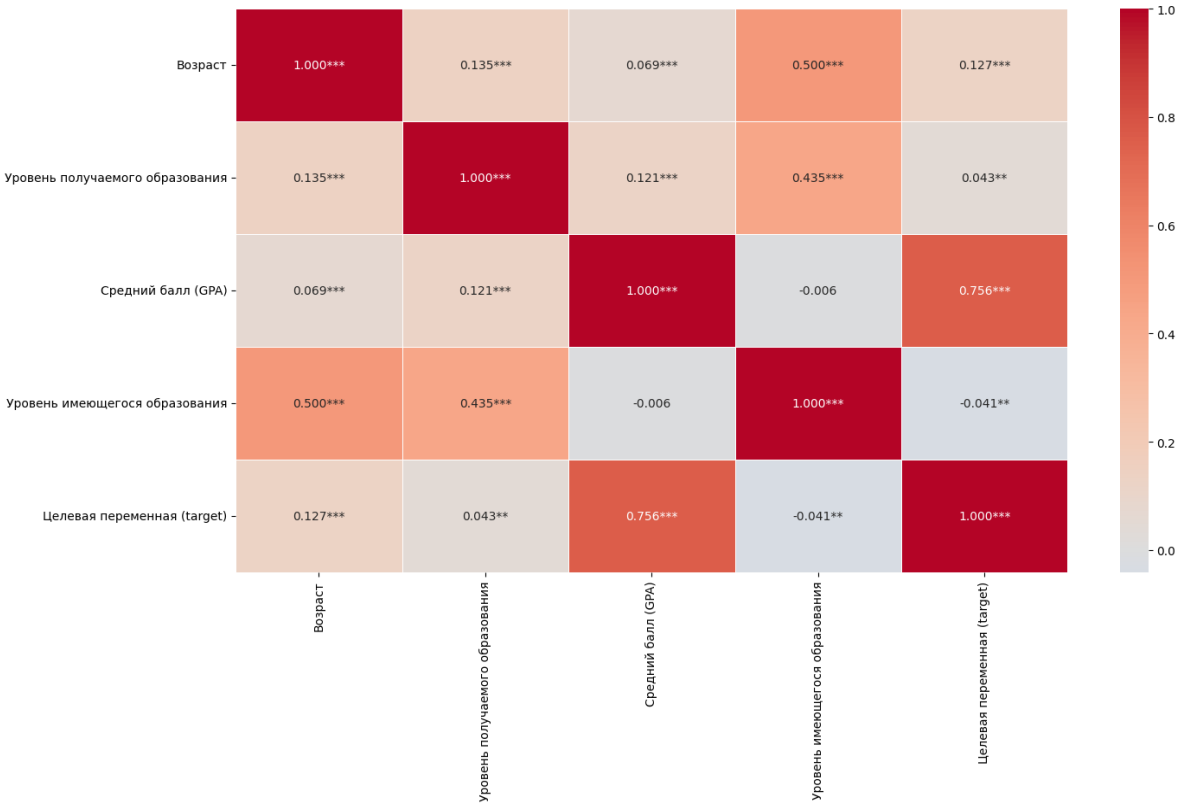

Рисунок 5 – Матрица корреляции числовых признаков и целевой переменной

МЕТОДЫ ИССЛЕДОВАНИЯ

Настоящее исследование базируется на анализе обезличенных данных из информационной среды Калининградского государственного технического университета за период с 2023 по 2025 год. Исходный объем данных (N = 14356) был представлен в виде таблиц, содержащих 24 уникальных признака о студентах, включая их демографические и социальные признаки, данные

о предыдущем образовании, академическую успеваемость. Качество данных находилось на низком уровне, из-за чего требовалась обработка датасетов перед работой с ними.

На первом этапе выполнена комплексная очистка датасетов от недостоверных данных. Из каждого набора данных удалены дубликаты записей, строки, содержащие ошибки. Строки с пропущенными значениями заполнены значением «Не указана». Записи, для которых не удалось восстановить критически важные поля, были исключены из дальнейшего анализа.

