УДК 004.89+004.048
DOI: 10.25729/ESI.2023.32.4.006

Применение машинного обучения для анализа образовательных результатов студентов вузов

Алпатов Алексей Викторович
Волжский политехнический институт (филиал) Волгоградского государственного технического университета, Волжский, Россия, *alpatov80@mail.ru*

Аннотация. В работе представлены результаты анализа и прогнозирования образовательных результатов студентов первого курса вуза при реализации отдельной дисциплины с использованием машинного обучения. Актуальность темы исследования обусловлена необходимостью вузов в современных условиях успешно конкурировать на рынке образовательных услуг, который характеризуется низким количеством абитуриентов и увеличением требований к качеству профессионального образования, как со стороны абитуриентов, так и со стороны государства. Важной составляющей для эффективного принятия решения в процессе управления качеством образовательного процесса является учебная аналитика, на основе которой можно выполнять прогнозирование академической успеваемости студентов, выявлять факторы, оказывающие существенное влияние на достижение высоких образовательных результатов. Исследование показало возможность прогнозирования сдачи экзамена по отдельной дисциплине студентов первого курса вуза на основе данных контрольных срезов, которые проводят деканаты в течение семестра для выявления групп студентов с повышенным риском возникновения академической задолженности. Точность прогнозирования, которую показали построенные модели (наивный байесовский классификатор и логистическая регрессия) оказалась вполне приемлемой, как на этапе проведения первого рубежного контроля, так и на этапе проведения второго. Результаты данной работы имеют практическое значение для администрации вузов и для преподавателей. Прогнозные модели можно использовать при прогнозировании отчисления студентов вследствие академической неуспеваемости. Модели могут быть встроены в образовательные информационные системы и быть помощником преподавателям для принятия решений в процессе реализации дисциплины.

Ключевые слова: прогнозирование образовательных результатов, учебная аналитика, интеллектуальный анализ образовательных данных, единый государственный экзамен, нейронные сети, дерево решений, логистическая регрессия, кластеризация

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

Введение. В сфере профессионального образования одной из приоритетных задач является поддержание высокого уровня образования с сохранением достаточно большой доли выпускников среди поступивших в учебное заведение. Это особенно актуально в современных условиях российского рынка образования, который характеризуется низким количеством абитуриентов и увеличением требований к качеству профессионального образования, как со стороны абитуриентов, так и со стороны государства. Важной составляющей для эффективного принятия решения в процессе управления качеством образовательного процесса является учебная аналитика, на основе которой можно выполнять прогнозирование академической успеваемости студентов, выявлять факторы, оказывающих существенное влияние на достижение высоких образовательных результатов.

Анализ научных публикаций, посвященных интеллектуальному анализу академических успехов студентов в высших учебных заведениях, показал, что существенное внимание уделено вопросам прогнозирования отчисления студентов [1, 2], выявления группы риска студентов образования академической задолженности по отдельной дисциплине [3, 4, 5], а также прогнозирования величины среднего балла [6]. Факторы, которые исследователи выбирают для построения прогнозных моделей академической успеваемости, можно разделить на такие группы, как образовательные, материальные, социальные, интеллектуальные и мотивационные. Ряд исследований посвящен изучению влияния различных факторов на образовательные

результаты. Например, воздействие социальных и демографических характеристик рассматривается в статье [7], в публикациях [8, 9] представлены результаты исследования по влиянию мотивации студентов на их образовательные результаты.

С развитием концепции индивидуальных образовательных траекторий, а также широким внедрением электронных образовательных платформ одним из направлений научных исследований стало прогнозирование образовательных траекторий отдельных студентов [4, 10, 11, 12]. В частности, в работе [11] рассматривался вопрос построения моделей прогнозирования успешности по дисциплине, которые позволяют в динамике «выявлять студентов с повышенным риском не аттестации по дисциплине». Прогнозирование осуществлялось на основе еженедельных данных об успеваемости студентов и активности в электронной образовательной среде с использованием различных методов машинного обучения. Наилучшую точность при прогнозировании показала модель Ансамбль –усреднение классификаторов Random Forest, XGBoost и Logistic Regression с регуляризацией L1. При этом в работе отмечено, что «данные цифрового следа уже в первой половине семестра позволяют выявлять студентов с высоким риском не успешности по дисциплине».

В настоящее время некоторые российские вузы вводят информационные образовательные системы для прогнозирования образовательных результатов [11]. Однако у большинства учебных заведений обычной практикой при анализе успеваемости студентов в течение семестра является проведения одного или двух контрольных срезов (рубежных контролей). Как правило, в этих срезах указываются данные о набранных баллах и о количестве пропусков. Эти сведения по всем дисциплинам передаются в деканат, объединяются и на основе агрегатных статистических показателей деканат выявляет студентов, которые находятся в группе риска на образование академической задолженности.

Целью данной работы является исследование возможности использования данных, полученных на контрольных срезах, для анализа и прогнозирования образовательных результатов студентов первого курса с помощью методов машинного обучения. Прогнозирование академических успехов обучающихся представляет собой задачу бинарной классификации. В качестве предиктной переменной выступает дамми-переменная, которая принимает значение 1, если студент сдал экзамен на любую положительную оценку и 0, если студент не сдал экзамен. Еще одной задачей в данной работе является выявление на основе кластерного анализа групп студентов со схожими итоговыми образовательными результатами по отдельной дисциплине.

