УДК 004.89

АНАЛИЗ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ МЕТОДОВ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ОБРАЗОВАТЕЛЬНОЙ АНАЛИТИКЕ

Д. А. Минуллин[1] [0000-0001-7713-5251], Ф. М. Гафаров[2] [0000-0003-4704-154X]

[1, 2]*Институт вычислительной математики и информационных технологий, Казанский (Приволжский) федеральный университет*

[1]minullin.dima@mail.ru, [2]fgafarov@yandex.ru

Аннотация

Проблема прогнозирования досрочного отчисления студентов российских вузов является актуальной, поэтому требуется разработка новых инновационных подходов для её решения. Для решения данной проблемы возможна разработка предиктивных систем на основе использования данных о студентах, имеющихся в информационных системах вузов. В настоящей работе исследованы модели машинного обучения для прогнозирования досрочного отчисления студентов, обученные на основе данных о характеристиках и успеваемости студентов. Основная научная новизна работы заключается в использовании методов объяснимого ИИ для интерпретации и объяснения функционирования обученных моделей машинного обучения. Методы объяснимого искусственного интеллекта позволяют понять, какие из входных признаков (характеристик студента) оказывают наибольшее влияние на результаты прогнозов обученных моделей, а также могут помочь понять, почему модели принимают те или иные решения. Полученные результаты расширяют понимание влияния различных факторов на досрочное отчисление студентов.

Ключевые слова*: образовательная аналитика, интеллектуальный анализ данных, машинное обучение, объяснимый искусственный интеллект*

ВВЕДЕНИЕ

Одной из ключевых проблем современного образования является досрочное отчисление из высших учебных заведений студентов по неуспеваемости. Дан-

ная проблема может привести не только к личностным трудностям самих студентов, но и повлечь за собой социальные проблемы, связанные с утратой образовательных ресурсов общества [1]. Досрочное отчисление студентов из-за проблем с успеваемостью сильно влияет на всю систему образования, т. к. приводит не только к уменьшению количества студентов в учебных группах, но и может повлечь за собой личные образовательные последствия, а также проблемы для самого университета, такие, например, как снижение рейтинга и увеличение затрат [2, 3]. Таким образом, необходимы стратегии и меры, направленные на решение данной проблемы и улучшение качества образования, с целью снижения количества отчислений студентов. Исследования названной проблемы основаны не только на данных об успеваемости, но и на широком спектре факторов, которые могут отказать влияние на решение студентов о продолжении обучения. В этих исследованиях анализируются такие аспекты, как мотивация, уровень интереса к обучению, успешная интеграция в университетскую среду, адаптация к новым условиям, удовлетворенность образовательным процессом. Цель исследователей – определить влияние этих факторов на решение студентов продолжить обучение. На основе полученных данных разрабатываются различные рекомендации для университетов по снижению количества отчисленных студентов и предлагаются меры по повышению качества образования [4–7].

В связи со сказанным разработка инновационных подходов к прогнозированию академической успешности студентов и выявлению студентов, которые могут быть отчислены, становится особенно актуальной. Одним из перспективных подходов здесь является принятие решений на основе данных. Использование методов машинного обучения для предсказания вероятности преждевременного отчисления студентов представляет собой современную стратегию [8]. Эти методы могут дать возможность оперативно выявлять студентов, находящихся в группе риска, и предотвращать их отчисление. Точные прогнозы, основанные на данных, позволят учебным заведениям предпринять шаги по оказанию помощи и адаптации, ориентированные на улучшение успеваемости студентов и повышение их удовлетворенности образовательным процессом.

Модели машинного обучения способны эффективно распознавать сложные закономерности и делать прогнозы на основе данных. В последнее время растет

интерес к интерпретации таких моделей как со стороны научного сообщества [9–11], так и со стороны пользователей различных предиктивных систем. Сфера интерпретируемого машинного обучения, также известная как «объяснимый ИИ» (Explainable AI), активно развивается, и в этой области появляется все больше исследований [12–17]. Такие подходы уже успешно применяется в различных сферах, включая медицину, энергетику и науку [18–21].

Для анализа моделей машинного обучения имеются различные инструменты и фреймворки, помогающие раскрыть внутреннюю работу моделей, делая их результаты более понятными (например, Captum [22, 23] и SHAPexplainer [24]). Использование этих инструментов позволяет визуализировать результаты работы алгоритмов интерпретации моделей, чтобы выяснить, как различные значения входных данных влияют на прогнозы моделей [25–28].

В настоящей работе показано использование методов Integrated Gradients, DeepLIFT, GradientSHAP и SHAP для анализа результатов применения моделей, обученных на решение задачи прогнозирования досрочного отчисления студентов. Для обучения моделей машинного обучения были использованы данные по выпускникам и досрочно выбывшим студентам Казанского (Приволжского) федерального университета в период с 2012 по 2019 годы. Для определения потенциальной группы риска среди обучающихся нами была использована модель прогнозирования бинарной классификации; выявлены факторы, которые больше всего способствуют досрочному отчислению студентов.

ДАННЫЕ И МЕТОДЫ

Был использован набор данных, представленный в статье [29]. Этот набор состоит из записей о выпускниках с полными данными за весь четырёхлетний цикл обучения, и досрочно выбывших студентах Казанского (Приволжского) федерального университета. Для обучения моделей был собран максимально полный и подробный набор данных, который отражает процесс обучения студентов в вузе, чтобы определить факторы, способствовавшие досрочному выбытию студента. Данные студента характеризуются набором параметров двух типов: числовые и категориальные. К числовым параметрам относятся баллы ЕГЭ, средний балл ЕГЭ, общий средний балл успеваемости за все полные семестры, средний балл за первый и второй семестры. К категориальным параметрам относятся пол

студента, форма обучения (очная\заочная), оплата обучения (бюджет\контракт), тип предыдущего образования, год поступления и специализация студента. Значения всех категориальных параметров были закодированы методом one-hot encoding в соответствии с [29].

