УДК 004.852

ПРОГНОСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ УСПЕВАЕМОСТИ СТУДЕНТОВ С ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ

Е. Н. ШНЕЙДЕРОВ, *доцент, к. т. н., доцент кафедры проектирования информационно-компьютерных систем БГУИР*
К. С. КРЕЗ, *магистрант 2-го курса БГУИР, специальность «Электронные системы и технологии»*
А. А. ШЕМЕРЕЙ, *студент 3-го курса БГУИР, специальность «Программируемые мобильные системы»*

В исследовании с использованием двух типов предикторов получена прогностическая модель успеваемости студентов. Особенность подхода заключается в том, что предикторами выступают данные о промежуточной успеваемости студентов, что позволяет получать прогноз успеваемости применительно к конкретному учебному семестру.
Формализованная задача исследования – бинарная классификация студентов на два целевых класса: «потенциально отчисленный студент» и «студент, потенциально сдавший сессию». Для решения данной задачи применяется многослойная архитектура искусственной нейронной сети с автоматической генерацией слоев и последующим сокращением количества нейронов. В качестве набора данных использованы две независимые выборки: студенты дневной и дистанционной форм получения образования, которые исследовались независимо друг от друга. Оценка качества прогностических моделей выполнялась с помощью стандартных метрик – Accuracy, Recall, Precision, Specificity и F1 Score, рассчитанных с использованием матрицы несоответствий. Результаты анализа указывают на высокую стабильность модели, подтвержденную методом кросс-валидации.
Исследование подтверждает возможность успешного применения простых моделей искусственных нейронных сетей для прогнозирования успеваемости студентов независимо от их специальности в рамках конкретного учебного заведения.
Ключевые слова: прогностическая модель, успеваемость студентов, образовательная аналитика, многослойная нейронная сеть, искусственная нейронная сеть.

ВВЕДЕНИЕ И ПОСТАНОВКА ЗАДАЧИ

Вопрос управления академической успеваемостью студентов учреждений высшего образования, без сомнения, является актуальным на различных уровнях управления образовательным процессом. Потеря контингента вследствие несвоевременного реагирования на неуспеваемость влияет как на трудоемкость факультетского администрирования, так и на планирование ресурсов учреждения высшего образования в целом. Введение в практику использования прогностического моделирования успеваемости может при надлежащем применении его результатов повысить управляемость процессом обучения.

Примерно десятилетие назад в литературе закрепился термин интеллектуального анализа образовательных данных (Educational Data Mining, EDM) – направления, целью которого является поиск скрытых закономерностей данных, полученных в рамках образовательного процесса, которые могут быть использованы при принятии стратегических решений. А в последние годы распространенное применение при обработке и анализе образовательных данных получили интеллектуальные методы [1–4].

Отдельное внимание авторов привлекла работа Гафарова Ф. М., Рудневой Я. Б. и Шарифова У. Ю. [5], в которой с использованием простой полносвязной искусственной нейронной сети с прямыми связями исследуется возможность получения прогноза академической успеваемости (отчислений) студентов Казанского (Приволжского) федерального университета. Описанный в статье метод получился нересурсоемким и достаточно эффективным для использования (точность прогнозирования составила не менее 80 %). Вместе с тем в исследовании [5] в качестве предикторов не используется информация о промежуточных показателях обучения студентов, а также полученные результаты не позволяют ответить на вопрос о потенциале отчисления студента в определенном учебном семестре при освоении образовательных программ.

В настоящей работе предлагается с использованием промежуточных показателей обучения с учетом выводов о значимости предикторов при прогнозировании успеваемости студентов и использованием простой полносвязной нейросети получить класс студентов, находящихся в группе риска (студенты, потенциально отчисленные за академическую неуспеваемость).

Для достижения указанной цели авторами были реализованы следующие этапы:

- формализация задачи прогнозирования для использования интеллектуальных методов обработки данных;
- сбор и подготовка образовательных данных, формирование экспериментальной выборки;
- выбор и реализация моделей искусственной нейронной сети;
- обучение нейронной сети, оценка качества моделей и точности прогнозирования;
- использование полученных моделей для получения прогноза для данных, не включенных в выборку.