Буквенные обозначения успеваемости студентов были переведены в числовые эквиваленты, а единичная запись «Опоздание» удалена как единичный случай. На основе полученных данных для каждого студента был рассчитан средний балл (GPA). Уровень получаемого и имеющегося образования закодированы порядковыми числами в соответствии с их иерархией.

Целевая переменная сформирована на основе последнего приказа студента (хронологически): значение «1» присвоено выпускникам, «0» – студентам, отчисленным по любому основанию (включая отчисление из аспирантуры, а также приёмной комиссии). Причины отчисления такие как академическая неуспеваемость или отчисление по собственному желанию не учитывались, так как по результатам анализа, студентов, отчислившихся по собственному желанию слишком мало (всего 295 записей или 6,6%). Кроме того, студенты, которые отчислились по собственному желанию имеют большое количество задолженностей и неявок и в будущем были бы отчислены за академическую неуспеваемость. Студенты, по которым отсутствовал итоговый приказ, исключены из анализа ввиду невозможности правильного заполнения данного признака. Студенты, которые перевелись в другой ВУЗ или на другую специальность не учитывались, так как перевод имеет другую суть и не может быть приравнен к отчислению или выпуску студента. Кроме того, важно упомянуть, что из-за малого количества студентов, обучающихся заочно (менее 5%), было принято решение не выделять их в отдельную группу.

После объединения таблиц и удаления записей с пропущенными значениями итоговая выборка составила N = 4422 уникальных наблюдений. Распределение классов оказалось умеренно сбалансированным: 2677 выпускников (60,5%) и 1745 отчисленных (39,4%). Это умеренный перекос, при котором большинство алгоритмов (логистическая регрессия, случайный лес и градиентный бустинг) способны хорошо обучаться и без балансировки, что подтверждает и результаты [5–7].

Для моделирования использовались следующие признаки: категориальные: пол, страна проживания, город проживания, основа обучения, факультет; числовые: возраст, уровень получаемого образования, GPA, уровень имеющегося образования. Выборка разделена хронологически (Time-Based Split), при котором более старые записи идут на обучение (40 %, 1768 записей), а будущие на валидационную (30 %, 1326 записей) и тестовую (30 %, 1328 записей) с сохранением пропорций классов (стратификация).

Для решения задачи бинарной классификации студентов по признаку завершения обучения были выбраны три алгоритма: логистическая регрессия [5], случайный лес (Random Forest) [6] и хист градиентный бустинг (HistGradientBoosting) [7]. Выбор данных алгоритмов обусловлен необходимостью сравнения линейных и нелинейных подходов.

Все модели реализованы на языке Python с применением библиотеки scikit-learn [8]. Предобработка данных, включая заполнение пропусков и кодирование категориальных признаков, организована в рамках единого пайплайна.

Для числовых признаков применяется заполнение пропусков медианой, для категориальных – заполнение модой с последующим порядковым кодированием (OrdinalEncoder). Все преобразования, включая подбор параметров заполнения и обучение кодировщика, выполняются исключительно на обучающей выборке и лишь затем применяются к валидационной и тестовой, что полностью исключает утечку данных.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Результатом проведенного исследования являются 3 модели классификации, обученные на данных, которые были получены из цифрового следа ВУЗа, и оценки их качества. Оценка ка-

чества моделей проводилась по следующим метрикам: точность (Accuracy), точность предсказания (Precision), полнота (Recall), среднее гармоническое между точностью предсказания и полнотой (F1-мера) и показатель того, насколько хорошо модель различает классы (ROC-AUC). Основной метрикой выбрана ROC-AUC, поскольку она устойчива к умеренному дисбалансу классов. Результаты тестирования обученных моделей на валидационной и тестовой выборках представлены в таблице 1 и на рисунке 6.