Для построения прогностической системы прогнозирования досрочного выбытия студентов были обучены и проанализированы модели двух типов (таблица 1). Первый тип моделей основан на искусственной нейронной сети (ИНС), состоящей из трёх полносвязных слоёв, второй тип – на алгоритме градиентного бустинга (XGBoost). Модели обучались на решениях задачи бинарной классификации, на выходе они давали прогноз того, что студент успешно завершит обучение (выход – 0) или же он будет досрочно отчислен (выход – 1).

Таблица 1. Описание используемых подходов и методов.

Подход\ Метод	**Описание**	**Преимущества**	**Недостатки**
Искусственные нейронные сети	Модели, имитирующие работу человеческого мозга, обучаются на основе выборки данных для предсказания результатов	Могут обрабатывать очень большое количество данных, выявлять сложные нелинейные зависимости, адаптивны и способны к самообучению.	Сложны в интерпретации, требуют значительных вычислительных мощностей и большого количества обучающих данных, подвержены переобучению.
XGBoost	Алгоритм градиентного бустинга, использующий в качестве базовых моделей деревья решений.	Относительно быстрый, эффективный, расширяемый, поддерживает разные типы данных, автоматиче-	Может переобучиться при некорректно подобранных параметрах, требует тщательной

		ски обрабатывает пропущенные значения, регулирует сложность модели.	настройки гиперпараметров.
Integrated Gradients	Метод интерпретации моделей машинного обучения, основанный на аппроксимации градиентов входных данных вдоль базовой линии.	Позволяет увидеть, какие признаки влияют на прогноз модели, позволяет интерпретировать сложные модели, может быть применён к моделям любого размера.	Подвержен проблеме «sparse gradients», вычислительно затратен для больших моделей
SHAP	Метод интерпретации предсказаний моделей машинного обучения, основанный на теории игр, позволяет объяснить влияние каждого признака на прогноз модели на основе их индивидуального вклада.	Объективен, может оценить важность признаков как для всей модели, так и для отдельных предсказаний, позволяет интерпретировать сложные модели.	Требует значительных вычислительных ресурсов для сложных моделей и требовать больше времени для вычислений, может быть сложен в понимании.
DeepLIFT	Метод интерпретации нейронных сетей, сравнивает активацию нейронов и присваивает им оценки.	Позволяет интерпретировать сложные нейронные сети, в частности глубокие сети, пытается решить проблему «sparse gradients», может быть применен к любым архитектурам нейронных сетей.	В некоторых случаях может быть неэффективным для моделей с большим количеством параметров

Gradient-SHAP	Комбинирует методы SHAP и Integrated Gradients для интерпретации моделей, предлагает объяснения на основе важности признаков.	Объединяет преимущества обоих методов, использует градиенты для ускорения SHAP, более точен, позволяет интерпретировать сложные модели.	Требует больше вычислительных ресурсов, чем SHAP или Integrated Gradients, может быть сложным в реализации.

Основным направлением настоящего исследования является применение методов интерпретации моделей машинного обучения, таких как Integrated Gradients, DeepLIFT, GradientSHAP и SHAP. Эти методы позволяют заглянуть внутрь «черного ящика» на основе графической визуализации влияния значений входных параметров на результаты прогноза обученных моделей. Методы Integrated Gradients, DeepLIFT и GradientSHAP относятся к типу градиентных. Эти методы присваивают важность каждому входному признаку, анализируя, как его изменения влияют на выход модели, используя градиентное разложение для количественной оценки этих эффектов. Для моделей XGBoost был применён метод атрибуции SHAP. Различные методы могут по-разному оценивать значимость признаков, и поэтому для получения максимально достоверного результата необходимо одновременно комбинировать разные методы. Эти методы позволяют выявить признаки, оказывающие наиболее существенное влияние на прогнозы модели, а также вводить числовые характеристики для количественной оценки их важности, что позволяет провести сравнение важности признаков.

Программный модуль анализа и интерпретация данных на основе интегрированных градиентов, DeepLIFT и GradientSHAP был разработан с использованием библиотеки Captum, а метод SHAP был реализован на основе одноимённой библиотеки (SHAP) для языка программирования Python. В библиотеку Captum встроены различные методы интерпретации, которые помогают объяснить модели глубокого обучения, разработанные с помощью фреймворка PyTorch. Для реализации модуля на основе SHAP был использован модуль KernelExplainer библиотеки Python SHAP, который работает универсально для любой модели прогнозирова-

ния. Этот модуль вычисляет прогностическую ценность признаков на основе значений SHAP. На вход модуля интерпретации подаются обученные модели, а также данные из тестового набора данных. На выходе модуля получаются атрибуции для каждого входного фактора, показывающие, насколько высока прогностическая сила каждого фактора.

РЕЗУЛЬТАТЫ

Был проведен сравнительный анализ интерпретации моделей нейронных сетей, обученных на полном наборе данных (40883 записей) и на данных в различных разрезах: по полу, специализации, предыдущему образованию, типу обучения. Средняя точность нейросетевой модели, обученной на полном наборе данных, составила 87%, средняя точность моделей, обученных на различных разрезах данных, – 84%. Для проверки точности моделей использовались данные из тестового набора данных, который составлял 30% от исходного.