ПРЕДИКТОРЫ АКАДЕМИЧЕСКОЙ УСПЕВАЕМОСТИ

Для построения прогностической модели успеваемости студентов предлагается использовать 2 типа предикторов: академические и неакадемические. Академические предикторы включают в себя средние баллы по итогам сессии в предыдущие семестры (соответственно, учитывается только тип аттестации, по которой студенту выставляется оценка), количество пересдач в предыдущих семестрах и результаты вступительных испытаний. Неакадемические предикторы включают в себя пол, прогнозируемый семестр, вид оплаты за обучение и форму получения образования.

В связи с тем, что авторами не ставилась цель получения одной универсальной модели, неакадемические предикторы «прогнозируемый семестр» и «форма получения образования» были учтены не в качестве входных параметров, а в качестве условия формирования различных моделей.

Детальное описание предикторов успеваемости приведено в таблице 1.

МОДЕЛЬ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ

Искусственные нейронные сети можно применять для решения базовых задач классификации, регрессии, кластеризации, генерации и другого, используя свои архитектурные особенности применительно к конкретной решаемой задаче, набору данных и используемым инструментам [6]. Целевая задача настоящего исследования сводится к задаче бинарной классификации объекта (студента) в следующие целевые классы: «Потенциально отчисленный студент» и «Студент, потенциально сдавший сессию». Эта задача относится к классу задач обучения с учителем [7, 8]. Входные параметры для обучения нейросети и, соответственно, описания объектов являются табличными подготовленными данными. На основе указанных характеристик для использования была выбрана многослойная архитектура искусственной нейронной сети с прямыми связями между слоями.

Принимая во внимание, что классификация объекта осуществляется на основании вариативного количества информации о нем, авторами использовалась реализация искусственной нейронной сети с автоматической генерацией слоев и послойным сокращением количества нейронов в них. Размерность первого слоя искусственной нейронной сети равна количеству параметров, описывающих объект.

Для программной реализации указанной модели был выбран DatRetClassifier, использующий библиотеки Tensorflow и Scikit-Learn [9]. Авторы использова-

Таблица 1. Предикторы успеваемости студентов для проведения эксперимента

Наименование предикторов	Вид предиктора	Значения предиктора
Пол	Категориальный	«Мужской» или «Женский»
Вид оплаты за обучение	Категориальный	«Бюджет» или «Оплата»
Результат ЦТ «Математика»	Числовой, целочисленный, неотрицательный	0 … 100
Результат ЦТ «Физика»		
Результат ЦТ «Английский язык»		
Результат ЦТ «Русский / белорусский язык»		
Средние баллы по итогам сессии студентов в предыдущие семестры (от 1 до 9 параметров в зависимости от прогнозируемого семестра)	Числовой, непрерывный	4,0 … 10,0
Количество пересдач студента в предыдущих семестрах (от 1 до 9 параметров в зависимости от прогнозируемого семестра)	Числовой, целочисленный, неотрицательный	0 и более

ли 500 нейронов в первом полносвязном (скрытом) слое, в каждом последующем слое количество нейронов в два раза меньше, а количество нейронов в выходном слое для задачи бинарной классификации равно двум. Все нейроны имеют функцию активации ReLU (линейный выпрямитель). Нейронная сеть обучалась с использованием оптимизатора Adam, функция потерь – бинарная перекрестная энтропия.

Детальное описание модели искусственной нейронной сети имеет архитектуру, представленную в таблице 2.

В таблице 2 значение X – количество параметров для описания объекта применительно к конкретному семестру (от 8 до 24 предикторов).

Программно устанавливаемые параметры для реализации искусственной нейронной сети приведены в таблице 3.

НАБОР ДАННЫХ

Для исследования были сформированы 2 выборки данных (студенты дневной и дистанционной форм получения образования), которые исследовались независимо друг от друга. Выборка не учитывала деления по образовательным программам (специальностям, специализациям и профилизациям). Для ее формирования учитывались следующие условия:

1. Базовое высшее образование (бакалавриат).
2. Дневная и дистанционная формы получения образования.