1. Данные и методы. Для построения прогностических моделей были использованы данные, которые представляли собой результаты контрольных срезов по дисциплине «Основы программирования» за первый семестр студентов очной формы обучения направлений подготовки 09.03.01 Информатика и вычислительная техника, а также 09.03.09 Программная инженерия. Моделирование осуществлялось на данных о 129 студентах, которые поступили в Волжский политехнический институт в 2020, 2021 и 2022 году.

Для обучения моделей использовались также результаты вступительных испытаний студентов при поступлении в Институт. В качестве вступительного испытания может быть принят результат Единого государственного экзамена (ЕГЭ), если студент поступил на базе среднего общего образования или результат экзамена, проводимого Институтом, если студент поступил на базе среднего профессионального образования. В большинстве исследований, в которых рассматривается вопрос прогнозирования образовательных результатов российских студентов, основным фактором влияния на успешность обучения являются результаты ЕГЭ [6; 13]. Однако, из ряда исследований, например, [14], известно, что результаты ЕГЭ наиболее сильное влияние оказывают на первых курсах обучения, а в дальнейшем эта связь уменьшается. В данной работе анализируются образовательные данные студентов первого курса, поэтому результаты вступительных испытаний предположительно будут оказывать наибольшее

влияние на успешность освоения учебных дисциплин, и их можно рассматривать как своего рода потенциал обучающегося.

Для мониторинга ритмичной работы студентов деканат в Институте проводит в течение семестра два контрольных среза (рубежных контроля) на шестой и двенадцатой неделях графика обучения. Данные о набранных баллах и количестве пропущенных студентами аудиторных часов, которые преподаватели передают в деканат, представлены в виде относительных величин. В таблице 1 представлены обозначения переменных, используемых при моделировании образовательных результатов, и их описание.

Таблица 1. Описание и обозначения переменных

Обозначение	Название переменной	Описание переменной
Exam	Результат сдачи экзамена по дисциплине	Бинарная переменная, которая принимает значение 1, если студент сдал экзамен и 0, если студент его не сдал
Group	Направление подготовки	Бинарная переменная, которая принимает значение 1, если студент обучается по направлению подготовки 09.03.09 Программная инженерия и 0, если студент обучается по направлению подготовки 09.03.01 Информатика и вычислительная техника
Sex	Пол студента	Бинарная переменная, которая принимает значение 1, если пол мужской и 0, если пол женский
Type_exam	Вид вступительных испытаний в институт	Бинарная переменная, которая принимает значение 1, если абитуриент был зачислен в институт по результатам ЕГЭ и 0, если он поступил в институт на основе экзамена, проводимого вузом
Entrance_ exam	Результаты вступительного экзамена	Переменная *Entrance_exam* представляет собой отношение суммы баллов за три вступительных испытания к их максимально возможной сумме, т.е. к 300
*Academ*1	Успеваемость к первому рубежному контролю	Доля баллов, набранных на этапе первого рубежного контроля, от максимально возможного количества баллов к данному этапу
Academ2	Успеваемость ко второму рубежному контролю	Доля баллов, набранных на этапе второго рубежного контроля, от максимально возможного количества баллов к данному этапу
*Attend*1	Посещаемость к первому рубежному контролю	Отношение количества посещений учебных занятий студентом к числу аудиторных занятий, проведенных на момент первого рубежного контроля
Attend2	Посещаемость ко второму рубежному контролю	Отношение количества посещений учебных занятий студентом к числу аудиторных занятий, проведенных на момент второго рубежного контроля
AcademF	Итоговая успеваемость за первый семестр	Доля баллов, набранных в течение первого семестра от максимально возможного
AttendF	Итоговая посещаемость за первый семестр	Отношение количества посещений учебных занятий студентом за первый семестр к числу аудиторных занятий, проведенных в течение семестра

Прогнозирование осуществляется дважды: на момент проведения первого и второго рубежного контроля. Показатели, характеризующие успеваемость и посещаемость, аккумулируют все другие возможные факторы, которые формируют знания по дисциплине. В связи с этим нет необходимости проведения дополнительного тестирования, сбора данных, которые касаются личности студента и его деятельности в процессе обучения.

При построении прогностических моделей широко используются методы машинного обучения. Как правило, исследователи строят несколько моделей и сравнивают точность прогнозирования. Например, в работе [15] применялись модели дерева решений (Decision Tree), случайный лес (Random Forest), наивный Байес (Naive Bayes), а также индукция правил (Rule induction). В целом наилучший результат показал метод дерева решений. В рамках проводимого исследование прогнозирование в задаче классификации осуществлялось на основе двух моделей: наивный байесовский классификатор и логистическая регрессия. Выбор был обусловлен тем, что объем выборки небольшой и наилучшие результаты при обучении будут давать простые модели. Для выявления закономерностей при формировании результатов обучения за первый семестр был проведен кластерный анализ с использованием иерархического подхода. При моделировании использовались библиотеки Python: Keras, sklearn, scipy. При проверке статистических гипотез применялись тесты из модуля stats библиотеки scipy.

2. Разведочный анализ данных. Проведем статистический анализ набора данных, который будет использован при прогнозировании. Анализируемый набор данных содержит 52,7% обучающихся по направлению подготовки «Программная инженерия» и, соответственно, 47,3% по направлению подготовки «Информатика и вычислительная техника». При этом, доля студенток составляет 13,2%, доля поступивших на базе среднего профессионального образования 7,0 %. По итогу освоения дисциплины в первом семестре в среднем экзамен сдают 59,7% студентов обоих направлений подготовки. Таким образом, в классах присутствует дисбаланс, но он незначительный.