Методы объяснимого ИИ позволяют получить числовые значения, показывающие степень влияния каждого входного признака на прогноз модели. Такие оценки способствуют пониманию, насколько похожи или отличаются значения атрибуций входных факторов, полученных на основе различных методов интерпретации (Integrated Gradients, DeepLIFT и GradientSHAP, SHAP). Величины весов, полученные в ходе применения методов интерпретации, могут иметь два направления – положительное и отрицательное. Эти направления несут информацию о влиянии конкретных входных параметров модели на выходное значение модели. Важность входных признаков, имеющих отрицательные и положительные числа, заключается в следующем: отрицательное значение показывает негативное влияние на вероятность досрочного отчисления студента, а положительное – увеличивает вероятность досрочного отчисления студента. Важность признака показывает, насколько он важен для классификации (рис. 1).

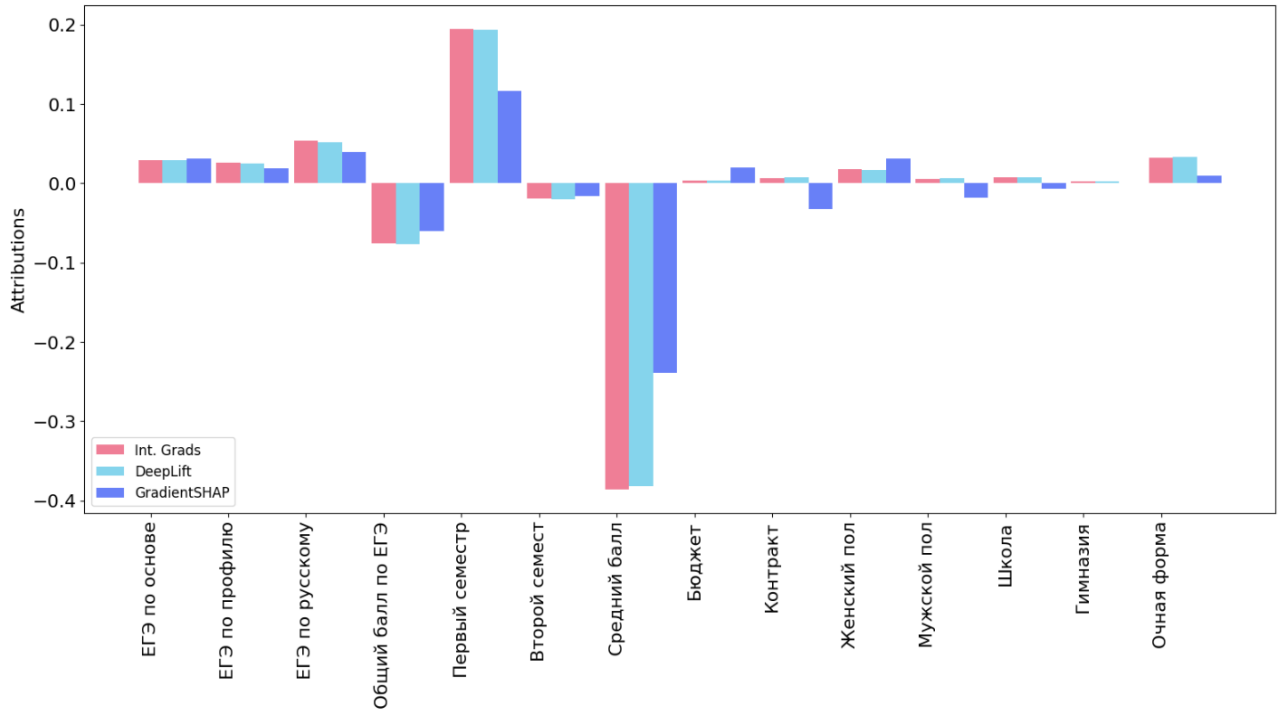

Рис. 1. Параметры атрибуции факторов для на основе методов Integrated Gradients, DeepLIFT и GradientSHAP

Сравнивая значения, полученные в ходе интерпретации нейронных сетей, обученных отдельно для студентов мужского и женского пола (Таблица 2), можно заметить, что в случае женского пола большее влияние на прогноз модели оказывают баллы ЕГЭ и оценки за второй семестр, в то время как у студентов мужского пола влияние баллов ЕГЭ ниже, а оценки за второй семестр имеют даже негативное влияние. Также, в отличие от случая женского пола, у студентов мужского пола начинает приобретать значимость признак – очная форма обучения. Остальные параметров имеют приблизительно одинаковы значения для обоих моделей.

Таблица 2. Значения интегрированных градиентов для нейронных сетей, обученных отдельно для студентов мужского и женского пола.

	Мужской	Женский
ЕГЭ по основе	0,02420	0,06589
ЕГЭ по профилю	0,03204	0,05651
ЕГЭ по русскому	0,03634	0,08342

Первый семестр	0,18182	0,13563
Второй семестр	–0,0099	0,00785
Очная форма	0,03290	0,00346

Также имеются явные различия в значениях интегрированных градиентов для нейронных сетей, обученных независимо для каждого из исследуемых типов обучения (таблица 3). Эти отличия наблюдались между всеми видами обучения, исключения составили такие параметры, как оценки за первый и второй семестры, которые всегда оказывали положительное влияние на прогнозы моделей, и факторы, значения интегрированных градиентов которых близки к нулю. Анализ моделей, обученных на данных студентов очной формы обучения, демонстрирует негативное влияние баллов ЕГЭ, положительное влияние признака пола студента, причём значение «женский пол» оказывает большее влияние на прогнозы модели, чем значение «мужской пол». Также значительное положительное влияние на результаты прогноза модели оказывает принадлежность к бюджетной форме обучения. Для студентов, обучающихся на очно-заочной форме, характерно отрицательное влияние баллов ЕГЭ по профильной дисциплине и русскому языку, и положительное – по основной учебной дисциплине. В отличие от очной формы обучения, для заочной формы обучения присуще отрицательное влияние фактора «женский пол» на прогнозы модели. Для студентов заочной формы обучения характерно положительное влияние баллов ЕГЭ и пола студента, в то время как принадлежность к бюджетной форме обучения имеет отрицательное влияние на прогнозы модели.

