УДК 004.048, 004.418

ПРОГНОЗИРОВАНИЕ СЛОЖНОСТИ КУРСА НА ОСНОВЕ ОЦЕНОК ПО ОБЕСПЕЧИВАЮЩИМ ДИСЦИПЛИНАМ С ПОМОЩЬЮ МЕТОДА ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ НА ПРИМЕРЕ КУРСА ПО ПРОГРАММИРОВАНИЮ НА PYTHON

Живетьев Александр Викторович[1], Белов Михаил Александрович[2]

[1]Аспирант;
Государственный университет «Дубна»;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: zhivetyev@gmail.com.

[2]Кандидат технических наук, доцент;
Государственный университет «Дубна»;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: belov@uni-dubna.ru.

В статье рассматриваются методы прогнозирования сложности учебных курсов на основе логистической регрессии с использованием оценок по обеспечивающим дисциплинам. Основной объект исследования — курс «Программирование на Python», для которого ключевыми обеспечивающими дисциплинами выбраны математика, информатика и английский язык. Целью исследования является разработка модели, позволяющей адаптировать учебные задания к индивидуальным потребностям студентов, повышая эффективность образовательного процесса. Для реализации модели использованы синтетические данные, что обусловлено ограничениями доступа к реальным образовательным данным. Применение методов машинного обучения, в частности логистической регрессии, позволяет получить не только классификацию курсов по сложности (легкий, средний, сложный), но и вероятностные оценки, отражающие степень уверенности модели в своих предсказаниях. Авторы рассматривают весовые коэффициенты признаков, что позволяет понять вклад каждой обеспечивающей дисциплины в прогнозирование сложности. Прогнозирование сложности курсов и заданий способствует более точному подбору учебных материалов, что улучшает качество образования и способствует развитию персонализированных образовательных траекторий. Таким образом, статья вносит вклад в развитие методов образовательной аналитики и подчеркивает необходимость перехода от прогнозирования успеваемости студентов к прогнозированию сложности курсов, что открывает новые перспективы для персонализации образовательного процесса и повышения его эффективности.

Ключевые слова: индивидуальные образовательные траектории, сложность курса, обеспечивающие дисциплины, машинное обучение, EDM, учебная аналитика, подбор заданий, логистическая регрессия.

Для цитирования:
Живетьев А. В., Белов М. А. Прогнозирование сложности курса на основе оценок по обеспечивающим дисциплинам с помощью метода логистической регрессии на примере курса по программированию на Python // Системный анализ в науке и образовании: сетевое научное издание. 2024. № 2. С. 91-97. EDN: HNKMGS. URL: https://sanse.ru/index.php/sanse/article/view/620.

PREDICTING COURSE DIFFICULTY BASED ON GRADES IN SUPPORTING DISCIPLINES USING LOGISTIC REGRESSION: A CASE STUDY OF A PYTHON PROGRAMMING COURSE

Zhivetyev Alexander V.[1], Belov Mikhail A.[1]

[1]PhD student;
Dubna State University,

19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: zhivetyev@gmail.com.

[2]PhD in Engineering sciences, associate professor;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: belov@uni-dubna.ru.

The article discusses methods for predicting the difficulty of academic courses based on logistic regression using grades from prerequisite subjects. The main subject of the study is the course "Programming in Python" for which key prerequisite subjects are mathematics, computer science, and English. The aim of the study is to develop a model that allows for the adaptation of academic assignments to the individual needs of students, thereby enhancing the effectiveness of the educational process. Synthetic data is used to implement the model due to limitations in access to real educational data. The application of machine learning methods, particularly logistic regression, allows not only for the classification of courses by difficulty (easy, medium, hard) but also for probabilistic assessments that reflect the model's confidence in its predictions. The authors examine the weight coefficients of the features, which allows for an understanding of the contribution of each prerequisite subject to the prediction of difficulty. Predicting the difficulty of courses and assignments facilitates more accurate selection of educational materials, improving the quality of education and promoting the development of personalized educational trajectories. Thus, the article contributes to the advancement of educational analytics methods and emphasizes the need to transition from predicting student performance to predicting course difficulty, opening up new prospects for the personalization and enhancement of the educational process.

Keywords: individual learning paths, course difficulty, supporting disciplines, machine learning, EDM, learning analytics, assignment selection, logistic regression.

