УДК 378.146.8

Клементьев Александр Александрович

соискатель кафедры специальной психологии и психолого-социальных технологий Московского государственного педагогического университета
https://orcid.org/0000-0002-5373-648X

Klementiev Aleksandr Aleksandrovich

External PhD student, Department of Special Psychology, Psychological and Social Technologies, Moscow Pedagogical State University
https://orcid.org/0000-0002-5373-648X

ИСПОЛЬЗОВАНИЕ ПРОГНОЗНЫХ МЕТОДОВ ОЦЕНКИ УСПЕВАЕМОСТИ АСПИРАНТОВ НА ОСНОВЕ ДАННЫХ LMS-ПЛАТФОРМ

USING PREDICTIVE METHODS FOR ASSESSING ACADEMIC PERFORMANCE OF GRADUATE STUDENTS BASED ON LMS PLATFORM DATA

Аннотация:
В статье оценивается перспективность использования LMS-платформ и современных методов статистического анализа данных – деревьев решений – для превентивной оценки академической успеваемости российских аспирантов и предсказания вероятности защиты ими кандидатской диссертации. В работе постулируется наличие проблемы низкой доли студентов, успешно заканчивающих обучение в аспирантуре и защищающих диссертацию. В процессе поиска инструмента для ее решения автор проводит обзор исследований по выявлению факторов успеваемости студентов, осуществляет краткое сравнение доступных статистических приемов и источников данных, подробно рассматривает пример практического применения выбранного инструментария зарубежными учеными. Постулируется вывод о необходимости проведения практического исследования по изучению возможности применения данных LMS и метода деревьев решений для предсказания успеваемости российских аспирантов.

Ключевые слова:
LMS, успеваемость, факторы успеваемости, деревья решений, статистическое моделирование, цифровое образование

Summary:
The paper assesses the prospects for using LMS platform data and modern methods of statistical data analysis – decision trees – for a proactive assessment of the academic performance of Russian graduate students and predicting the likelihood of them successfully defending their thesis. The paper postulates the problem of only a small proportion of Russian graduate students successfully complete postgraduate studies and defend their thesis. In the process of finding a solution to this issue, the author conducts a review of relevant studies to identify student performance factors, provides a brief comparison of the available statistical tools and data sources, examines a case of practical application of the selected tools by western researchers. The author comes to the conclusion that a practical study is necessary to access the possibility of using LMS data along with decision trees to predict Russian graduate students' academic performance.

Keywords:
LMS, academic performance, factors of academic performance, decision trees, statistical modeling, digital education

Введение. Одной из ключевых проблем послевузовского образования в России является низкая доля аспирантов, успешно заканчивающих обучение и защищающих диссертацию. По последним данным службы государственной статистики Российской Федерации, за 2018 г. только 65,6 % обучающихся от числа поступивших в аспирантуру завершили свое обучение в ней с положительным результатом. Это самый низкий показатель за 8 лет.

Вместе с тем снижается и доля аспирантов, успешно защищающих диссертацию. В 2018 г. только 12,4 % обучавшихся в аспирантуре граждан успешно прошли через процедуру защиты. Такая тенденция наблюдалась последние несколько лет, а в 2018 г. отмечено наименьшее количество защит с 2010 г. [1]

Тем не менее потребность в сохранении числа аспирантов остается актуальной, что подтверждается принятием в первом чтении закона «О внесении изменений в отдельные законодательные акты Российской Федерации в части подготовки научно-педагогических кадров в аспирантуре (адъюнктуре)», который обязывает аспирантов подготовить и предоставить полноценную диссертацию для прохождения итоговой аттестации и получения диплома об окончании аспирантуры [2].

Директивные меры и законодательные инициативы, однако, не являются единственно возможным способом повысить количество выпускников аспирантуры и защит. Развитие инструментов статистического моделирования и анализа данных, а также цифровизация образовательного процесса дают современным исследователям в сфере образования мощный инструмент для

предсказания будущей успеваемости учащихся и превентивного выявления студентов, которые с высокой вероятностью не сумеют освоить программу послевузовского образования.