Таблица 3. Значения интегрированных градиентов для нейронных сетей, обученных отдельно для разных типов обучения.

	Очная	**Очно-заочная**	**Заочная**
ЕГЭ по основе	–0,05182	0,15287	0,02765
ЕГЭ по профилю	–0,06643	–0,1274	0,01641
ЕГЭ по русскому	–0,04512	–0,0948	0,00523

Женский пол	0,02742	–0,00125	0,02123
Мужской пол	0,00562	0,00762	0,00673
Бюджет	0,02456	0,002542	–0,00512
Контракт	0,00812	0,00856	0.03144

Основное различие между значениями интерпретаций форм обучения (бюджет/контракт) (таблица 4) проявляется в отношении влияния оценок за второй семестр: у бюджетной формы обучения наблюдается отрицательное влияние на прогноз модели, при контрактной – положительное. Важно отметить, что для студентов контрактной формы обучения более значимое влияние на прогнозы модели оказывают баллы ЕГЭ по профильной дисциплине, чем для студентов бюджетной формой обучения. Также можно выделить такие факторы, как «пол студента», «обучение на очной основе» и «женский пол», которые оказывает существенное влияние на прогнозы модели. Влияние же остальных признаков несущественно.

Таблица 4. Значения интегрированных градиентов для нейронных сетей, обученных отдельно для бюджетных и контрактных студентов.

	Бюджет	**Контракт**
ЕГЭ по профилю	0,017248	0,049231
Второй семестр	–0,00909	0,010561
Женский пол	0,012281	0,018576
Мужской пол	0,002932	0,001745
Очное обучение	0,01106	0,024567

Интерпретация моделей, обученных на данных, принадлежащих студентам, разделенным на основании вида их предыдущего образования (таблица 5), показала, что основными факторами, положительно влияющими на прогнозы моделей, являются следующие параметры: баллы по ЕГЭ, оценки за первый и второй семестры, пол студента, очное обучение, бюджетная и контрактная формы обуче-

ния. Во всех случаях, за исключением принадлежности предыдущего образования к «Лицею», «Вузу» и «Колледжу», можно наблюдать положительное влияние фактора «оценки за второй семестр» на прогнозы модели. Также для характеристики «Вуза» характерно негативное влияние на прогноз модели показателей ЕГЭ по основной и по профильным дисциплинам.

Таблица 5. Значения интегрированных градиентов для нейронных сетей, обученных отдельно по типу предыдущего образования.

	Школа	**Гимназия**	**Лицей**	**Вуз**	**Колледж**	**Техникум**
ЕГЭ по основе	0,07532	0,06231	0,05121	–0,01246	0,02541	0,02963
ЕГЭ по профилю	0,08031	0,06912	0,04592	–0,07125	0,00341	0,09863
Второй семестр	0,01834	0,02657	–0,06122	–1,14231	–0,0185	0,06187

Далее было проведено исследование возможностей метода SHAP для интерпретации моделей, обученных на основе градиентного бустнга (XGBoost). На рис. 2 показана диаграмма “waterfall”, которая позволяет визуализировать индивидуальные SHAP-значения, вычисленные для отдельного объекта из набора данных. Все параметры, включённые в набор данных, расположены на оси «*y*» и ранжированы сверху вниз на основе их значений-SHAP для данного конкретного прогноза. Фактическое значение объекта также отображается на оси «*y*». Влияние каждого параметра на прогноз представлено синей или красной стрелками и соответствующим значением-SHAP, которое отражает влияние этого признака на прогноз модели. Каждый прогноз начинает свою точку отсчёта с некоторого «базового значения» и влияние каждого признака смещает прогноз.

Рис. 2. Диаграмма значений SHAP для студента, успешно окончившего обучение (а) и для досрочно отчисленного студента (б).

В библиотеке SHAP имеется также другой тип представления локальных объяснений в виде графика силы – «force plot». Графики силы для двух студентов представлены на рис. 3.

Рис. 3. График силы (force plot) значений SHAP для студента, успешно окончившего обучение (а) и для досрочно отчисленного студента (б).

Для сравнения на рисунках 2 и 3 представлены интерпретации моделей (значения SHAP) для двух возможных классов данных: для студента, окончившего обучение, и досрочно отчисленного студента. Для студента, успешно окончившего обучение, основными показателями, влияющими на продолжение обучения, являются: общая средняя успеваемость, оценки за первый семестр и баллы за ЕГЭ по русскому языку. Оценки за второй семестр и то, что его предыдущим образованием являлся колледж, в свою очередь, оказали наибольшее влияние в сторону досрочного выбытия студента и «двигали» прогноз модели в сторону возможного отчисления (см. рис. 2а и 3а). Для студента (см. рис. 2б и 3б), который

был досрочно отчислен основными признаками, повлиявшим на выбытие, оказались его средний балл (несмотря на его достаточно высокое среднее значение), а также год поступления (2018). Различия SHAP-значений указывают на то, что каждый из признаков по-особенному влияет на результат прогноза модели, что укрепляет доверие к процессу принятия управленческих решений на основе их прогнозов.