For citation:
Zhivetyev A. V., Belov M. A. Predicting Course Difficulty Based on Grades in Supporting Disciplines Using Logistic Regression: A Case Study of a Python Programming Course. *System analysis in science and education*, 2024;(2):91-97 (in Russ). EDN: HNKMGS. Available from: https://sanse.ru/index.php/sanse/article/view/620.

Введение

В множестве научных исследований, посвященных построению индивидуальных образовательных траекторий, сделан акцент на предсказании академической успеваемости студентов на основе их предыдущих оценок. Например, в [1] представлены методы интеллектуального анализа данных, используемые для прогнозирования успеваемости студентов. В частности, работа [2] посвящена прогнозированию успеваемости с использованием методов классификации. В статье [3] прогнозирование успеваемости осуществляется на основе логов из *LMS Moodle*. В [4] решается частная задача предсказания успеваемости – предсказание неудовлетворительной оценки по учебной дисциплине с помощью генетических алгоритмов. Работы посвящены использованию популярных методов машинного обучения, сравнению точности, производительности и прочих полученных метрик, однако не учитывают практическое применение полученных результатов. Около 70% статей посвящено прогнозированию риска академической неуспешности студентов, что подразумевает или риск, связанный с возможностью не сдать вовремя конкретный предмет учебного плана, или риск, связанный с неполучением диплома [5]. Однако на практике, кроме предупреждений о попадании в группу риска и бесед со студентами и их родителями, иных реальных стратегий, как правило, не предлагается [6]. Фактически, от подобных прогнозов нет почти никакой пользы.

В связи с описанными проблемами подобный подход к прогнозированию может кажется недостаточно эффективным и даже неуместным в контексте принятия образовательных решений. Преды-

дущие оценки часто не учитывают множество факторов, влияющих на успех студента, таких как мотивация, качество преподавания и индивидуальные особенности обучающегося. Следовательно, использование прогнозов успеваемости, основанных исключительно на предыдущих оценках, может оказаться несостоятельным и недостаточно информативным для принятия обоснованных решений в образовательном процессе.

Вместо того чтобы прогнозировать успеваемость студентов, стоит обратить внимание на прогнозирование сложности отдельных дисциплин и конкретных учебных материалов. Этот подход позволяет более эффективно управлять образовательным процессом, адаптируя содержание и методы обучения к уровню подготовки и потребностям студентов. Прогноз сложности позволяет персонализировать обучение, учитывая индивидуальные особенности студентов и обеспечивая им оптимальные условия для усвоения знаний. Таким образом, переход от прогнозирования успеваемости к прогнозированию сложности может значительно улучшить эффективность образовательного процесса. Этот подход позволяет более точно адаптировать обучение к потребностям студентов и создавать более подходящие условия для их успешного обучения и развития.

1. Выбор обеспечивающих дисциплин

В качестве объекта исследования была выбрана дисциплина «Программирование на *Python*». Язык *Python* – не только востребованный инструмент в научных исследованиях благодаря своей современной экосистеме и богатым библиотекам для анализа данных, но и широко используемый в различных областях современного мира. Кроссплатформенный Python может быть запущен на различных операционных системах, что делает его универсальным выбором для разработки программного обеспечения, веб-приложений, научных вычислений и многого другого.

Обеспечивающая дисциплина в контексте образовательных курсов представляет собой область знаний и навыков, которая служит основой для успешного освоения материала и достижения учебных целей. Хорошие оценки по этим дисциплинам могут способствовать успешному освоению рассматриваемой дисциплины. Для курса «Программирование на *Python*» были выбраны следующие обеспечивающие дисциплины:

- Математика.
- Информатика.
- Английский язык.

Математика развивает аналитическое мышление, способность к абстракции и решению сложных задач, что является основой для эффективного программирования. Алгоритмы и структуры данных, ключевые аспекты программирования, часто требуют понимания математических концепций, таких как графы, комбинаторика и теория вероятностей. Кроме того, логическое мышление, формализованное через математическое образование, помогает программистам разрабатывать и оптимизировать код, а также решать проблемы, возникающие в процессе разработки.

Для работы с графикой потребуются знания о дифференциальных уравнениях и владение геометрией, математический анализ, физика, вычислительная математика нужны для моделирования естественных процессов, а без дискретной математики не получится писать базы данных или создавать поисковые системы [7].