Настоящая работа ставит перед собой цель продемонстрировать, как внедрение в российские вузы систем управления обучением (LMS) и активное их использование позволяют создать инструмент статистического анализа образовательных траекторий учащихся, который может быть использован для улучшения показателей окончания аспирантуры и числа защищенных диссертаций.

История изучения факторов успеваемости студентов. Попытки использовать статистический анализ для предсказания успеваемости студентов предпринимались с начала XX в. Пионером подобных работ стал Дж. Джонсон, который в 1926 г. попробовал предсказать вероятность успешного окончания университета в момент поступления [3, с. 82].

На протяжении следующих десятилетий было предложено множество теоретических моделей, объясняющих взаимосвязь между академической успеваемостью и различными социальными, экономическими, демографическими и поведенческими характеристиками учащихся [4, с. 588–589].

На рубеже XXI в. интерес исследователей в значительной степени сместился в сторону академических предикторов будущей успеваемости: было установлено, что одним из наиболее достоверных средств прогнозирования будущих достижений и успехов в обучении являются предыдущие и текущие академические результаты, а также личная вовлеченность учащегося в образовательный процесс, которую можно оценить через поведенческие показатели, такие как частота посещения занятий, полнота, своевременность и качество выполнения учебных заданий [5, с. 62–63].

К сожалению, работа большинства исследователей, которые занимались этой проблемой в прошлом, была в значительной степени осложнена проблемой качества данных. Недоступность достаточного объема точной, релевантной и быстро получаемой информации об образовательном процессе затрудняла выявление взаимосвязей между изучаемыми показателями и делала невозможным проведение оперативного моделирования для оценки ситуации, касающейся конкретных учащихся. Из-за недостатка данных анализ, как правило, ограничивался рассмотрением таблиц распределений, либо, в лучшем случае, построением регрессионных моделей в попытке выделить ключевые детерминанты успеваемости [6, с. 589].

Однако в результате проникновения информационных технологий в систему образования ситуация кардинально изменилась. Цифровизация процесса обучения, появление и расширение баз данных учеников и развитие онлайн-образования значительно увеличили доступный для анализа объем информации. Ключевым элементом в процессе сбора и накопления необходимых данных выступили электронные системы управления образованием (LMS), которые получают все более широкое распространение в высших учебных заведениях России [7, с. 294].

LMS способны собирать и хранить большие объемы информации о пользовательской активности и взаимодействии. Перечень данных, которыми позволяют оперировать электронные системы управления образованием, включает в себя показатели прошлой и текущей успеваемости, количество и продолжительность онлайн-сессий (посещений студентами сайтов онлайн-курсов), факт и частоту использования инструментов LMS, чтение или публикацию сообщений в системе, регулярность загрузки работ и заданий. Все эти сведения собираются в режиме реального времени и могут быть выгружены в любой момент. При этом сбор информации ненавязчив для пользователя и не требует вмешательства преподавателей или специального персонала. И хотя набор показателей, детализация данных и историческая глубина информации могут различаться от системы к системе и от учреждения к учреждению, эти данные характеризуют аспекты поведения учащихся, которые трудно или невозможно проанализировать другими способами, например, вовлеченность в образовательный процесс или стратегии и траектории обучения [8, с. 222].

Исследователи начали использовать возможности систем управления образованием для предсказания успеваемости почти сразу после возникновения LMS. Так, уже в 2002 г. Ванг и Ньюлин попытались выявить факторы, определяющие успешность обучения учащихся в рамках дистанционного образования с использованием системы LMS. Эти работы положили начало целому виду исследований, направленных на использование появившихся возможностей, предоставленных системами управления образованием, для поиска наиболее значимых детерминант успеха в обучении и предсказания успеваемости учащихся [9, с. 21].

Минай-Бидголи, Ромеро и Церезо предложили использовать данные систем LMS для предсказания итоговых оценок учащихся на основании показателей посещаемости ими сетевых курсов. Авторы также включили в анализ такие переменные, как количество онлайн-сессий, частота входов в систему, количество прочитанных и созданных оригинальных сообщений, число просмотренных страниц контента [10].

В последующих исследованиях список переменных для анализа расширился. Константиас и Пинтелас, а также Оладокун объединяли данные LMS с демографическими параметрами студентов [11, с. 75].