В таблице 2 приведены основные описательные статистики для количественных показателей. Значения статистических показателей переменной *Entrance_exam* представлены в процентах и в баллах. Средний балл вступительных испытаний при поступлении на специальности «Программная инженерия» и «Информатика и вычислительная техника» составляет 193,5 со среднеквадратическим отклонением 24,0. Крайние значения изменяются от 127,0 до 264,0 баллов. На рис. 1 показана гистограммы распределения результатов вступительных испытаний студентов, представленных в датасете для двух групп студентов: сдавших экзамен и не сдавших его. Визуально распределение для сдавших экзамен близко к нормальному. Распределение для *Exam* = 0 ограниченно значением *Entrance_exam* = 0,74. Это означает, что студенты, которые набрали примерно более 222 баллов, всегда сдают экзамен вовремя.

Таблица 2. Описательная статистика

Показатель	***Entrance_exam***		***Academ1***	***Attend1***	***Academ2***	***Attend2***
	доля	**балл**				
mean	0,645	193,5	0,522	0,826	0,519	0,758
std	0,080	24,0	0,242	0,268	0,254	0,279
min	0,423	127,0	0,000	0,000	0,000	0,000
25%	0,593	178,0	0,375	0,800	0,353	0,667
50%	0,643	193,0	0,538	0,933	0,529	0,857
75%	0,700	210,0	0,706	1,000	0,711	0,966
max	0,880	264,0	1,000	1,000	0,974	1,000

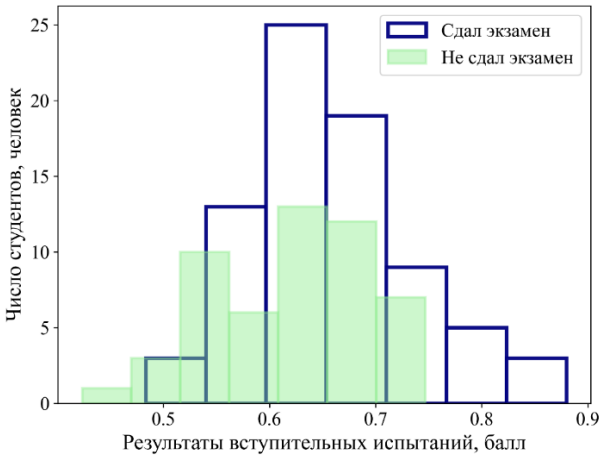

Рис. 1. Распределения результатов вступительных испытаний студентов, поступивших на направления подготовки «Программная инженерия» и «Информатика и вычислительная техника»

В таблице 3 показаны средние значения mean и среднеквадратические отклонения std переменной *Entrance_exam* для различных классов студентов. Можно отметить, что средний балл студенток, поступивших в вуз, выше, чем у студентов (203,1 и 192,1, соответственно). Проверка гипотезы о равенстве средних с помощью *t*-теста показала, что нулевая гипотеза отклоняется на уровне значимости 0,05 (*P*-значение = 0,039). При проверке данной гипотезы предполагалось, что обе выборки имеют распределения, близкие к нормальному. Поскольку количество студенток небольшое, то для проверки гипотезы о нормальности использовался критерий Шапиро. *P*-значение оказалось равным 0,064 и нулевая гипотеза не отклоняется, если уровень значимости принять равным 0,05. Для проверки гипотезы о равенстве средних использовался критерий Флигнера-Килина (*P*-значение = 0,743). На основе результатов вступительного экзамена можно сделать предположение о том, что базовые знания у девушек первого курса в целом выше, чем у юношей (однако данное предположение не является в достаточной степени надежным). Среднее значение баллов за вступительные испытания у студентов направления подготовки «Программная инженерия» существенно выше, чем у студентов, обучающихся по направлению «Информатика и вычислительная техника». Проведенный *t*-тест о равенстве средних подтвердил данное различие (*P*-значение < 0,001). Таким образом, уровень подготовки студентов, поступивших на направление подготовки «Программная инженерия», превышает уровень подготовки тех, кто поступил на направление «Информатика и вычислительная техника». Среднее значение *Entrance_exam* для сдавших экзамен по дисциплине, ожидаемо выше почти на 11 баллов по сравнению с теми, кто не сдал экзамен.

Таблица 3. Результаты вступительного испытания, балл

Показатель	***Sex***		***Group***		***Exam***	
	1	***0***	***1***	***0***	***1***	***0***
mean	192,1	203,1	199,9	186,5	197,9	187,0
std	23,7	24,3	21,8	24,4	24,2	22,3

В таблице 4 представлены значения коэффициентов корреляции между различными переменными анализируемого набора данных. Рассмотрим линейную корреляцию между прогнозируемым признаком *Exam* и факторными признаками. Наиболее сильное влияние оказывают показатели, характеризующие академическую успеваемость и посещаемость учебных занятий. При этом для второго рубежного контроля эта связь выражена в большей степени, вследствие близости промежуточной аттестации по дисциплине. Корреляции *Exam* с переменной *Sex* слабая и отрицательная, что может свидетельствовать о том, что студентки имеют не намного больше шансов сдать экзамен. Линейная связь между прогнозируемой переменной и

переменной *Type_exam* практически отсутствует. Проверка статистической значимости коэффициента корреляции показала, что нулевая гипотеза на уровне значимости 0,01 не отклоняется. В связи с этим данная переменная была исключена из датасета.