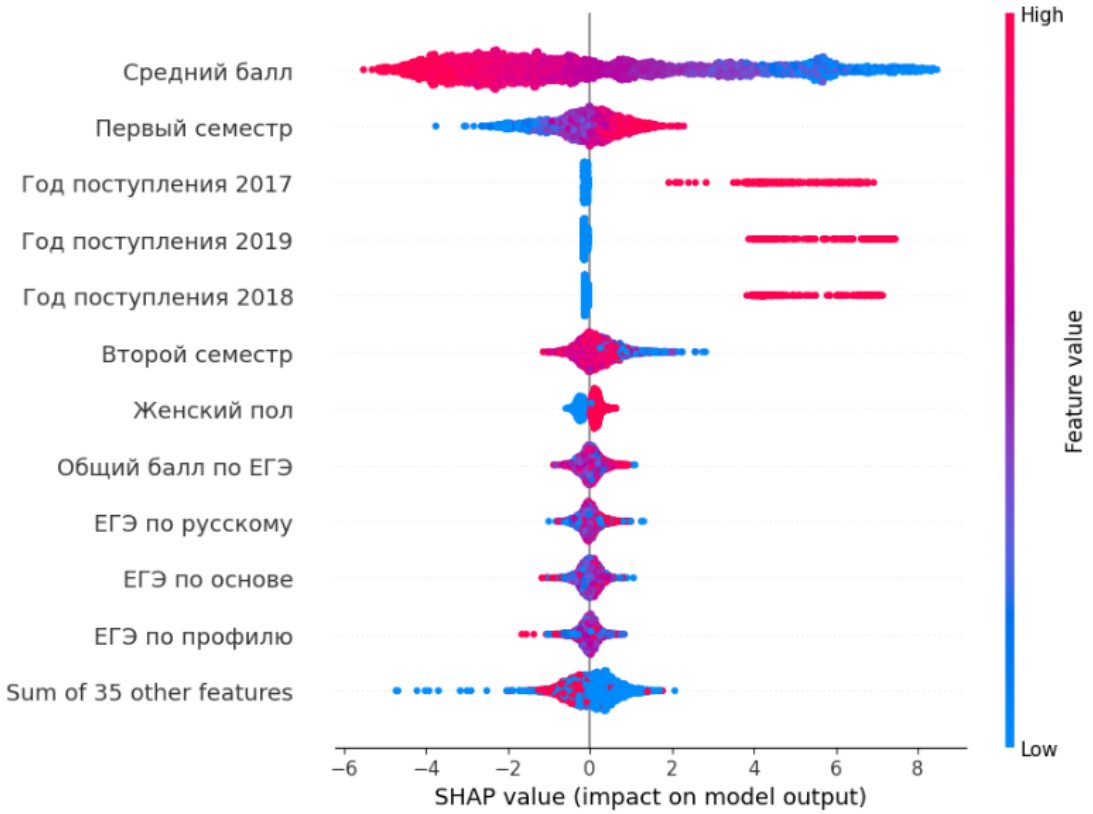

Рис. 4. Сводный график (beeswarm plot) SHAP для модели XGBoost

Метод SHAP предоставляет возможность не только построить графическую визуализацию влияния различных признаков на прогнозы модели для отдельных объектов данных, но и получить представление важности каждого фактора на основе полного набора данных (глобальные объяснения). На рис. 4 показано поведение модели на глобальном уровне на основе графика, который называется «beeswarm plot». На таком типе графиков синие точки означают объекты с малым значением признака, красные – с большим значением признака. Данный график ранжирует параметры по степени их влияния на прогнозы модели сверху вниз от

наибольшего к наименьшему. Графики такого типа помогают понять влияние каждого фактора на прогноз модели в зависимости от расположения точек, их разброса и плотности. Широкий разброс или высокая плотность точек указывают на более значительную изменчивость или более существенное влияние на прогнозы модели. Например, анализируя сводный график на рис. 3, можно прийти к выводу, что высокие оценки за первый семестр и в то же время низкие оценки за второй семестр и низкий средний бал за последующие семестры, скорее всего, могут привести к досрочному отчислению студента из вуза.

Чтобы определить, какие признаки, в целом, являются наиболее важными для прогнозов, выдаваемых моделью, можно использовать функцию shap.plots.bar инструмента SHAP, отражающую результаты усреднения значений Шепли по всем наблюдениям. На рис. 5 представлены гистограммы важности факторов на основе их средних значений Шепли для двух моделей. Первая модель была обучена на данных студентов женского пола, а вторая модель – на данных студентов мужского пола. Графики предоставляют возможность наглядно сравнить, какие факторы оказывают наибольшее значения на прогнозы обученных моделей. Для модели XGBoost со значительным перевесом наибольшую значимость на прогнозы модели оказывает общий средний балл за все семестры обучения, следующий по важности фактор — это средний балл за первый семестр.

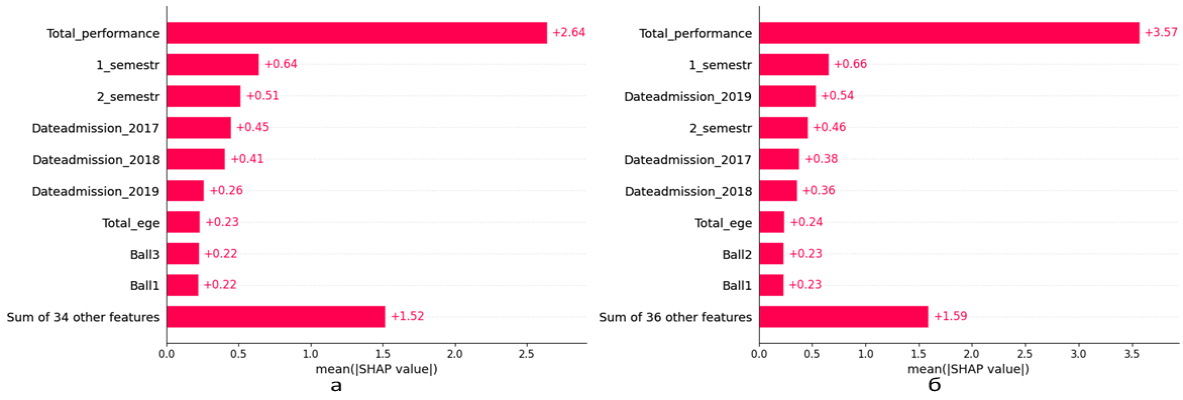

Рис. 5. Столбчатая диаграмма (shap.plots.bar), отражающую результаты усреднения SHAP-значений для модели XGBoost по всем объектам (а – студенты женского пола, б – студенты мужского пола)