Ученые также начали уделять больше внимания не количественным, а качественным показателям вовлеченности студентов в обучение. Так, в дополнение к частоте обращения к системе в анализ были включены параметры длительности и регулярности учебных сессий, что позволило добиться более надежных результатов [12].

Параллельно с развитием предиктивной аналитики на основе данных LMS появился подход, объединяющий информацию систем управления обучением с более традиционными опросными данными. Так, Саркер и коллеги предложили три предсказательные модели на основе комбинации нескольких источников данных. В их работе предлагается объединение опросных сведений, полученных в результате анкетирования студентов, и данных системы LMS. В результате сравнения моделей исследователи пришли к выводу, что включение опросных данных в модель увеличивает точность предсказания в сравнении с моделями, построенными исключительно на информации LMS [13, с. 413].

В целом, корпус проведенных к настоящему времени исследований позволяет заключить, что системы управления образованием являются ценным источником качественно новых данных, которые могут быть эффективно использованы для предсказания успеваемости учащихся и построения прогностических моделей. Подобные модели потенциально способны заранее определить академические возможности конкретного студента, установить вероятность того, что он не справится с академической нагрузкой и не сможет завершить учебную программу.

Таким образом, использование данных систем LMS, потенциально вместе с опросными данными, позволяет создать инструмент, способный частично решить проблему снижающейся доли выпускников и защит в системе аспирантуры России путем превентивного выявления студентов, находящихся под угрозой отчисления, и вмешательства в их учебный процесс.

Выбор инструментария. В большей части работ, рассмотренных в настоящей статье, использовались традиционные для исследовательской среды методы прогнозирования успеваемости – построение линейных или логистических регрессионных моделей. Регрессионный анализ получил широкое распространение в научной среде из-за возможностей метода обеспечить точное выявление связей между отдельными переменными в ситуации сложной взаимозависимости. Регрессионный анализ позволяет выявить и численно выразить степень влияния каждого из множества предикторов на целевой показатель, что делает этот метод исключительно полезным, например, при определении факторов, влияющих на успеваемость учащихся. Однако у регрессионного анализа есть несколько особенностей, ограничивающих его эффективность для решения прикладных задач в сфере образования.

Регрессионные модели создают большое количество предпосылок и ограничений на исходные данные и результаты построения моделей. Так, распределение категорий переменных, используемых в анализе, должно соответствовать нормальному, «остатки», т. е. ошибки модели, должны быть гомоскедастичными и не смещенными относительно нуля. Кроме того, такие модели чувствительны к пропускам данных: отсутствие хотя бы одного значения в параметрах респондента обычно приводит к исключению его из анализа. При использовании регрессий также осложнено нахождение эффектов взаимодействия данных, т. е. совместного влияния предикторов. Процесс поиска этих эффектов не автоматизирован, требует большого количества манипуляций с моделью и объемных вычислений.

Регрессионный анализ изначально создавался как способ нахождения линейных связей. И хотя исследователь может использовать регрессии для установления более сложных видов корреляций между переменными, это потребует от него дополнительных усилий и инициативы. Наконец, модели, построенные в результате применения регрессионного анализа, сложны для понимания и интерпретации людьми, не обладающими хотя бы базовым образованием в области статистики и анализа данных, что сильно ограничивает сферу практического применения этих моделей [14].

Однако существует группа статистических методов, которые позволяют решать задачу прогнозирования успеваемости и которые при этом лишены части недостатков, присущих регрессионным моделям, – деревья решений.

Получившие признание в 1980-х гг. они предполагают построение автоматизированных индукционных алгоритмов. Классическое дерево решений – это многомерный граф, который поэтапно разделяет все попавшие в него объекты на отдельные группы в зависимости от значений переменных-предикторов.

Деревья решений обладают рядом преимуществ по сравнению с регрессионными моделями. Во-первых, в результате их построения создается легко интерпретируемая схема, а поэтапный характер принятия решений, который реализован в деревьях, схож с образом мышления человека. Это свойство позволяет исследователю делать выводы о принадлежности конкретного наблюдения тому или иному классу, не обладая специальными знаниями в области статистики. Во-вторых, в отличие от регрессий, деревья решений не накладывают жестких ограничений на тип и качество исходных данных, что позволяет включать в анализ случаи с пропущенными значениями показателей и статистическими «выбросами» [15, с. 91–94].