Таблица – Метрики качества моделей на валидационной выборке

Модель	Точность	Полнота	Точность предсказания	F1	ROC-AUC
Логистическая регрессия (Logistic Regression)	0.886	0.979	0.864	0.918	0.941
Случайный лес (Random Forest)	0.9	0.969	0.886	0.926	0.94
Хист Градиентный бустинг (Hist Gradient Boosting)	0.88	0.954	0.877	0.914	0.936

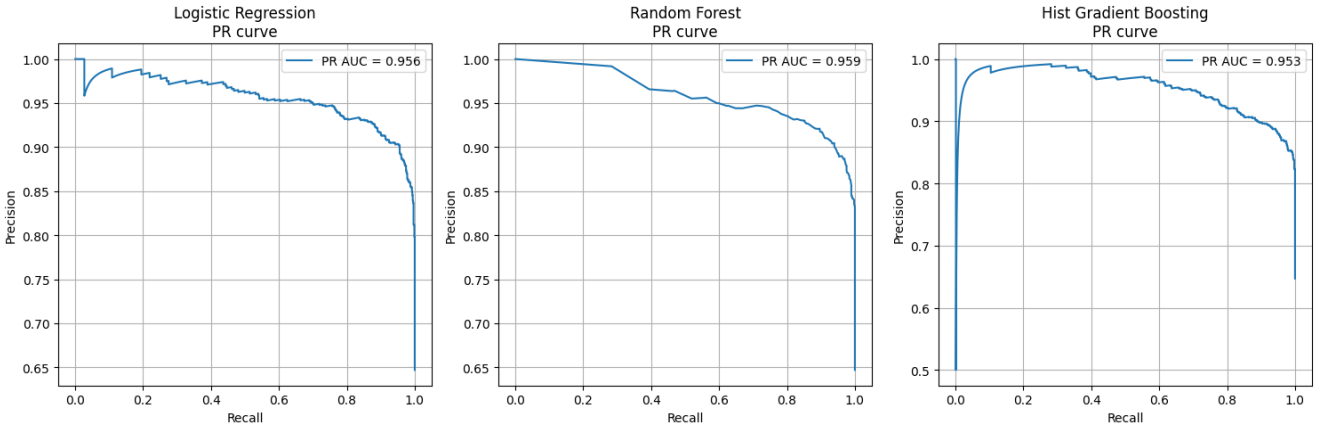

Рисунок 6 – Сводка по валидационной выборке

Анализ качества построенных моделей показывает, что все три модели склонны несколько чаще ошибаться в сторону ложного предсказания выпуска (False Positive), чем пропускать реальных выпускников (False Negative), о чем свидетельствуют матрицы ошибок (Рис. 7). С практической точки зрения это означает, что система будет излишне оптимистична в отношении небольшой доли студентов, однако подавляющее большинство рисков будет выявлено.