ЗАКЛЮЧЕНИЕ

В работе проведено исследование возможностей применения методов машинного обучения в образовательной аналитике на основе построения и обучения моделей досрочного отчисления студентов из университета. Построенные модели способны с высокой точностью определять студентов в группе риска. Основное внимание было сосредоточено на применении методов интерпретации прогнозов обученных моделей. С помощью данных методов удалось выявить наиболее важные входные признаки модели – характеристики студентов, которые максимально влияют на вероятность их досрочного отчисления.

Методы интерпретации моделей помогают не только понять принципы работы моделей машинного обучение, но и проверить, правильно ли соотносится наше собственное мышление с выводами, полученными с их помощью. На основе выявленных факторов можно настроить эффективный процесс обучения, подстраиваться под индивидуальные характеристики студентов и их потребности.

Благодарности

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета ("ПРИОРИТЕТ-2030").

СПИСОК ЛИТЕРАТУРЫ

1. *Груздев И. А., Горбунова Е. В., Фрумин И. Д.* Студенческий отсев в российских вузах: к постановке проблемы // Вопросы образования. 2013. № 2. С. 67–81. https://doi.org/10.17323/1814-9545-2013-2-67-81

2. *Терентьев Е.А., Груздев И.А., Горбунова Е.В.* Суд идёт: дискурс преподавателей об отсеве студентов // Вопросы образования. 2015. № 2. С. 129–151. https://doi.org/10.17323/1814-9545-2015-2-129-151

3. *Горбунова Е.В.* Выбытия студентов из вузов: исследования в России и США // Вопросы образования. 2018. № 1. С. 110–131. https://doi.org/10.17323/1814-9545-2018-1-110-131

4. *Горбунова Е.В.* Влияние адаптации первокурсников к университету на вероятность их отчисления из вуза // Universitas. Журнал о жизни университетов. 2013. № 2 (1). С. 59–84.

5. *Климова Т.А., Ким А.Т., Отт М.А.* Индивидуальные образовательные траектории студентов как условие качественного университетского образования // Университетское управление: практика и анализ. 2023. 27 (1). С. 23–33. https://doi.org/10.15826/umpa.2023.01.003

6. *Мещеряков А.О., Баянова Н.А., Калинина Е.А., Денисов В.А.* Предикторы выбытия студентов медицинского вуза // Медицинское образование и профессиональное развитие. 2022. № 3 (47). URL: https://www.medobr.ru/ru/jarticles/736.html?SSr=0101348cba14ffffffff27c__07e60b0e0e0130-1843 (дата обращения 01.03.2024)

7. *Шмелева Е. Д.* Факторы отсева студентов инженерно-технического профиля в российских вузах // Вопросы образования. 2020. № 3. С. 110–136.

8. *Мухамадиева К.Б.* Машинное обучение в совершенствовании образовательной среды // Образование и проблемы развития общества. 2020. № 4 (13). С. 70–77.

9. *Shrikumar A., Greenside P., Kundaje A.* Learning important features through propagating activation differences // ICML'17. 2017. P. 3145–3153. URL: https://proceedings.mlr.press/v70/shrikumar17a.html (дата обращения 01.03.2024)

10. *Apley D. W., Jingyu Zhu* Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models // Journal of the Royal Statistical Society Series B: Statistical Methodology. 2020. No 4 (82). P. 1059–1086. https://doi.org/10.1111/rssb.12377

11. *Linardatos P., Papastefanopoulos V., Kotsiantis S.* Explainable AI: A Review of Machine Learning Interpretability Methods // Entropy. 2020. https://doi.org/10.3390/e23010018

12. *Rachha A., Seyam M.* Explainable AI in Education: Current Trends, Challenges, And Opportunities // SoutheastCon. 2023. P. 232–239. https://doi.org/10.1109/SoutheastCon51012.2023.10115140

13. *Fan F.L., Xiong J., Li M., Wang G.* On Interpretability of Artificial Neural Networks: A Survey // IEEE Trans Radiat. Plasma Med Sci. 2021. No. 5 6). P. 741–760. https://doi.org/10.1109/trpms.2021.3066428

14. *Fiore U.* Neural Networks in the Educational Sector: Challenges and Opportunities // Balkan Region Conference on Engineering and Business Education. 2019. No 1 (1). P. 332–337. https://doi.org/10.2478/cplbu-2020-0039

15. *Montavon G., Samek W., Müller K.-R.* Methods for interpreting and understanding deep neural networks // Digital Signal Processing. 2018. No. 73. P. 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

16. *Saranya A., Subhashini R.* A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends // Decision Analytics Journal. 2023. No. 7. https://doi.org/10.1016/j.dajour.2023.100230

17. *Murdoch W.J., Singh C., Kumbier K., Abbasi-Asl R., Yu B.* Definitions, methods, and applications in interpretable machine learning // Proceedings of the National Academy of Sciences. 2019. No. 16 (44). P. 22071–22080

18. *Meyer Lauritsen S. et al.* Explainable artificial intelligence model to predict acute critical illness from electronic health records // Nature Communications. 2020. № 11 (1).