Таблица 2. Описание модели используемой искусственной нейронной сети

	1-й слой	2-й слой	3-й слой	4-й слой	5-й слой	6-й слой	7-й слой	8-й слой	9-й слой	10-й слой
Функция потерь	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU
Количество нейронов	X	500	250	125	62	31	15	7	3	2
Оптимизатор	Adam									
Количество эпох при обучении	До 50									
Функция потерь	Бинарная перекрестная энтропия									

Таблица 3. Параметры программной реализации DatRetClassifier

Параметр	Значение	Описание
epoch: int	до 50	Количество эпох для обучения модели
optimizer: str	Adam(learning_rate=0.001)	Имя оптимизатора или экземпляра оптимизатора
loss: str	CategoricalCrossentropy()	Функция потери
verbose	0	Вывод обучение модели по эпохам
number_neurons: int	500	Количество слоев в первом полносвязном слое
validation_split: float	0	Доля данных обучения, которые будут использоваться в качестве данных проверки
batch_size: int	1	Количество выборок на обновление градиента
shuffle	True	«Перемешивание» обучающей выборки
callback: []	[EarlyStopping(monitor='loss', mode='auto', patience=7, verbose=1), ReduceLROnPlateau(monitor='loss', factor=0.2, patience=3, min_lr=0.00001, verbose=1)]	Утилиты, вызываемые в определенные моменты во время обучения модели

3. Наличие результатов централизованного тестирования по трем предметам (для БГУИР это математика, физика и один из государственных языков).

4. Наличие полных данных в информационной системе университета о результате освоения учебных дисциплин с момента начала обучения.

5. Обучение по полному циклу образовательных программ за 2014–2018/2019 (для дневной формы / для дистанционной формы), 2015–2019/2020, 2016–2020/2021, 2017–2021/2022, 2018–2022/2023.

Для указанного в предыдущем разделе классификатора набор данных не должен содержать пропущенных и ошибочных значений [10], поэтому на первом этапе подготовки с использованием языка программирования Python была выполнена очистка данных от неполной, недостоверной или ошибочной информации. Данные собирались таким образом, чтобы информация по предикторам, которая содержится в информационной системе университета, агрегировалась в уникальной записи каждого обучающегося. Формат файла полученных данных – *.json. Общее количество уникальных записей о студентах – 10 897 записей.

Представление выборок данных по обучающимся сведено в таблицы 4, 5.

Таблица 4. Представление выборки данных для исследования студентов дневной формы получения образования

Характеристики		Семестр						
		2	3	4	5	6	7	8
Количество уникальных записей студентов в выборке данных, чел.		9848	9457	9096	8934	8777	8708	8319
Количество сдавших сессию студентов в выборке данных, чел.		9457	9096	8934	8777	8708	8319	8107
Количество отчисленных студентов в выборке данных, чел.		391	361	162	157	69	389	212
Количество в выборке данных	мужчин, чел.	7438	7093	6786	6655	6615	6454	6116
	женщин, чел.	2410	2364	2310	2279	2262	2254	2203
Количество студентов, имеющих оценку по ЦТ «Математика», чел.		9848	9457	9096	8934	8777	8708	8319
Количество студентов, имеющих оценку по ЦТ «Физика», чел.		9085	8708	8370	8213	8066	8000	7611
Количество студентов, имеющих оценку по ЦТ «Англ. язык», чел.		763	749	726	721	711	708	708
Количество студентов, имеющих оценку по ЦТ «Рус. / бел. язык», чел.		9848	9457	9096	9096	8777	8708	8319
Количество студентов, имеющих пересдачи, чел.		2363	3257	3887	4334	4729	5114	5159

Таблица 5. Представление выборки данных для исследования студентов дистанционной формы получения образования

Характеристики		Семестр								
		2	3	4	5	6	7	8	9	10
Количество уникальных записей студентов в выборке данных, чел.		1049	957	867	684	562	511	420	378	324
Количество сдавших сессию студентов в выборке данных, чел.		957	867	684	562	511	420	378	324	297
Количество отчисленных студентов в выборке данных, чел.		92	90	183	122	51	91	42	54	27
Количество в выборке данных	мужчин, чел.	767	695	684	490	406	363	294	260	223
	женщин, чел.	282	262	183	194	156	148	126	118	101

