УДК 004.942
DOI 10.17513/snt.40046

ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ УСПЕВАЕМОСТИ АБИТУРИЕНТОВ В ВУЗЕ

[1]Щербин С.И., [1]Харитонов И.М., [1]Огар Т.П., [1]Панфилов А.Э., [2]Кравец А.Г.

[1]Камышинский технологический институт (филиал) ФГБОУ ВО «Волгоградский государственный технический университет», Камышин, e-mail: asoiu@kti.ru;

[2] ФГБОУ ВО «Волгоградский государственный технический университет», Волгоград, e-mail: AllaGKravets@yandex.ru

Рассматривается проблема выбора будущей профессии выпускниками средней школы и воздействия данного выбора на успешность образования в высших учебных заведениях. В исследовании рассматривается вопрос о значимости осознанного выбора профессионального пути на начальной стадии образовательного процесса и его влиянии на дальнейшее развитие студентов в системе высшего образования. Авторами в проводимом исследовании ставится задача по улучшению точности прогнозирования успеваемости студентов в высшем учебном заведении (на примере Камышинского технологического института (филиала) Волгоградского государственного технического университета). Исследование направлено на разработку методов и подходов, способствующих более точной оценке возможной успеваемости студентов. Для достижения поставленной задачи применяются методы кластерного анализа и многомерной линейной регрессии на данных архивов приемной комиссии университета. В результате анализа архивных данных приемной комиссии университета делается вывод о том, что спектральный метод кластеризации обеспечивает высокую точность прогнозирования. Авторами отмечается, что применение кластерного анализа дает прирост в точности прогнозирования, что подчеркивает эффективность данного подхода для повышения надежности прогнозирования успеваемости студентов в высших учебных заведениях. Исследование показывает, что осознанный выбор профессии на начальном этапе образовательного процесса может являться одним из факторов повышения успеваемости студентов.

Ключевые слова: регрессионный анализ, кластеризация, прогнозирование, успеваемость, высшее образование, анализ данных

APPLICATION OF CLUSTER ANALYSIS METHODS IN THE TASK OF PREDICTING THE ACADEMIC PERFORMANCE OF APPLICANTS AT THE UNIVERSITY

[1]Shcherbin S.I., [1]Haritonov I.M., [1]Ogar T.P., [1]Panfilov A.E., [2]Kravets A.G.

[1]Kamyshin Technological Institute (branch of) Volgograd State Technical University, Kamyshin, e-mail: asoiu@kti.ru;

[2]Volgograd State Technical University, Volgograd, e-mail: AllaGKravets@yandex.ru

The problem of choosing a future profession by high school graduates and the impact of this choice on the success of education in higher education institutions is considered. The study examines the importance of a conscious choice of a professional path at the initial stage of the educational process and its impact on the further development of students in the higher education system. The authors of the study set the task of improving the accuracy of forecasting student academic performance in higher education (using the example of the Kamyshin Technological Institute (branch) Volgograd State Technical University). The research is aimed at developing methods and approaches that contribute to a more accurate assessment of students' possible academic performance. To achieve this task, methods of cluster analysis and multidimensional linear regression are used on the data of the archives of the university admissions committee. As a result of the analysis of the archival data of the university admissions committee, it is concluded that the spectral clustering method provides high prediction accuracy. The authors note that the use of cluster analysis gives an increase in forecasting accuracy, which emphasizes the effectiveness of this approach to increase the reliability of forecasting student academic performance in higher education institutions. Research shows that a conscious choice of profession at the initial stage of the educational process can be one of the factors in improving student performance.

Keywords: regression analysis, clustering, forecasting, academic performance, higher education, data analysis

После окончания средней школы выпускник встает перед проблемой выбора будущей профессии. Осознанность этого выбора является важным шагом для достижения высоких результатов в профессиональной сфере. На данный выбор может влиять мнение родителей выпускника, его знакомых, сложившаяся мода в обществе, пропаганда профессии в СМИ и т.д. Данные факторы не учитывают индивидуальные способности абитуриента, а также его интересы. Поэтому, с большой вероятностью, неосознанный

выбор неподходящей специальности сделает дальнейшее обучение в вузе малоэффективным. Актуальность данной проблемы доказывают различные исследования в этой области, описанные ниже.