Зарубежный опыт. Зарубежные исследователи уже продемонстрировали эффективность предсказательных моделей, построенных при помощи деревьев решений на основе данных систем LMS. Так, в 2017 г. коллектив авторов из Университета Западной Шотландии предпринял попытку предсказать успеваемость студентов этого учебного заведения и факт получения ими диплома в будущем.

В качестве источников данных для модели авторы использовали следующие ресурсы.

1. Система LMS. Из LMS в анализ были включены показатели общего времени, проведенного на платформе, а также количество посещенных ресурсов, число попыток сдачи онлайн-тестов, количество пройденных курсов, онлайн-обсуждений, в которых студент принял участие.
2. Система SRS (Student Record System – система фиксации информации о студентах). Из этой системы была извлечена информация о следующих показателях: пол, возраст, национальность, место проживания, кампус, тип программы, предыдущий уровень образования, наличие инвалидности.
3. Онлайн-опрос студентов. С его помощью были собраны данные о следующих показателях: социально-экономический статус родителей студентов; количество часов, проведенных на работе; количество часов, проведенных за учебой; текущий уровень образования; объем поддержки со стороны семьи; удовлетворенность качеством образования; тип обучения; состояние здоровья; уровень адаптации к университетской среде; наличие предыдущих знаний по изучаемым дисциплинам.

Помимо деревьев решений в анализ были включены такие статистические алгоритмы, как нейронная сеть и метод опорных векторов. Авторы оценили возможность совместного использования алгоритмов для повышения точности предсказания.

Создавая различные комбинации источников данных и статистических алгоритмов, исследователи разработали 7 предсказательных моделей, которые оценивают академическую успеваемость студентов Университета Западной Шотландии. Перечень моделей и их наполнение представлены в таблице 1.

Таблица 1 – Описание статистических моделей

№ модели	Модель 1	Модель 2	Модель 3	Модель 4	Модель 5	Модель 6	Модель 7
Источники данных	Система SRS	Система LMS	Онлайн-опрос	SRS и LMS	SRS и онлайн-опрос	LMS и онлайн-опрос	SRS, LMS, онлайн-опрос
Алгоритм	Дерево решений			Дерево решений, нейронная сеть, метод опорных векторов			
Целевая переменная	Успеваемость студента						

После построения моделей была проведена оценка их эффективности в предсказании успеваемости студентов посредством разделения выборки на две части: тренировочную и тестовую. Модели строились на тренировочной части выборки, но проверялись на тестовой – для исключения возможности «переобучения», т. е. искусственного завышения показателей точности.

Итоговые показатели точности различных моделей и алгоритмов представлены в таблице 2.

Таблица 2 – Сравнение эффективности моделей

Модель	Точность, %	RMSE	Чувствительность, %	Специфичность, %	Доля ошибок, %
Модель 1	34,81	0,751	10,78	13,84	65,19
Модель 2	50,33	0,685	25,76	25,36	49,67
Модель 3	78,05	0,464	13,66	15,67	21,95
Модель 4	81,62	0,421	77,74	74,52	26,90
Модель 5	73,10	0,503	66,37	60,17	26,90
Модель 6	80,81	0,421	70,00	71,94	19,19
Модель 7	81,67	0,396	79,62	75,86	18,33

Помимо стандартного показателя точности, т. е. доли студентов, успеваемость которых была предсказана верно, был оценен показатель RMSE (корень среднеквадратичной ошибки модели), чувствительность (доля истинно-положительных результатов), специфичность (доля истинно-отрицательных результатов) и доля ошибочных предсказаний.

В результате анализа данных исследователи заключили, что использование деревьев решений позволяет предсказать успеваемость студентов с высокой точностью, при этом эффективность модели в значительной степени зависит от качества исходных данных. Так, деревья решений корректно оценили уровень успеваемости 78 % студентов на основании данных онлайн-опроса. Кроме того, объединение источников данных и подключение к анализу дополнительных статистических алгоритмов позволяет добиться еще более высокой точности. К примеру, модель № 7, в которую были включены и данные электронных систем университета (LMS и SRS), и результаты онлайн-опроса, достигла уровня точности в 81,67 %.