Студенты и профессионалы, обладающие глубокими знаниями в области информатики, часто могут более эффективно разрабатывать, тестировать и оптимизировать программное обеспечение. Знание информатики позволяет им не только понимать, как работает программный код, но и каким образом данные и алгоритмы могут быть оптимизированы для решения конкретных задач, что существенно повышает качество и производительность финальных программных продуктов. Связь между успехом в программировании и навыками в информатике является очень прямой, поскольку информатика занимается изучением методов и процессов, которые находят свое применение в программи-

рования. Например, в исследовании [8] было выявлено, что наиболее сильная взаимосвязь имеется между баллом ЕГЭ по информатике и текущей успеваемостью по алгоритмизации вычислений. В результате анализа удалось доказать, что результаты ЕГЭ по информатике оказывают более существенное влияние на успеваемость студентов, чем баллы по физике и по русскому языку.

Английский язык признается самым популярным языком в программировании – все типы данных, функции, методы являются английскими фразами, словами или сокращениями. Каждый язык программирования имеет свой алфавит и словарь, свой синтаксис и семантику. При необходимости изучения открытого исходного кода, важно знание языка, на котором этот код описан все наиболее популярные среды основываются на английском языке. Современная литература, посвященная программированию, также преимущественно издается изначально на этом языке [9]. У англоговорящих программистов лучше развита реакция общения с операционной системой и программами в процессе интерактивного диалога, они быстрее решают проблемы отладки и редактирования программных продуктов, быстрее находят ошибки и осваивают еще не переведенную документацию [10].

Следует отметить, что выбор оптимальных обеспечивающих дисциплин не является основным объектом исследования данной статьи. В контексте конкретной образовательной задачи (например, если речь идет про IT-колледж или дополнительное образование) выбор дисциплин может быть и другим. Например, для образовательных учреждений, специализирующихся на преподавании именно информационных технологий, обеспечивающими дисциплинами для курса «Программирование на Python» могут быть: «Алгоритмы и структуры данных», «Базы данных», «Веб-разработка».

2. Логистическая регрессия для прогнозирования сложности

Для проведения настоящего исследования были использованы синтетические данные учащихся. Принятие этого решения было вынужденной мерой, поскольку на момент написания статьи авторам не удалось получить доступ к реальным данным по успеваемости и образовательным траекториям студентов из-за действия жестких законодательных норм по защите персональных данных. Авторы осознают определенные ограничения, связанные с использованием синтетических данных, однако приложили все усилия к тому, чтобы смоделировать наиболее реалистичные сценарии. Следует также отметить, что основной акцент исследования был сделан на разработке и тестировании новых методов персонализации образовательных траекторий, а не на анализе конкретных наборов данных. Таким образом, использование синтетических данных позволило сфокусироваться на алгоритмической составляющей без риска нарушения конфиденциальности персональных сведений.

Пример данных, экспортированных в формат *.csv*, представлен в табл. 1.

Табл. 1. Оценки по математике, информатике и английскому

Математика	Информатика	Английский	Сложность
4	5	4	Легкий
3	3	4	Сложный
4	4	5	Легкий
3	5	4	Легкий
3	3	4	Сложный
4	4	4	Легкий
5	5	5	Легкий
5	5	5	Легкий

4	5	4	Легкий

В столбце «Сложность» хранится метка о сложности освоения курса «Программирование на *Python*». Доступны три значения метки «легкий», «средний», «сложный».

Логистическая регрессия – это метод машинного обучения, который используется для решения задач классификации, то есть для разделения данных на категории или классы. Для реализации модели была выбрана библиотека *scikit-learn*, которая представляет собой мощный инструмент машинного обучения и анализа данных в *Python*. Ее преимущества включают в себя широкий выбор алгоритмов, реализацию эффективных методов обучения, а также простоту использования и интеграции.

После обучения модели получились следующие веса признаков (см. табл. 2).

Табл. 2. Веса признаков для каждого класса

Метка	Математика	Информатика	Английский
Легкий	0.354115	1.073301	0.199614
Средний	-0.431581	-0.975298	-0.636782
Сложный	0.077466	-0.098003	0.437168

Веса признаков позволяют понять, какие признаки вносят наибольший вклад в прогнозы модели для каждого класса. Большие положительные или отрицательные веса могут указывать на то, что определенные признаки сильно влияют на вероятность принадлежности к определенному классу.