Материалы и методы исследования

В работе [1] исследуется использование методов машинного обучения для анализа и прогнозирования образовательных результатов студентов первого курса (г. Волжский, Россия). Прогнозирование осуществляется с помощью методов машинного обучения для выявления групп студентов с высоким риском возникновения академической задолженности. Исследование показало, что точность прогнозирования сдачи экзамена по дисциплине приемлема как на этапе первого, так и на этапе второго контрольного среза. Основные переменные, использованные для моделирования, включали результаты вступительных испытаний, академическую успеваемость и посещаемость занятий.

Коллеги из университета Амити (Индия) в работе [2] выделяют пять основных показателей эффективности: удовлетворенность выпускников, студентов, работодателей, занятость выпускников и уровень знаний выпускников. Целью данного исследования ставится успешное трудоустройство выпускников и удовлетворение потребностей работодателей. Авторы стремятся охватить всестороннее представление об успеваемости студентов, а также выяснить детали своевременной успеваемости студентов. Объект исследования – 300 студентов, генерация кластеров производится программой MATLAB. Выделенные кластеры – это меры, основанные на оценках, навыках и т.д. Выбранный метод кластеризации – метод нечеткого с-среднего. В процессе исследования исходные данные разделены на 3 кластера. По результатам данного исследования авторы делают вывод о возможности предсказать будущее трудоустройство выпускников, а также о выявлении студентов, которым требуется больше внимания от преподавателей.

Коллегами из Испании в работе [3] проведено исследование эффективности онлайн-обучения в среде Moodle (система управления курсами), для дальнейшего предсказания успеваемости студентов. Причиной исследования указывается резкий переход на дистанционную форму обучения, в связи с вирусом COVID-19. Преподаватели не могли предугадать вовлеченность студента в изучение дисциплин, а, следовательно, не могли предсказать его возможные результаты.

Данная работа представляет собой набор моделей для раннего прогнозирования успеваемости студентов. Эти модели построены на основе данных взаимодействия 802 студентов полного онлайн-обучения. В результате работы моделей авторов выявлены основные факторы, влияющие на прогноз успешности обучения: возможность доступа к курсу, количество обращений и результаты опросов по теме, количество обсуждений и выполнений задач, возрастной фактор. Фактор возраста является отрицательным предиктором успеваемости, что в свою очередь означает обратную зависимость между возрастом и успеваемостью. Авторы выделили 5 групп студентов и подтвердили, что количество взаимодействий студентов с Moodle тесно связано с их оценками.

Раннее прогнозирование успеваемости студентов считают важным также и авторы работы [4]. Статья посвящена анализу и прогнозированию академической успеваемости студентов с помощью методов Data Mining. Целью исследования является построение моделей прогнозирования академической успеваемости студентов, принимаются во внимание их демографические характеристики, особенности зачисления и учебной деятельности. Исследование базируется на данных, собранных из университетов г. Пенза. Данные включают информацию о зачислении студентов, а также их активности в электронной информационно-образовательной среде (ЭИОС). Для анализа использовались различные методы классификации, такие как наивный байесовский классификатор и метод опорных векторов (SVM).

Эксперименты подтвердили, что модели, обученные на данных подгрупп студентов, показывают лучшие результаты по сравнению с моделями, обученными на всех данных. Учет особенностей зачисления и учебной деятельности позволяет более точно выявлять студентов, подверженных риску неуспеваемости.

Целью работы является повышение качества прогнозирования возможной успеваемости студентов в вузе. Данное прогнозирование направлено на повышение эффективности выбора абитуриентом будущего направления обучения. Для достижения поставленной цели производится поиск взаимосвязи между результатами обучения в школе (далее используется термин «Предикторы») и в вузе (далее используется термин «Критериальная переменная»).

Способом решения задачи выбрана линейная регрессионная модель. Для повышения точности прогнозирования исходные данные делятся по группам с внутренними схожими признаками с помощью следую-

щих методов кластерного анализа: метод k-средних, иерархический метод, спектральный метод [5; 6].

Научная новизна работы заключается в применении статистических методов анализа данных для слабоформализованной задачи прогнозирования будущей успеваемости студентов.