Таблица 4. Матрица коэффициентов корреляции

	Group	***Sex***	***Type_ exam***	***Entrance_ exam***	***Academ1***	***Attend1***	***Academ2***	***Attend2***	***Exam***
Group	1	-0,185	0,167	0,279	0,277	0,100	0,213	0,197	0,330
Sex	-0,185	1	-0,107	-0,155	-0,098	-0,123	-0,069	-0,157	-0,18
Type_ exam	0,167	-0,107	1	0,042	0,054	0,036	-0,065	0,075	-0,039
Entrance_ exam	0,279	-0,155	0,042	1	0,036	0,027	0,119	0,128	0,224
Academ1	0,277	-0,098	0,054	0,036	1	0,798	0,792	0,743	0,59
Attend1	0,100	-0,123	0,036	0,027	0,798	1	0,716	0,887	0,523
Academ2	0,213	-0,069	-0,065	0,119	0,792	0,716	1	0,847	0,679
Attend2	0,197	-0,157	0,075	0,128	0,743	0,887	0,847	1	0,672
Exam	0,330	-0,180	-0,039	0,224	0,59	0,523	0,679	0,672	1

Коэффициенты корреляции между посещаемостью и успеваемостью довольно высокие, что является вполне закономерным, поскольку высокий уровень посещаемости, как правило, свидетельствует о высокой мотивации студентов к получению знаний и навыков по дисциплине «Основы программирования». Кроме того, основная часть баллов, набирается при проведении аудиторных занятий. Включение коррелирующих факторов, например, *Academ1* и *Attend1*, в линейную регрессионную модель приведет к появлению мультиколлинеарности. Поскольку корреляция с прогнозируемой переменной ***Exam*** выше у тех переменных, которые показывают долю набранных баллов к рубежному контролю, то в прогнозные модели будут включаться только *Academ1* и *Academ2*, а переменные *Attend1* и *Attend2* не будут включены. Связь между прогнозируемой переменной и переменной *Entrance_exam* довольно слабая и составляет 0,224. Это обусловлено несколькими причинами. Во-первых, в качестве вступительных испытаний принимались результаты ЕГЭ или результаты внутреннего экзамена Института. Базовой дисциплиной ЕГЭ может быть одна из трех дисциплин: информатика, физика или английский язык. Поэтому по профильным дисциплинам у студентов базовый уровень подготовки может отличаться. Во-вторых, студенты первого курса имеют разный опыт программирования до поступления в вуз.

3. Прогнозные модели. При прогнозировании переменной *Exam* на момент проведения первого рубежного контроля в качестве факторных признаков были использованы переменные *Group, Sex, Entrance_exam, Academ*1. На этапе прохождения второго рубежного – *Group, Sex, Entrance_exam, Academ*2.

Для сравнения были получены разные модели: наивный байесовский метод и логистическая регрессия. При обучении моделей данные делились на обучающую и тестовую выборки. Доля данных в тестовой выборке составляла 20 %.

Рассмотрим вопрос выбора метрики для оценки качества полученных моделей. При прогнозировании бинарной переменной возможны ошибки двух типов: ошибки первого рода (False Positive) и ошибки второго рода (False Negative). В рассматриваемой задаче ошибка первого рода заключается в неверном предсказании того, что студент сдаст экзамен по дисциплине. Соответственно, ошибка второго рода состоит в неверном предсказании того, что студент не сдаст экзамен. В таблице 5 показана матрица ошибок для задачи прогнозирования сдачи экзамена по дисциплине.

Таблица 5. Матрица ошибок

		Фактическое значение *FY*	
		1	0
Прогнозное значение *FY*	1	True Positive (*TP*) Верный прогноз, что студент сдаст экзамен	False Positive (*FP*) Неверный прогноз, что студент сдаст экзамен
	0	False Negative (*FN*) Неверный прогноз, что студент не сдаст экзамен	True Negative (*TN*) Верный прогноз, что студент не сдаст экзамену

Для оценки качества работы алгоритмов в задачах бинарной классификации часто используются такие метрики, как меткость *Accuracy*, точность *Precision* и полнота *Recall*.

Precision можно интерпретировать, как долю студентов, для которых модель сделала верный прогноз о сдаче экзамена. Чем выше значение *Precision*, тем ниже доля ошибок первого рода. Метрика *Recall* показывает долю верно прогнозируемых сдач экзамена среди всех фактически сдавших экзамен студентов. Чем выше *Recall*, тем ниже доля ошибок второго рода. Метрика *Accuracy* актуальна в том случае, когда оба класса имеют одинаковое значение для исследователя. В рассматриваемой задаче наиболее важным является контролировать ошибки первого рода, поскольку в этом случае последствия для процесса управления численностью контингента будут более негативными. Ведь в этом случае заведомо неуспевающий студент будет признан успевающим и на него не будет оказано вовремя воспитательное воздействие. Безусловно, желательно не оставлять без внимания и ошибки второго рода, поскольку при наличии большого числа обучающихся преподаватель будет тратить существенное количество времени для мотивации студентов, которые являются успевающими. Лучше сконцентрировать больше внимания и усилий для работы с действительно отстающими.

С учетом данных аргументов для сравнения качества моделей была выбрана *F*-мера с β −коэффициентом равным 0,5:

$$F_\beta = (1+\beta^2)\cdot\frac{Precision\cdot Recall}{(\beta^2\cdot Precision)+Recall}$$

В таблице 6 представлены показатели качества моделей прогнозирования, *Precision* и *F*-мера для тестовой выборки. Лучшие результат показала модель «дерево решений». Можно отметить, что прогнозирование уже на первом рубежном контроле имеет приемлемую точность. Это дает возможность на ранних этапах обучения выявлять отстающих студентов. При обучении на данных, полученных на втором рубежном контроле, модели оказались немного точнее, что объясняется близостью окончания семестра.