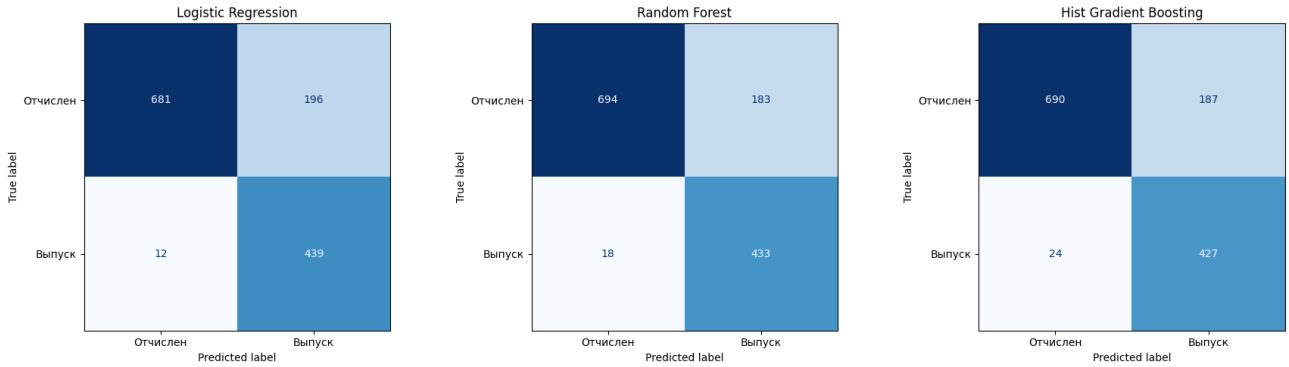

Рисунок 7 – Матрица ошибок

Для выявления факторов, вносящих наибольший вклад в прогноз, был проведен анализ важности признаков с помощью метода «feature_importances_» доступного в библиотеке scikit-learn (Рис. 8). В качестве основной модели была выбрана модель случайного леса (Random Forest), так как она предоставляет наиболее высокие (в совокупности) метрики качества. Установлено, что средний балл успеваемости или же GPA является доминирующим предиктором для данной модели (63%), что полностью соответствует результатам корреляционного анализа и данным литературы [3,9]. Вторым по значимости признаком стал возраст обучающегося (16.8%). Данный результат объясняется неоднородностью контингента по форме обучения, жизненному опыту и уровню мотивации. Наименьший вклад в прогноз внесли социально-демографические характеристики: страна проживания, факультет, пол (вклад менее 10% для каждого отдельного признака).

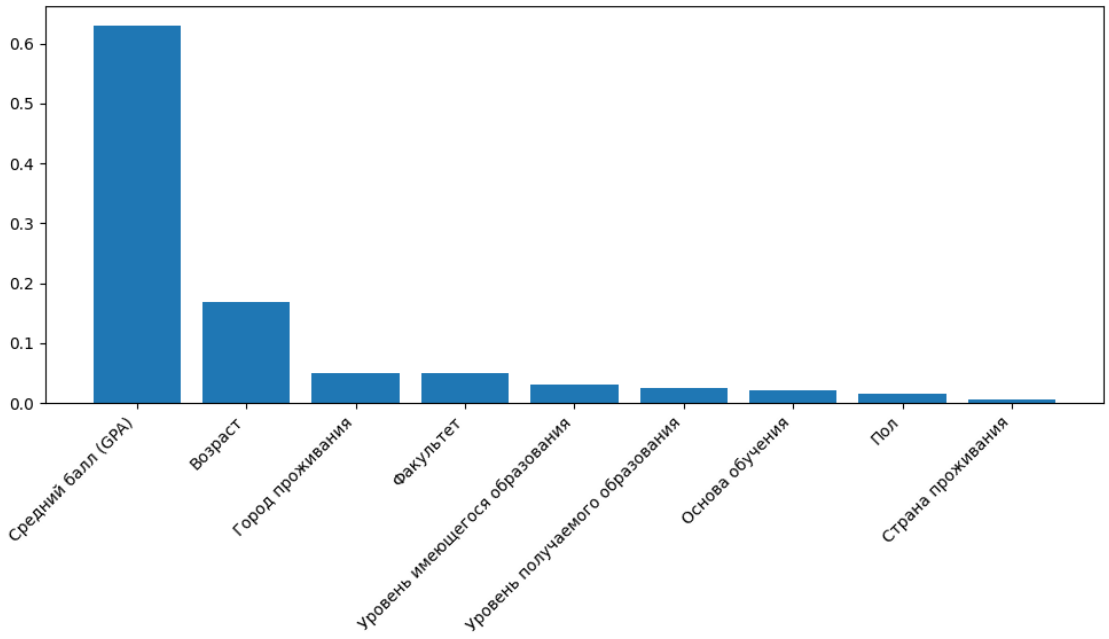

Рисунок 8 – Гистограмма важности признаков

ЗАКЛЮЧЕНИЕ

В ходе выполнения данного исследования был разработан прототип классификатора оперативного прогнозирования академических рисков на основе цифрового следа студентов. Полученные результаты подтверждают, что цифровой след, формируемый в информационной среде вуза, обладает значимым предсказательным потенциалом в отношении итогового статуса студента (выпуск или отчисление) [4]. Доминирующим предиктором выступает текущая академиче-