Применение логистической регрессии в задачах классификации позволяет не только определять текстовые названия классов, такие как «легкий», «средний» и «сложный», но и получать значения вероятностей принадлежности к каждому из этих классов. Эти вероятности варьируются в диапазоне от 0 до 1 и отражают степень уверенности модели в своём предсказании. Текстовые названия классов удобны для отображения результатов в пользовательских интерфейсах (например, в *LMS*), предоставляя легко интерпретируемую информацию, а численное значение вероятности принадлежности к классу можно использовать как коэффициенты для различных расчетов, например, в можно применить эти значения для определения сложности конкретного задания в рассматриваемом курсе. Это позволяет более точно адаптировать учебный материал к индивидуальным потребностям студентов, распределяя задания в зависимости от вероятностной оценки их сложности.

Заключение

Как подчеркивалось во вступлении, внимание акцентируется не на прогнозе успеваемости студента, а именно на потенциальной сложности курса. Что же делать с полученным прогнозом и как использовать его в практических задачах, направленных на повышение эффективности образовательного процесса в контексте задачи управления индивидуальной образовательной траекторией?

Прогнозирование общей сложности курса, очевидно, может быть полезно при подборе конкретных заданий того или иного уровня сложности. Существуют модели формирования адаптивной индивидуальной образовательной траекторией на основе динамического управления сложностью курса, например, в [11] используется математический аппарат нечеткой логики. Значение, полученное прогнозированием общей сложности курса, может использовано в качестве некоторого «метакоэффициента», на который может быть умножена вероятностная сложность конкретного задания, предлагаемого студенту.

Таким образом, прогнозирование сложности дисциплины (расчет коэффициента сложности) в общем процессе управления индивидуальной образовательной траекторией, можно отразить на следующей схеме (модель верхнеуровневого процесса в нотации *BPMN,* см. рис. 1).

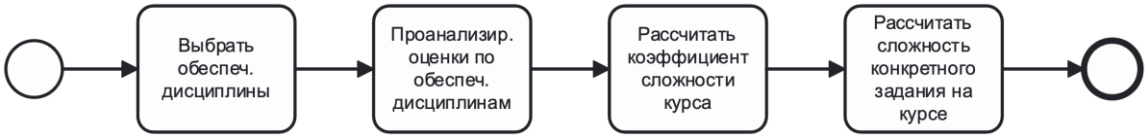

Рис. 1. Шаги процесса расчета сложности

Очевидно, что все шаги процесса могут быть частично или полностью автоматизированы, наибольшую сложность для автоматизации представляет только первый шаг, в котором могут потребоваться значительные аналитические усилия для подбора подходящих обеспечивающих дисциплин, тем не менее, даже этот шаг представляется скорее как разовое действие (составление матрицы связи дисциплин, по которым будет строиться прогноз сложности, и обеспечивающих дисциплин), возможно, пересматриваемое время от времени.

В процессе внедрения в систему образования новых федеральных государственных образовательных стандартов высшего образования (ФГОС 229 ВО) при формировании учебных планов и рабочих программ необходимо учитывать связи между изучаемыми дисциплинами. Отражаются эти связи с помощью таких понятий как «пререквизиты» и «постреквизиты» (дисциплины, обязательные для освоения до и после изучения данной дисциплины). В большинстве случаев пререквизиты и постреквизиты дисциплины указываются в учебном плане на усмотрение преподавателя без вычисления каких-либо взаимосвязей между ними. Следовательно, учебный план может не в полной мере отражать взаимозависимости между дисциплинами, что приведет к противоречивым оценкам, а в дальнейшем ─ к претензиям со стороны аудиторов системы менеджмента качества образования [12]. Коэффициент сложности может быть также полезен и при решении данной проблемы.