Авторы заверяют, что использование нескольких источников данных, включая LMS университета, позволяет построить «очень эффективную и точную модель предсказания академической успеваемости», а также «выявить студентов, находящихся в опасности отчисления» [16, с. 61–75].

Перспективы применения метода в России. Внедрение подобной системы предсказания успеваемости и вероятности успешной защиты диссертации в российской аспирантуре может стать одним из инструментов для увеличения доли аспирантов, успешно завершающих обучение. Однако, как и в случае с любой инновацией, использование такой системы может быть осложнено некоторыми трудностями. Для понимания потенциальных перспектив и возможных сложностей создания и внедрения системы прогнозных методов оценки успеваемости проведем анализ сильных и слабых сторон имплементации такой системы, а также исследование потенциальных угроз, которые могут повлиять на её внедрение (SWOT-анализ).

Сильные стороны: успешное применение алгоритмов построения деревьев решений могло бы позволить создать легкую в интерпретации и применении модель, при помощи которой сотрудники университетов могли бы с высокой точностью определять, принадлежит ли конкретный студент к группе риска (то есть имеет высокие шансы не защитить диссертацию). Наличие подобной модели позволило бы превентивно вносить точечные изменения в образовательный процесс, совершенствовать механизмы контроля или иным образом влиять на ход обучения таких студентов и, возможно, повысить долю аспирантов, заканчивающих обучение защитой диссертации.

Слабые стороны: для эффективной работы прогнозных алгоритмов необходим большой объем точных и разнообразных данных, касающихся как можно большего числа аспектов жизни и учебы аспирантов. Хотя системы LMS существенно увеличили объем доступных исследователям данных, их может оказаться недостаточно для составления достаточно точных прогнозов. В этом случае исследователю придется обогащать данные LMS опросной информацией, что внесет дополнительные организационные трудности в создание модели.

Возможности: внедрение режима самоизоляции и удаленного обучения в вузах России существенно ускорило цифровизацию образовательного процесса. Необходимость переводить большую часть взаимодействий со студентами в Интернет привела к значительному увеличению количества данных об активности и успеваемости последних. Создание системы предсказания успеваемости в момент максимальной доступности данных предоставит ей намного больше возможностей для формирования точных прогнозов.

Угрозы: принятие европейского закона о защите информации (GDPR) и ужесточение российского ФЗ «О персональных данных» демонстрируют общемировую тенденцию к совершенствованию законодательства в сфере манипуляции персональной информацией. Хотя текущий статус законодательства не запрещает исследователю оперировать обезличенными данными студентов с целью построения прогностических моделей, дальнейшее ужесточение требований может осложнить получение необходимых данных.

Заключение. В целом, использование прогнозных методов оценки успеваемости аспирантов на основе данных LMS-платформ может рассматриваться как перспективный инструмент повышения показателей эффективности аспирантуры. Однако для проверки возможности использования деревьев решений в этой ситуации необходимо провести предварительное исследование. Имеет смысл оперировать уже имеющимися в распоряжении высших учебных заведений данными о результатах защит или невыходе на защиту бывших аспирантов, а также сведениями о них, хранящимися в системах LMS, для построения нескольких предсказательных моделей при помощи разных алгоритмов деревьев решений с последующей оценкой их эффективности.

Ссылки:

1. Основные показатели деятельности аспирантуры и докторантуры [Электронный ресурс] // Федеральная служба государственной статистики. URL: https://www.gks.ru/storage/mediabank/asp-1.xls (дата обращения: 21.10.2020).