4. Кластерный анализ. Для выявления закономерностей в результатах обучения студентов за первый семестр был проведен иерархический кластерный анализ. Использовались характеристики объектов: *Academ*1, *Attend*1, *Academ*2, *Attend*2, *Sex*, *Exam*, *Group*, итоговая успеваемость за первый семестр *AcademF*, итоговая посещаемость за первый семестр *AttendF*, а также бинарная переменная *Y*.

Таблица 6. *F*-мера ($\beta = 0.5$) различных моделей прогнозирования на тренировочной и тестовой выборках

Модель	**Первый рубежный контроль**		**Второй рубежный контроль**	
	Тренировочная	**Тестовая**	***Тренировочная***	**Тестовая**
Наивный байесовский классификатор	0.83	0.82	0.85	0.89
Логистическая регрессия	0,82	0,84	0,85	0,88

В процессе выявления кластеров комбинировались различные метрики оценки расстояний между объектами и метрики расчета расстояний между кластерами. Анализировались различные варианты разбиения объектов на группы. Было принято решение остановиться на 6 кластерах, поскольку при таком делении результаты кластеризации можно проинтерпретировать наилучшим образом. Данное разбиение было получено при выборе метода Варда (ward) для оценки степени сходства объектов евклидового расстояния (euclidean) и для оценки расстояний между кластерами. На рис. 2 показана дендрограмма, которая иллюстрирует обоснованность разделения студентов по образовательным результатам на шесть групп.

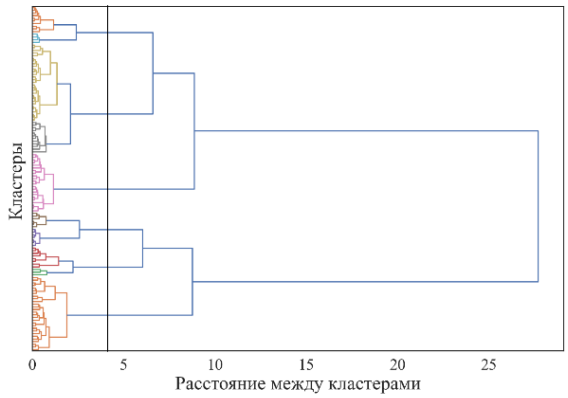

Рис. 2. Дендрограмма

В таблице 7 приведены средние характеристики кластеров. Количество объектов по кластерам распределено неравномерно: изменяется от 11 до 41. Студенты, которые сдали экзамен по дисциплине и студенты, которые его не сдали, образуют по три кластера. Среди тех, кто сдал экзамен, средний уровень посещаемости заметно выше. При этом эффективность на занятиях у них также более высокая, о чем свидетельствуют отношения *AcademF*/*AttendF*, которые имеют значения около 90%. Посещаемость в более поздних контрольных срезах уменьшается, а доля набранных баллов, напротив, увеличивается. В кластерах, в которых студенты сдали экзамен, средний результат на вступительных испытаниях выше, чем в кластерах со студентами, которые не сдали экзамен по дисциплине.

Таблица 7. Средние показатели кластеров

Номер кластера	Число объектов в кластере	*Exam*	*Group*	*Sex*	*Entrance_exam*	Первый РК		Второй РК		Итог за первый семестр		
						Academ1	*Attend1*	*Academ2*	*Attend2*	*AcademF*	*AttendF*	*AcademF/ AttendF*, %
1	28	0	0,000	1,000	183,2	0.423	0.823	0.393	0.662	0.386	0.534	72,2
2	11	0	1,000	0,727	197,8	0.500	0.828	0.408	0.705	0.479	0.655	73,1
3	13	0	0,461	1,000	186,2	0.061	0.152	0.048	0.099	0.026	0.071	36,0
4	22	1	0,000	1,000	189,3	0.601	0.961	0.666	0.913	0.786	0.888	88,6
5	41	1	1,000	1,000	200,4	0.659	0.924	0.675	0.916	0.801	0.886	90,3
6	14	1	0,714	0,000	204,2	0.639	0.957	0.610	0.893	0.796	0.852	93,5

Первый кластер (рис. 3), среди образовавших академическую задолженность студентов, – самый многочисленный и представлен только направлением подготовки «Информатика и вычислительная техника».

Второй кластер – это студенты-задолжники, которые обучались по направлению «Программная инженерия». Доля итоговой успеваемостью у второго кластера выше, чем у первого, при одинаковой эффективности работы на занятиях. Третий кластер – студенты с

крайне низкой посещаемостью и академической успеваемостью, которые не проявляли интереса к изучению дисциплины.

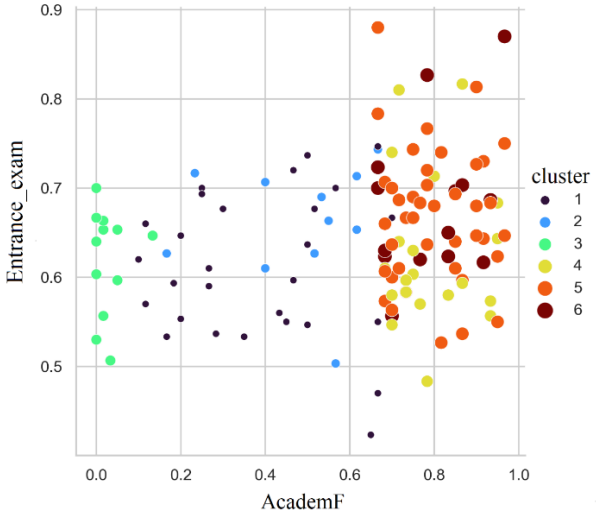

Рис. 3. Кластеры в координатах *AcademF*, *Entrance_exam*

Доля студенток, сдавших экзамен, составляет 14/17≈76,5%, что заметно выше, чем число сдавших экзамен студентов (63/112 ≈ 56,2%). Одним из факторов данного различия, возможно, является тот факт, что студентки первого курса имели на момент поступления в ВУЗ более высокий средний балл вступительных испытаний. Кроме того, как правило, девушки более ответственно подходят к выполнению заданий в рамках изучаемого курса. Однако для проверки статистической гипотезы о равенстве долей данных недостаточно.