В целом, следует отметить, что оценка сложности курсов и конкретных заданий имеет первостепенное значение для построения эффективной индивидуальной образовательной траектории. Правильное определение уровня сложности позволяет выбирать для студентов оптимальные курсы и задания, которые соответствуют их текущему уровню знаний и способностей, а также способствуют их постепенному развитию и прогрессу. Это не только повышает мотивацию и вовлеченность обучающихся, но и обеспечивает более глубокое и прочное усвоение материала. Кроме того, адекватная оценка сложности позволяет преподавателям и разработчикам курсов создавать более сбалансированные и логически выстроенные программы обучения, в которых каждый последующий этап органично вытекает из предыдущего. Такой подход обеспечивает плавный переход от одного уровня к другому, предотвращая резкие скачки сложности и снижая риск потери интереса или демотивации учащихся. В конечном итоге, грамотное построение индивидуальной образовательной траектории на основе оценки сложности способствует более эффективному и гармоничному процессу обучения, максимизируя потенциал каждого студента.

Список источников

1. Агаев Ф. Т., Мамедова Г. А., Меликова Р. Т. Прогнозирование успеваемости студентов в электронном образовании с использованием методов data mining. // Информатизация образования и методика электронного обучения: цифровые технологии в образовании: Материалы V Международной научной конференции. В 2-х частях, Красноярск, 21-24 сентября 2021 года / Под общей редакцией М. В. Носкова. Том Часть 2. Красноярск: Сибирский федеральный университет, 2021. – С. 19-23. – EDN: BREKVL.
2. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data / C. Márquez, A. Cano, C.Romero, S. Ventura // Applied intelligence. – 2013. – Vol. 38, № 3. – P. 315–330. – DOI: 10.1007/s10489-012-0374-8.

3. Mgala, M., Mbogho A. Data-driven intervention-level prediction modeling for academic performance // Proceeding ICTD '15 Proceedings of the Seventh International Conference on Information and Communication Technologies and Development. –2015. – Article No. 2. – DOI: 10.1145/2737856.2738012.

4. Shahiri A. M., Husain W., R. N. Abdul. A Review on Predicting Student's Performance Using Data Mining Techniques // Procedia Computer Science. – 2015. – Vol. 72, No.14. – P. 414–422. – DOI: 10.1016/j.procs.2015.12.157

5. Цифровая образовательная история как составляющая цифрового профиля обучающегося в условиях трансформации образования / Р. В. Есин, Т. В. Зыкова, Т. А. Кустицкая, А. А. Кытманов // Перспективы науки и образования. 2022. – № 5 (59). – С. 566-584. – DOI: 10.32744/pse.2022.5.34.

6. Кустицкая Т. А., Носков М. В., Вайнштейн Ю. В. Прогнозирование успешности обучения: проблемы и задачи // Наука и школа. –2023. – № 4. – С. 71-83. – DOI: 10.31862/1819-463X-2023-4-71-83.

7. Аксентов В. А. Важность математики в программировании // Вестник науки. – 2023. – №2 (59). – С. 201-203. – URL: https://cyberleninka.ru/article/n/vazhnost-matematiki-v-programmirovanii (дата обращения: 20.03.2024).

8. Ерохина Е. А., Хруслова Д. В. Влияние результатов ЕГЭ на успеваемость студентов вузов // Информационные технологии в науке, образовании и управлении: Труды международной конференции IT + S&E'16 (Гурзуф, 22.05. – 01.06.2016). М.: ИНИТ, 2016. — С. 265-272.

9. Микитченко С. П., Разинкин В. Б. Английский язык в деятельности программиста // Психология и педагогика: методика и проблемы практического применения. – 2016. – №49-2. –С. 45-49. – URL: https://cyberleninka.ru/article/n/angliyskiy-yazyk-v-deyatelnosti-programmista (дата обращения: 16.03.2024).

10. Баранова М. В. О необходимости изучения английского языка студентами – будущими программистами. // Известия ПГПУ им. В. Г. Белинского. – 2011. – № 24. – С. 540-543.

11. Применение методов нечеткой логики для формирования адаптивной индивидуальной траектории обучения на основе динамического управления сложностью курса/ М.А. Белов, С. И. Гришко, А. В Живетьев [и др.] // Моделирование Оптимизация И Информационные Технологии. – 2022. – Том 10, № 4 (39). – С. 7–8.

12. Черняева Н. В. Модели и алгоритмы исследования корреляции между дисциплинами учебного плана специальности в вузах. // Актуальные проблемы современной науки: взгляд молодых: сборник трудов V Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых, 26 апреля 2016 г. / [науч. ред. О. С. Нагорная]. Челябинск: Violitprint, 2016. – С. 228-234.