19. *Linden T., Jong J., Lu C., Kiri V., Haeffs K., Fröhlich H.* An explainable multimodal neural network architecture for predicting epilepsy comorbidities based on administrative claims data // Frontiers in Artificial Intelligence. 2021. No. 4. https://doi.org/10.3389/frai.2021.610197

20. *Lu Y., Murzakhanov I., Chatzivasileiadis S.* Neural network interpretability for forecasting of aggregated renewable generation // In Proceedings of 2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids. 2021. P. 282–288.

21. *Mai-Anh T. Vu et al.* A shared vision for machine learning in neuroscience // Journal of Neuroscience. 2018. 38 (7). P. 1601–1607.

22. *Sundararajan M., Taly A., and Yan Q.* Axiomatic attribution for deep networks // CoRR. 2017. URL: https://arxiv.org/abs/1703.01365

23. *Kokhlikyan N. et al.* Captum: A unified and generic model interpretability library for pytorch // CoRR. 2020. URL: https://arxiv.org/abs/2009.07896 (дата обращения 01.03.2024)

24. *Lundberg S.M., Lee S.-I.* A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. 2017. No. 30. P. 4765–4774.

25. *Linardatos P., Papastefanopoulos V., Kotsiantis S.* Explainable AI: A Review of Machine Learning Interpretability Methods // Entropy. 2020. https://doi.org/10.3390/e23010018

26. *Sahakyan M., Aung Z., Rahwan T.* Explainable Artificial Intelligence for Tabular Data: A Survey // IEEE Access. 2021. No. 9. P. 135392–135422. https://doi.org/10.1109/ACCESS.2021.3116481

27. *Шобонов Н.А., Булаева М.Н., Зиновьева С.А.* Искусственный интеллект в образовании // Проблемы современного педагогического образования. № 79 (4). 2023. С. 288–290.

28. *Khosravi H. et al.* Explainable Artificial Intelligence in education // Computers and Education: Artificial Intelligence. 2022. No. 3. https://doi.org/10.1016/j.caeai.2022.100074

29. *Гафаров Ф.М., Руднева Я.Б., Шарифов У.Ю.* Прогностическое моделирование в высшем образовании: определение факторов академической успеваемости // Высшее образование в России. 2023. Т. 32. № 1. С. 51–70. https://doi.org/10.31992/0869-3617-2023-32-1-51-7

ANALYSING MACHINE LEARNING MODELS BASED ON EXPLAINABLE ARTIFICIAL INTELLIGENCE METHODS IN EDUCATIONAL ANALYTICS

D. A. Minullin[1] [0000-0001-7713-5251], F. M. Gafarov[2] [0000-0003-4704-154X]

[1, 2]*Kazan federal university*

[1]minullin.dima@mail.ru, [2]fgafarov@yandex.ru

Abstract

The problem of predicting early dropout of students of Russian universities is urgent and therefore requires the development of new innovative approaches to solve it. To solve this problem, it is possible to develop predictive systems based on the use of student data, available in the information systems of universities. This paper investigates machine learning models for predicting early student dropout trained on the basis of student characteristics and performance data. The main scientific novelty of

the work lies in the use of explainable AI methods to interpret and explain the performance of the trained machine learning models. The Explainable AI methods allow us to understand which of the input features (student characteristics) have the greatest influence on the results of the machine learning models. (student characteristics) have the greatest influence on the prediction results of trained models, and can also help to understand why the models make certain decisions. The findings expand the understanding of the influence of various factors on early dropout of students.

***Keywords**: educational analytics, data mining, machine learning, explainable AI*

REFERENCES

1. *Gruzdev I.A., Gorbunova E.V., Frumin I.D.* Studencheskij otsev v rossijskih vuzah: k postanovke problemy [Student dropout in russian higher education institutions: the problem statement] // Educational Studies Moscow. 2013. № 2. P. 67–81. https://doi.org/10.17323/1814-9545-2013-2-67-81

2. *Terentyev E.A., Gruzdev I.A., Gorbunova E.V.* Sud idyot: diskurs prepodavatelej ob otseve studentov [The Court Is Now in Session: Professor Discourse on Student Attrition] // Educational Studies Moscow. 2015. №2. S. 129–151. https://doi.org/10.17323/1814-9545-2015-2-129-151

3. *Gorbunova E.V.* Vybytiya studentov iz vuzov: issledovaniya v Rossii i SSHA [Research on Student Departure in Russia and the U.S.] // Educational Studies Moscow. 2018. №1. S. 110–131. https://doi.org/10.17323/1814-9545-2018-1-110-131

4. *Gorbunova E.V.* Vliyanie adaptacii pervokursnikov k universitetu na veroyatnost' ih otchisleniya iz vuza [The Effect of Adaptation of Freshmen to the University on the Probability of Their Expulsion from the University] // Universitas. 2013. № 2 (1). S. 59–84.

5. *Klimova T.A., Kim A.T., Ott M.A.* Individual'nye obrazovatel'nye traektorii studentov kak uslovie kachestvennogo universitetskogo obrazovaniya [Individual Educational Trajectories as a Condition for High-Quality University Education] // University Management: Practice and Analysis. 2023. 27(1). S. 23–33. https://doi.org/10.15826/umpa.2023.01.003

6. *Meshcheryakov A.O, Bayanova N.A., Kalinina E.A.* Prediktory vybytiya studentov medicinskogo vuza // Medicinskoe obrazovanie i professional'noe razvitie. 2022. № 3 (47).
URL: https://www.medobr.ru/ru/jarticles/736.html?SSr=0101348cba14ffffffff27c__07e60b0e0e0130-1843

7. *Shmeleva E.D.* Faktory otseva studentov inzhenerno-tekhnicheskogo profilya v rossijskih vuzah [Factors of Attrition among Computer Science and Engineering Undergraduates in Russia] // Educational Studies Moscow. 2020. №3. S. 110–136.