На рис. 3 показана диаграмма, визуализирующая кластеры в координатах *Entrance_exam* и *AcademF*. Из данной диаграммы видно, что разброс значений *Entrance_exam* для студентов, сдавших экзамен, выше, чем у студентов, которые экзамен не сдали. При итоговой успеваемости выше 0,667 (т.е. от 40 баллов и выше) почти отсутствуют представители кластеров с академической задолженностью. Это обусловлено тем, что студенты, которые к дате проведения экзамена не наберут хотя бы 40 баллов из 60 возможных и не выполнят все установленные учебные задания, к экзамену не допускаются и у них образуется академическая задолженность. Студенты, которые набрали 40 баллов и выше, были допущены к экзамену, хорошо осваивают учебную программу и, как правило, успешно сдают экзамен.

Заключение. Основным результатом работы является тот факт, что была показана возможность прогнозирования сдачи экзамена по отдельной дисциплине студентов первого курса вуза на основе данных контрольных срезов, которые проводят деканаты для выявления групп студентов с повышенным риском возникновения академической задолженности. Дополнительным фактором, который использовался для построения прогнозных моделей, выступал результат вступительных испытаний в Институт (ЕГЭ или внутренний экзамен). Точность прогнозирования, которую показали рассматриваемые модели (наивный байесовский классификатор и логистическая регрессия) оказалась вполне приемлемой как на этапе проведения первого рубежного контроля, так и на этапе второго.

Результаты данной работы имеют практическое значение для администрации ВУЗов и для преподавателей. Прогнозные модели можно использовать при выявлении группы студентов, имеющих высокий риск отчисления вследствие академической неуспеваемости. Модели могут быть встроены в образовательные информационные системы и быть помощником преподавателям для принятия решений в процессе реализации дисциплины.

Кластерный анализ, проведенный с использованием иерархического подхода, показал, что в кластерах, в которых студенты не сдали экзамен по дисциплине, существенно ниже уровень посещаемости студентов. Кроме того, студенты, допустившие дефолт (т.е. образование

академической задолженности по дисциплине) менее эффективно работали на учебных занятиях в течение семестра.

Список источников

1. Горбунова Е.В. Построение модели выбытия студентов по данным университетов с разной периодичностью рубежного контроля / Е.В. Горбунова, В.В. Ульянов, К.К. Фурманов // Прикладная эконометрика, 2017. – Т. 45. – С. 116-135.
2. Гафаров Ф.М. Прогностическое моделирование в высшем образовании: определение факторов академической успеваемости / Ф.М. Гафаров, Я.Б. Руднева, У.Ю. Шарифов // Высшее образование в России, 2023. – Т. 32. – №1. – С. 51-70. – DOI: 10.31992/0869-3617-2023-32-1-51-70.
3. Czibula G., Mihai A., Crivei L. S PRAR: A novel relational association rule mining classification model applied for academic performance prediction. Procedia computer science, 2019, no. 159, pp. 20-29, DOI: 10.1016/j.procs.2019.09.156.
4. Есин Р. В. Прогнозирование успешности обучения по дисциплине на основе универсальных показателей цифрового следа LMS Moodle. / Р. В. Есин, Т. А. Кустицкая, М. В. Носков // Информатика и образование, 2023. – № 38(3). – С. 31-41. – DOI: 10.32517/0234-0453-2023-38-3-31-41.
5. Русаков С.В. Нейросетевая модель прогнозирования группы риска по успеваемости студентов первого курса / С.В. Русаков, О.Л. Русакова, К.А. Посохина // Современные информационные технологии и ИТ-образование, 2018. – Т. 14. – № 4. – С. 815-822. – DOI: 10.25559/SITITO.14.201804.815-822.
6. Шухман А.Е. Анализ и прогнозирование успеваемости обучающихся при использовании цифровой образовательной среды / А.Е. Шухман, Д.И. Парфенов, Л.В. Легашев, Л.С. Гришина // Высшее образование в России, 2021. – Т. 30. – № 8-9. – С. 125-133. – DOI: 10.31992/0869-3617-2021-30-8-9-125-13.
7. Егорова Е.С. Data Mining в образовании: прогнозирование успеваемости учащихся / Е.С. Егорова, Н.А. Попова // Моделирование, оптимизация и информационные технологии, 2023. – №11(2). – DOI: 10.26102/2310-6018/2023.41.2.003. – URL: https://moitvivt.ru/ru/journal/pdf?id=1325 (дата обращения: 20.08.2023).
8. Шармин В. Г. Определение степени влияния различных факторов на академическую успеваемость студентов на основе их самооценки, в том числе с учетом пола студента / В. Г. Шармин, Т. Н. Шармина, Д. В. Шармин // Science for Education Today, 2022. – Т. 12. – № 3. – С. 92–114. – DOI: 10.15293/2658-6762.2203.05.
9. Шмарихина Е.С. Исследование факторов успеваемости обучающихся / Е.С. Шмарихина // Вестник НГУЭУ, 2018. – № 3. – С. 130-143.
10. Куприянов Р.Б. Повышение качества модели прогнозирования образовательных результатов студентов университетов / Р.Б. Куприянов, Д. Ю. Звонарев // Информатика и образование, 2021. – Т. 36(9). – С. 40–46. – DOI: 10.32517/0234-0453-2021-36-9-40-46.
11. Куприянов Р. Б. Разработка модели прогнозирования образовательных результатов обучающихся для университетов / Р.Б. Куприянов, Д.Ю. Звонарев // Искусственный интеллект и принятие решений, 2021. – №2. – С. 11-20. – DOI: 10.14357/20718594210202.
12. Носков М.В. Прогностическая модель оценки успешности предметного обучения в условиях цифровизации образования / М.В. Носков, Ю.В. Вайнштейн, М.В. Сомова, И.М. Федотова // Вестник Российского университета дружбы народов. Серия: Информатизация образования, 2023. – Т. 20. – № 1. – С. 7–19. – DOI:10.22363/2312-8631-2023-20-1-7-19.
13. Накарякова Н.Н. Прогнозирование группы риска (по успеваемости) среди студентов первого курса с помощью дерева решений / Н.Н. Накарякова, С.В. Русаков, О.Л. Русакова // Прикладная математика и вопросы управления. – 2020. – № 4. – С. 121-136. – DOI: 10.15593/2499-9873/2020.4.08.
14. Ерохина Е.А. Влияние результатов ЕГЭ на успеваемость студентов ВУЗ / Е.А. Ерохина, Д.В. Хруслова // Информационные технологии в науке, образовании и управлении. – Москва, 2016. – С. 265-272. – URL: https://elibrary.ru/download/elibrary_26377412_40315072.pdf.
15. Surbhi A., Santosh K. Vishwakarma, Akhilesh K. Sharma Using Data Mining classifier for predicting student's performance in UG level. International journal of Computer applications, 2017, v. 172, no.8, pp. 39-44, available at: https://www.researchgate.net/publication/319172745_Using_Data_Mining_Classifier_for_Predicting_ Student's_ Performance_in_UG_Level (accessed: 08/20/2023).