Результаты исследования и их обсуждение

Алгоритм решения задачи

Для построения модели предсказания успешности обучения в вузе использовались статистические данные, полученные из архивов приемной комиссии Камышинского технологического института (филиала) Волгоградского государственного технического университета за 5 лет. Важно отметить, что в данном исследовании делается акцент на объективные факторы, такие как результаты аттестаций и ЕГЭ, не принимая во внимание личностные характеристики преподавательского состава и другие субъективные аспекты образовательного процесса. Это позволяет сосредоточиться на количественно измеримых данных, обеспечивая таким образом объективность и воспроизводимость результатов. В исходные данные модели предсказания успеваемости входили:

1. Аттестационная оценка по математике.
2. Аттестационная оценка по геометрии.
3. Аттестационная оценка по физике.
4. Результаты основного государственного экзамена (ОГЭ) по математике.
5. Результаты единого государственного экзамена (ЕГЭ) по физике.
6. Результаты единого государственного экзамена (ЕГЭ) по математике.
7. Усредненная оценка аттестата о среднем общем образовании по всем предметам.

В качестве модели связи между предикторами и критериальной переменной использовалась многомерная линейная регрессионная модель. Выбор данной модели обусловлен несколькими факторами:

1. Простота интерпретации результатов: линейная регрессия обладает преимуществом в виде наглядности и простоты интерпретации коэффициентов, что важно для первоначального понимания зависимостей между образовательными показателями и успеваемостью студентов.

2. Наличие предварительных данных: предварительный анализ данных показал, что основные предикторы (школьные оценки и результаты ЕГЭ) демонстрируют линейные зависимости с итоговой успеваемостью студентов, что делает использование линейной регрессии адекватным выбором для начального этапа исследования.

3. Оценка значимости переменных: линейная регрессия позволяет оценить значимость каждого предиктора отдельно, что является важным аспектом при анализе влияния различных учебных и социальных факторов на успеваемость.

Однако, несмотря на применение линейной регрессии, авторы осознают потенциальные ограничения данного метода, связанные с игнорированием возможной мультиколлинеарности и взаимодействий между предикторами. В связи с этим в дальнейшем планируется разработка более сложной статистической модели, которая позволит учитывать множественные взаимодействия между переменными и нелинейности в данных. Этот этап будет включать использование многомерных регрессионных моделей, методов регуляризации, а также возможное применение машинного обучения для более точного и комплексного анализа влияния образовательного процесса на успеваемость студентов.

В ранних исследованиях, описанных в [7], применение линейной регрессии на всей выборке данных показало точность прогноза будущей успеваемости, равной 72%. Данное значение считается удовлетворительным, поэтому для увеличения точности прогноза произведено предварительное разделение выборки на группы с помощью методов кластерного анализа.

Пример реализации предлагаемого алгоритма

Алгоритм прогнозирования успеваемости состоял из двух фаз:

1. Фаза 1: анализ исходных данных и распределение значений по кластерам.

2. Фаза 2: расчет регрессионной статистики для каждого кластера и анализ результатов.

В первой фазе работы производилась кластеризация исходных данных. На данном этапе возможно определение внутренних признаков распределения студентов по различным кластерам (в исследовании проводилось разделение по 2, 3, 4 и 5 кластерам). Указанное количество кластеров обосновывается имеющимся количеством исходных данных для анализа, а также удобством логического обоснования принципов распределения.

Во второй фазе производился расчет регрессионных статистик, показывающих степень возможности успешного прогнозирования будущих данных. В регрессионную статистику входят следующие показатели:

1. R-квадрат.
2. Нормированный R-квадрат.

R-квадрат – коэффициент детерминации, трактующийся следующим образом: регрессионная модель в n% случаев объясняет зависимость между изучаемыми объектами. Чем выше данный коэффициент, тем качественнее модель. Приемлемое качество модели при значениях данного коэффициента выше 80%. Данный параметр может показывать завышенное значение при малом количестве данных для анализа. Для снижения завышенного значения использовался параметр – нормированный R-квадрат.

Далее следует работа описанного алгоритма, проделанного над данными, разделенными спектральным методом на 3 кластера.

Первым этапом набор данных был разделен на 3 кластера. Каждый из них можно логически трактовать таким образом:

- «слабоуспевающие» – студенты с низкими баллами;
- «хорошисты» – студенты, имеющие стабильно средние баллы по всем предметам;
- «отличники» – студенты, имеющие высокие баллы по всем предметам.