8. *Mukhamadieva K.B.* Mashinnoe obuchenie v sovershenstvovanii obrazovatel'noj sredy // Obrazovanie i problemy razvitiya obshchestva. 2020. № 4(13). S. 70–77.

9. *Shrikumar A., Greenside P., Kundaje A*. Learning im-portant features through propagating activation differences // ICML'17. 2017. P. 3145–3153. URL: https://proceedings.mlr.press/v70/shrikumar17a.html

10. *Apley D.V., Zhu J.* Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models // Journal of the Royal Statistical Society Series B: Statistical Methodology. 2020. No. 4 (82). P. 1059–1086. https://doi.org/10.1111/rssb.12377

11. *Linardatos P., Papastefanopoulos V., Kotsiantis S.* Explainable AI: A Review of Machine Learning Interpretability Methods // Entropy. 2020.
https://doi.org/10.3390/e23010018

12. *Rachha A., Seyam M.* Explainable AI In Education: Current Trends, Challenges, And Opportunities // SoutheastCon. 2023. P. 232–239.
https://doi.org/10.1109/SoutheastCon51012.2023.10115140

13. *Fan F.L., Xiong J., Li M., Wang G.* On Interpretability of Artificial Neural Networks: A Survey // IEEE Trans Radiat Plasma Med Sci. 2021. No. 5 (6). P. 741–760. https://doi.org/10.1109/trpms.2021.3066428

14. *Fiore U.* Neural Networks in the Educational Sector: Challenges and Opportunities // Balkan Region Conference on Engineering and Business Education. 2019. No. 1 (1). P. 332–337. https://doi.org/10.2478/cplbu-2020-0039

15. *Montavon G., Samek W., Müller K.-R.* Methods for interpreting and understanding deep neural networks // Digital Signal Processing. 2018. No. 73. P. 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

16. *Saranya A., Subhashini R.* A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends // Decision Analytics Journal. 2023. No. 7. https://doi.org/10.1016/j.dajour.2023.100230

17. *Murdoch W.J., Singh C., Kumbier K., Abbasi-Asl R., Yu B.* Definitions, methods, and applications in interpretable machine learning // Proceedings of the National Academy of Sciences. 2019. No. 16 (44). P. 22071–22080

18. *Lauritsen S.M. et al.* Explainable artificial intelligence model to predict acute critical illness from electronic health records // Nature Communications. 2020. No. 11 (1).

19. *Linden T., Jong J., Lu C., Kiri V., Haeffs K., Fröhlich H.* An explainable multimodal neural network architecture for predicting epilepsy comorbidities based on administrative claims data // Frontiers in Artificial Intelligence. 2021. No. 4. https://doi.org/10.3389/frai.2021.610197

20. *Lu Y., Murzakhanov I., Chatzivasileiadis S.* Neural network in-terpretability for forecasting of aggregated renewable generation // In Proceedings of 2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids. 2021. P. 282–288.

21. *Mai-Anh T. Vu et al.* A shared vision for machine learning in neuroscience // Journal of Neuroscience. 2018. 38(7). P. 1601–1607.

22. *Sundararajan M., Taly A., and Yan Q.* Axiomatic attribution for deep networks // CoRR. 2017. URL: https://arxiv.org/abs/1703.01365

23. *Kokhlikyan N. et al.* Captum: A unified and generic model interpretability library for pytorch // CoRR. 2020. URL: https://arxiv.org/abs/2009.07896 (last access 01.03.2024)

24. *Lundberg S.M., Lee S.-I.* A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. 2017. No. 30. P. 4765–4774.

25. *Linardatos P., Papastefanopoulos V., Kotsiantis S.* Explainable AI: A Review of Machine Learning Interpretability Methods // Entropy. 2020. https://doi.org/10.3390/e23010018

26. *Sahakyan M., Aung Z., Rahwan T.* Explainable Artificial Intelligence for Tabular Data: A Survey // IEEE Access. 2021. No. 9. P. 135392-135422.

https://doi.org/10.1109/ACCESS.2021.3116481

27. *Shobonov N.A., Bulaeva M.N., Zinovieva S.A.* Iskusstvennyj intellekt v obrazovanii [Artificial Intelligence in Education] // Problemy sovremennogo pedagogicheskogo obrazovaniya. № 79 (4). 2023. S. 288–290.

28. *Khosravi H. et al.* Explainable Artificial Intelligence in education // Computers and Education: Artificial Intelligence. 2022. No. 3.
https://doi.org/10.1016/j.caeai.2022.100074

29. *Gafarov F.M., Rudvena Ya.B., Sharifov U.Yu.* Prognosticheskoe modelirovanie v vysshem obrazovanii: opredelenie faktorov akademicheskoj uspevaemosti [Predictive Modeling in Higher Education: Determining Factors of Academic Performance] // Higher Education in Russia. 2023. V. 32. No. 1. S. 51–70.
https://doi.org/10.31992/0869-3617-2023-32-1-51-7

СВЕДЕНИЯ ОБ АВТОРАХ

МИНУЛЛИН Дмитрий Артурович – аспирант, Казанский федеральный университет.

Dmitriy Arturovich MINULLIN – Postgraduate student, Kazan Federal University.

email: minullin.dima@mail.ru

ORCID: 0000-0001-7713-5251.

ГАФАРОВ Фаиль Мубаракович – кандидат физ.-мат. наук, доцент, Казанский федеральный университет.

Fail Mubarakovich GAFAROV – Candidate of Science in Physics and Mathematics, Docent, Kazan Federal University.

email: fgafarov@yandex.ru

ORCID: 0000-0003-4704-154X.

Материал поступил в редакцию 2 мая 2024 года