Алпатов Алексей Викторович. *К. ф.-м. н., доцент, доцент кафедры Информатика и технология программирования, Волжский политехнический институт (филиал). Научные интересы: Моделирование социально-экономических процессов на основе эконометрического похода и методов машинного обучения. AuthorID: 148251, ORCID: 0000-0002-3344-5984, alpatov80@mail.ru, 404121, Волгоградская область, Волжский, ул. Энгельса, 42а.*

UDC 004.89+004.048
DOI: 10.25729/ESI.2023.32.4.006

Application of machine learning to analyze academic performance of university students

Aleksey V. Alpatov

Volzhsky polytechnic institute (branch) of Volgograd state technical university, Russian Federation, Volzhskiy, *alpatov80@mail.ru*

Abstract. The paper presents the results of analyzing and predicting the educational results of first-year university students in the implementation of a separate discipline using machine learning. The relevance of the research topic is due to the need for universities in modern conditions to successfully compete in the educational services market, which is characterized by a low number of applicants and an increase in requirements for the quality of vocational education both on the part of applicants and on the part of the state. An important component for effective decision-making in the process of quality management of the educational process is educational analytics, on the basis of which it is possible to predict the academic performance of students, to identify factors that have a significant impact on achieving high educational results. The study showed the possibility of predicting the exam in a particular discipline of first-year university students based on the data of control sections conducted by deans during the semester to identify groups of students with an increased risk of academic debt. The prediction accuracy shown by the constructed models (neural network, decision tree and logistic regression) turned out to be quite acceptable both at the stage of the first boundary control and at the stage of the second. The results of this work are of practical importance for the administration of universities and for teachers. Predictive models can be used to predict the expulsion of students due to academic failure. Models can be embedded in educational information systems and be an assistant to teachers for decision-making in the process of implementing the discipline.

Keywords: students' performance prediction, learning analytics, educational data mining, unified state exam, neural networks, decision tree, logistic regression, clustering