Значения R-квадрата и нормированного R-квадрата для трех кластеров приведены в таблице 1.

На основе полученных данных был сделан вывод, что качество прогнозирования увеличилось. В среднем коэффициенты выше 0.8, значит прогноз соответствует требованиям к качеству.

Таблица 1

Регрессионные статистики исследуемых кластеров

№	R-квадрат	Нормированный R-квадрат
Кластер 1	0,88	0,81
Кластер 2	0,84	0,81
Кластер 3	0,82	0,80

Помимо этого, проанализированы полученные коэффициенты зависимости прогноза от тех или иных предикторов. Полученные коэффициенты представлены в таблице 2.

В данном случае величина влияния предиктора на результат определяется модулем полученного коэффициента. Вывод – наибольшее влияние имеет средняя оценка по аттестату, значительное влияние оказывают итоговые оценки по математике и в меньшей степени по геометрии и физике, а результаты по единому государственному экзамену имеют слабое влияние.

Полученные в результате всего исследования усредненные показатели регрессионной статистики представлены в таблице 3.

Таблица 2

Значения полученных коэффициентов

Предиктор	1 кластер	2 кластер	3 кластер
Алгебра	9	3,20	3,49
Геометрия	-2,28	1,04	3,41
Физика	-1,79	-3,23	-1,71
Физика (ЕГЭ)	-0,54	0,63	0,22
Математика (ЕГЭ)	-0,07	0,29	-0,19
Математика (ОГЭ)	5,54	2,60	-0,13
Средняя оценка по аттестату	-22,2	-6,38	-8,08

Таблица 3

Сводная таблица по всем методам

Количество кластеров	Min R-квадрат, %	Max R-квадрат, %	Средний R-квадрат, %	Min нормированный R-квадрат, %	Max нормированный R-квадрат, %	Средний нормированный R-квадрат, %
K-средних	86,41	87,31	86,86	79,16	84,41	81,785
Иерархический	84,71	100	91,07	65,53	81	75,51
Спектральный	87,17	100	95	65,53	92,12	76

Заключение

В результате анализа полученных данных авторами сделан вывод, что, в среднем, спектральный метод предлагает разбиение на кластеры, точность предсказания для которых составляет более 90%. Для решения поставленной задачи данный метод кластеризации подходит в большей степени по сравнению с другими, рассмотренными выше.

При сравнении значения регрессионной статистики с полученными по необработанному набору данных отмечено, что имеется прирост значения коэффициента детерминации в 23%. Вывод – применение методов кластерного анализа для предварительной подготовки анализируемых данных позволяет увеличить точность дальнейшего прогнозирования.

Полученные в ходе исследования результаты показывают относительно небольшое влияние школьной подготовки по математике и физике на успеваемость студентов, что может быть связано со спецификой выборки абитуриентов факультета информационных технологий. Однако необходимо провести дополнительные исследования с использованием более обширных данных для подтверждения этих выводов и оценки их обобщаемости.

Список литературы

1. Алпатов А.В. Применение машинного обучения для анализа образовательных результатов студентов вузов // Информационные и математические технологии в науке и управлении. 2023. № 4(32). С. 67-78. DOI: 10.25729/ESI.2023.32.4.006.

2. Singh Ishwank, Sabitha Sai, Choudhury Tanupriya, Aggarwal Archit, Dewangan Kumar Bhupesh. Mapping Student Performance with Employment Using Fuzzy C-Means // International Journal of Information System Modeling and Design. 2020. Vol. 11. P. 36-52.

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

4. Егорова Е.С., Попова Н.А. Data Mining в образовании: прогнозирование успеваемости учащихся // Моделирование, оптимизация и информационные технологии. 2023. № 11(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1325 (дата обращения: 05.04.2024). DOI: 10.26102/2310-6018/2023.41.2.003.

5. Ibanez Alfonso, Larranaga Pedro, Bielza Concha. Cluster methods for assessing research performance: exploring Spanish computer science // Scientometrics. 2013. Vol. 97. P. 571-600.

6. Гранков М.В., Аль-Габри В.М., Горлова М.Ю. Анализ и кластеризация основных факторов, влияющих на успеваемость учебных групп вуза // Инженерный вестник Дона. 2016. № 4. URL: ivdon.ru /ru/magazine/ archive/n4y2016/3775. (дата обращения: 05.04.2024).

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