2. О внесении изменений в отдельные законодательные акты Российской Федерации в части подготовки научно-педагогических кадров аспирантуре (адъюнктуре)» : федеральный закон [Электронный ресурс] // Система обеспечения законодательной деятельности. URL: http://sozd.duma.gov.ru/download/870DCA96-F704-42E2-82F0-E797A06DF734. (дата обращения: 21.10.2020).
3. Johnson J.B. Predicting Success in College at Time of Entrance // School and Society. 1926. No. 23. P. 82–88.
4. Macfadyen L.P., Dawson S. Mining LMS Data to Develop an «Early Warning System» for Educators: A Proof of Concept // Computers & Education. 2010. Vol. 54, iss. 2. P. 588–599. https://doi.org/10.1016/j.compedu.2009.09.008.
5. Adejo O.W., Connolly T. Predicting Student Academic Performance Using Multi-model Heterogeneous Ensemble Approach // Journal of Applied Research in Higher Education. 2018. Vol. 10, iss. 1. P. 61–75. https://doi.org/10.1108/jarhe-09-2017-0113.
6. Macfadyen L.P., Dawson S. Op. cit. P. 589.
7. Гамбеева Ю.Н., Сорокина Е.И. Развитие электронного обучения как новой модели образовательной среды // Креативная экономика. 2018. Т. 12. № 3. С. 285–304. https://doi.org/10.18334/ce.12.3.38897.
8. Dawson S.P., McWilliam E., Tan J. Teaching Smarter: How Mining ICT Data Can Inform and Improve Learning and Teaching Practice // Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education. Melbourne, 2008. P. 221–230.
9. Wang A.Y., Newlin M.H. Predictors of Performance in the Virtual Classroom: Identifying and Helping at-risk Cyber-students // The Journal of Higher Education. Academic Matters. 2002. № 29 (10). P. 21–25.
10. Cerezo R., Sanchez-Santillan M., Paule-Ruiz M.P., Nunez J.C. Students' LMS Interaction Patterns and Their Relationship with Achievement: a Case Study in Higher Education // Computers & Education. 2016. Vol. 96. P. 42–54. https://doi.org/10.1016/j.compedu.2016.02.006 ; Minaei-Bidgoli B., Kashy D.A., Kortemeyer G., Punch W.F. Predicting Student Performance: an Application of Data Mining Methods with an Educational Web-based System // 33rd Annual Frontiers in Education. Westminster, 2003. P. T2A-13. https://doi.org/10.1109/fie.2003.1263284 ; Romero C., López M.I., Luna J.M., Ventura S. Predicting students' final performance from participation in on-line discussion forums // Computers & Education. 2013. Vol. 68. P. 458–472. https://doi.org/10.1016/j.compedu.2013.06.009.
11. Oladokun V.O., Adebanjo A.T., Charles-Owaba O.E. Predicting Students' Academic Performance Using Artificial Neural Network: a Case Study of an Engineering Course // The Pacific Journal of Science and Technology. Vol. 9, iss. 1. P. 72–79 ; Kotsiantis S.B., Pintelas P.E. Predicting Students Marks in Hellenic Open University // Advanced Learning Technologies, ICALT. Washington, 2005. P. 664–668. https://doi.org/10.1109/icalt.2005.223.
12. Agudo-Peregrina A.F., Iglesias-Pradas S., Conde-Gonzalez M.A., Hernandez-Garcia A. Can We Predict Success from Log Data in VLEs? Classification of Interactions for Learning Analytics and Their Relation with Performance in VLE-supported F2F and Online Learning // Computers in Human Behavior. Vol. 31. P. 542–550. https://doi.org/10.1016/j.chb.2013.05.031 ; Cerezo R., Sanchez-Santillan M., Paule-Ruiz M.P., Nunez J.C. Op. cit. P. 48–49.
13. Sarker F., Tiropanis T., Davis, H.C. Exploring Student Predictive Model that Relies on Institutional Databases and Open Data Instead of Traditional Questionnaires // Proceedings of the 22nd International Conference on World Wide Web ACM. Rio de Janeiro, 2013. P. 413–418. https://doi.org/10.1145/2487788.2487955.
14. Демиденко Е.З. Линейная и нелинейная регрессия. М., 1981. 302 с. ; James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning. New York, 2013. 418 p. https://doi.org/10.1007/978-1-4614-7138-7.
15. Long W.J., Griffith J.L., Selker H.P., D'Agostino R.B. A Comparison of Logistic Regression to Decision-Tree Induction in a Mediacal Domain // Computers and Biomedical Research. 1993. Vol. 26, iss. 1. P. 74–97. https://doi.org/10.1006/cbmr.1993.1005.
16. Adejo O.W., Connolly T. Op. cit. P. 61–75.

Редактор: Ситникова Ольга Валериевна
Переводчик: Кочетова Дарья Андреевна