References

1. Gorbunova E.V., Ulyanov V.V., Furmanov K.K. Postroenie modeli vybytija studentov po dannym universitetov s raznoj periodichnost'ju rubezhnogo kontrolja [Using data from universities with different structure of academic year to model student attrition]. Prikladnaja jekonometrika [Applied Econometrics], 2017, v. 45, pp. 116-135.
2. Gafarov F.M., Rudneva Ya.B., Sharifov U.Yu. Prognosticheskoe modelirovanie v vysshem obrazovanii: opredelenie faktorov akademicheskoj uspevaemosti [Predictive modeling in higher education: determining factors of academic performance]. Vysshee obrazovanie v Rossii [Higher education in Russia], 2023, v. 32, no. 1, pp. 51-70, DOI: 10.31992/0869-3617-2023-32-1-51-70.
3. Czibula G., Mihai A., Crivei L. S PRAR: A novel relational association rule mining classification model applied for academic performance prediction. Procedia computer science, 2019, no 159, pp. 20-29, DOI: 10.1016/j.procs.2019.09.156.
4. Esin R.V., Kustitskaya T.A., Noskov M.V. Prognozirovanie uspeshnosti obuchenija po discipline na osnove universal'nyh pokazatelej cifrovogo sleda LMS Moodle [Predicting academic performance in a course by universal features of LMS Moodledigital footprint]. Informatika i obrazovanie [Informatics and Education], 2023, v. 38, no. 3, pp. 31-41, DOI: 10.32517/0234-0453-2023-38-3-31-41.
5. Rusakov S.V., Rusakova O.L., Posokhina K.A. Nejrosetevaja model' prognozirovanija gruppy riska po uspevaemosti studentov pervogo kursa [Neural network model of predicting the risk group for the accession of students of the first course]. Sovremennye informacionnye tehnologii i IT-obrazovanie [Modern Information Technologies and IT-Education], 2018, v. 14, no 4, pp. 815-822, DOI: 10.25559/SITITO.14.201804.815-822.
6. Shukhman A.E., Parfenov D.I., Legashev L.V., Grishina L.S. Analiz i prognozirovanie uspevaemosti obuchajushhihsja pri ispol'zovanii cifrovoj obrazovatel'noj sredy [Analysis and forecasting students' academic performance using a digital educational environment]. Vysshee obrzovanie v Rossii [Higher education in Russia], 2021, v. 30, no. 8-9, pp. 125-133, DOI: 10.31992/0869-3617-2021-30-8-9-125-133.
7. Egorova E.S., Popova N.A. Data Mining v obrazovanii: prognozirovanie uspevaemosti uchashhihsja [Data Mining in education: predicting student performance]. Modelirovanie, optimizacija i informacionnye tehnologii [Modeling, Optimization and Information Technology], 2023, v. 11, no. 2, DOI: 10.26102/2310-6018/2023.41.2.003, available at: https://moitvivt.ru/ru/journal/pdf?id=1325 (accessed: 08/20/2023).
8. Sharmin V.G., Sharmina T.N., Sharmin D.V. Opredelenie stepeni vlijanija razlichnyh faktorov na akademicheskuju uspevaemost' studentov na osnove ih samoocenki, v tom chisle s uchetom pola studenta [Identifying the degree of influence of various factors on students' academic performance based on their self-assessment, taking

into account students' gender]. Science for education today, 2022, v. 12, no. 3, pp. 92-114, DOI: 10.15293/2658-6762.2203.05.

9. Shmarikhina E.S. Issledovanie faktorov uspevaemosti obuchajushhihsja [Investigation the factors of students performance]. Vestnik NGUJeU [Vestnik NSUEM], 2018, v. 3, pp. 130-143.
10. Kupriyanov R.B., Zvonarev D.Yu. Povyshenie kachestva modeli prognozirovanija obrazovatel'nyh rezul'tatov studentov universitetov [Improving the quality of the university students' academic performance prediction model]. Informatika i obrazovanie [Informatics and education], 2021, v. 36, no. 9, pp. 40-46, DOI: 10.32517/0234-0453-2021-36-9-40-46.
11. Kupriyanov R.B., Zvonarev D.Yu. Razrabotka modeli prognozirovanija obrazovatel'nyh rezul'tatov obuchajushhihsja dlja universitetov [Developing of the student's educational success prediction model for universities]. Iskusstvennyj intellekt i prinjatie reshenij [Artificial intelligence and decision making], 2021, v.2, pp. 11-20, DOI: 10.14357/20718594210202.
12. Noskov M.V, Vaynshteyn Yu.V, Somova M.V, Fedotova I.M. Prognosticheskaja model' ocenki uspeshnosti predmetnogo obuchenija v uslovijah cifrovizacii obrazovanija [Prognostic model for assessing the success of subject learning in conditions of digitalization of education]. Vestnik Rossijskogo universiteta druzhby narodov. Serija: Informatizacija obrazovanija [RUDN Journal of informatization in education], 2023, vol. 20(1), pp.7-19, DOI: 10.22363/2312-8631-2023-20-1-7-19.
13. Nakaryakova N.N., Rusakov S.V., Rusakova O.L. Prognozirovanie gruppy riska (po uspevaemosti) sredi studentov pervogo kursa s pomoshh'ju dereva reshenij [Prediction of the risk group (by academic performance) among first course students by using decision tree method]. Prikladnaja matematika i voprosy upravlenija [Applied mathematics and control sciences], 2020, no. 4, pp. 121-136, DOI: 10.15593/2499-9873/2020.4.08.
14. Erokhina E.A., Khruslova D.K. Vliyaniye rezul'tatov YEGE na uspevayemost' studentov VU [Influence of results of use on progress of students higher education institution]. Informacionnye tehnologii v nauke, obrazovanii i upravlenii [Information technologies in science, education and management], 2016, p., 265-272, available at: https://elibrary.ru/download/elibrary_26377412_40315072.pdf.
15. Surbhi A., Santosh K. Vishwakarma, Akhilesh K. Sharma. Using Data Mining Classifier for Predicting Student's Performance in UG Level. International journal of Computer applications, 2017, v. 172, no.8, pp. 39-44, available at: https://www.researchgate.net/publication/319172745_Using_Data_Mining_Classifier_for_Predicting_Student's_Performance_in_UG_Level. (accessed:08/20/2023).

Alexey Viktorovich Alpatov. *Candidate of physics and mathematics sciences, associate professor of the department of computer science and programming technology, Volzhsky Polytechnic Institute (branch) of Volgograd state technical university. Research interests: Modeling of socio-economic systems based on the econometric approach and machine learning. AuthorID (RSCI): 148251, Researcher ID(Scopus): 57170357100, ORCID: 0000-0002-3344-5984, alpatov80@mail.ru, Russia, Volgograd region, Volzhskiy, Engels str., 42a.*

Статья поступила в редакцию 06.12.2023; одобрена после рецензирования 13.12.2023; принята к публикации 14.12.2023.
The article was submitted 12/06/2023; approved after reviewing 13/12/2023; accepted for publication 12/14/2023.