ская успеваемость (GPA) что подтверждается как собственными расчётами, так и выводами других авторов [3,9]; дополнительное влияние оказывает возраст, тогда как социально-демографические факторы имеют второстепенное значение. Наилучшее качество классификации на тестовой выборке по совокупности метрик выборке продемонстрировала модель случайного леса (Random Forest) ее метрика ROC-AUC составила 0,94, однако и более простая логистическая регрессия показывает приемлемые метрики.

Направлением для дальнейших исследований может быть включение динамических переменных, отображающих временные тренды образовательной активности: изменение среднего балла в течение семестра, количество и динамику академических пересдач, регулярность и интенсивность работы в электронных курсах. Для дополнительного повышения точности прогнозов могут быть использованы более сложные ансамблевые архитектуры и оптимизация гиперпараметров.

СПИСОК ЛИТЕРАТУРЫ

1. *Мальков И.А.* Система прогнозирования академической задолженности на основе текущей успеваемости и процента пропусков // Международный научный журнал «Вестник науки». – 2025. – № 6 (87). – Т. 1. – С. 1495–1505.
2. *Булычева П.А., Ошмарина О.Е., Шадрина Е.В.* Выявление академически неуспешных студентов на первом году обучения в университете на примере НИУ ВШЭ – Нижний Новгород // Вопросы образования. – 2016. – № 4. – С. 141–162.
3. *Ильин А.Р., Хусаинов Р.М.* Интеллектуальная система оценки рисков академической неуспеваемости на основе машинного обучения // Известия Балтийской государственной академии рыбопромыслового флота. – 2026. – № 1(75). – С. 129–136.
4. *Попова Н.А., Егорова Е.С.* Интеллектуальный анализ образовательных данных для прогноза успеваемости студентов вуза // Известия Кабардино-Балкарского научного центра РАН. – 2023. – № 2(112). – С. 18–29.
5. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. 2nd ed. – Springer, 2009.
6. *Breiman L.* Random Forests // Machine Learning. – 2001. – Vol. 45, № 1. – P. 5–32.
7. *Friedman J.H.* Greedy Function Approximation: A Gradient Boosting Machine // Annals of Statistics. – 2001. – Vol. 29, № 5. – P. 1189–1232.
8. *Pedregosa F. et al.* Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.
9. *Огий О.Г., Кметь А.Д.* Анализ взаимосвязи параметров «на входе» и академической результативности студентов // Известия Балтийской государственной академии рыбопромыслового флота. – 2026. – № 1(75). – С. 13–24.
10. *Кочергина Е.В., Прахов И.А.* Взаимосвязь между отношением к риску, успеваемостью студентов и вероятностью отчисления из вуза // Вопросы образования. – 2016. – № 4. – С. 206–228.

PREDICTING STUDENTS' ACADEMIC RISKS BASED ON DIGITAL TRACKS USING MACHINE LEARNING METHODS

V.D. Cherpakov, Student,
email: cherpakov_vlad@mail.ru
Kaliningrad State Technical University

Y.V. Kashporov Associate Professor,
email: yurij.kashporov@klgtu.ru
Kaliningrad State Technical University

One of the most pressing challenges for modern universities is the early identification of students at academic risk. This study investigates the possibility of predicting student dropout based on digital footprint data from the university information system. The research employs logistic regression, random forest, and XGBoost gradient boosting methods. Using these techniques, classification models were built to predict successful student graduation. The analysis was conducted on anonymized data from the Kaliningrad State Technical University (KSTU), comprising N = 4422 records after dataset cleaning. The results show that the developed classification models can predict university student dropout with high accuracy, achieving a ROC-AUC of 0.94.

Keywords*: academic risks, digital footprint, machine learning, classification, and academic performance prediction.*