Характеристики	Семестр								
	2	3	4	5	6	7	8	9	10
Количество студентов, имеющих оценку по ЦТ «Математика», чел.	1049	957	867	684	562	511	420	378	324
Количество студентов, имеющих оценку по ЦТ «Физика», чел.	958	876	790	610	497	449	365	327	278
Количество студентов, имеющих оценку по ЦТ «Англ. язык», чел.	91	81	77	74	65	62	55	51	46
Количество студентов, имеющих оценку по ЦТ «Рус. / бел. язык», чел.	1049	957	867	684	562	511	420	378	324
Количество студентов, имеющих пересдачи, чел.	430	598	667	532	478	461	381	347	297

Предварительный анализ очищенного набора данных показал, что в БГУИР превалируют две причины отчисления: «академическая неуспеваемость» и «по собственному желанию». В настоящем исследовании причины отчисления обучающихся не входили в качестве параметров в реализованную модель, однако это создает дополнительные условия для возникновения гипотез о наличии закономерностей в образовательных данных.

ОЦЕНКА КАЧЕСТВА МОДЕЛИ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ

В задачах оценки качества и сравнения моделей классификации, как правило [11], используется матрица несоответствий (confusion matrix), содержащая в себе показатели корректно и некорректно классифицированных объектов. Для бинарной классификации матрица несоответствий приведена в таблице 6.

Таблица 6. Общий вид матрицы несоответствий для бинарной классификации

Категория		Фактический класс	
		Студент, потенциально сдавший сессию	Потенциально отчисленный студент
Прогнозный класс	Студент, потенциально сдавший сессию	Истинно-положительный (True Positive – TP)	Ложноположительный (False Positive – FP)
	Потенциально отчисленный студент	Ложноотрицательный (False Negative – FN)	Истинно-отрицательный (True Negative – TN)

С помощью матрицы несоответствий были рассчитаны основные метрики качества нейронной сети: общая точность прогноза (Accuracy), полнота (Recall), точность (Precision), специфичность (Specificity) и F1-мера (F1 Score, среднее гармонической общей точности и полноты). Метрики качества для выборки студентов дневной формы получения образования приведены в таблице 7, а для студентов дистанционной формы обучения – в таблице 8.

Таблица 7. Метрики качества для различных реализаций модели (семестров) выборки студентов дневной формы получения образования

Метрика качества	Модель (прогнозный семестр)						
	2	3	4	5	6	7	8
Accuracy	0,871	0,852	0,882	0,867	0,853	0,829	0,816
Recall	0,842	0,851	0,820	0,861	0,832	0,847	0,805
Precision	0,854	0,846	0,885	0,890	0,867	0,818	0,832
Specificity	0,934	0,972	0,937	0,956	0,940	0,931	0,922
F1 Score	0,847	0,848	0,851	0,875	0,849	0,832	0,818

Таблица 8. Метрики качества для различных реализаций модели (семестров) выборки студентов дистанционной формы получения образования

Метрика качества	Модель (прогнозный семестр)								
	2	3	4	5	6	7	8	9	10
Accuracy	0,885	0,840	0,810	0,860	0,857	0,872	0,892	0,789	0,800
Recall	0,840	0,827	0,801	0,833	0,831	0,798	0,830	0,803	0,809
Precision	0,863	0,832	0,856	0,824	0,821	0,876	0,900	0,819	0,847
Specificity	0,970	0,968	0,963	0,954	0,920	0,980	0,930	0,942	0,925
F1 Score	0.851	0.829	0,827	0.828	0,825	0,835	0,863	0,810	0,827

Для проверки устойчивости результатов на выборке студентов дистанционной формы получения образования была проведена оценка общей точности модели с использованием метода кросс-валидации. Вся выборка была разбита на 10 подвыборок с равным количеством записей в каждой. Далее было последовательно проведено 10 экспериментов, в которых каждая выделенная подвыборка была использована в качестве контрольной, а остальные 9 – в качестве обучающих. Среднее значение общей точности прогноза по экспериментам для различных выборок попадает в диапазон 0,845 … 0,855, что составляет отклонение от частных значений, полученных базовым методом разбивки набора данных на обучающую и тестовую выборки, не более 10 %.

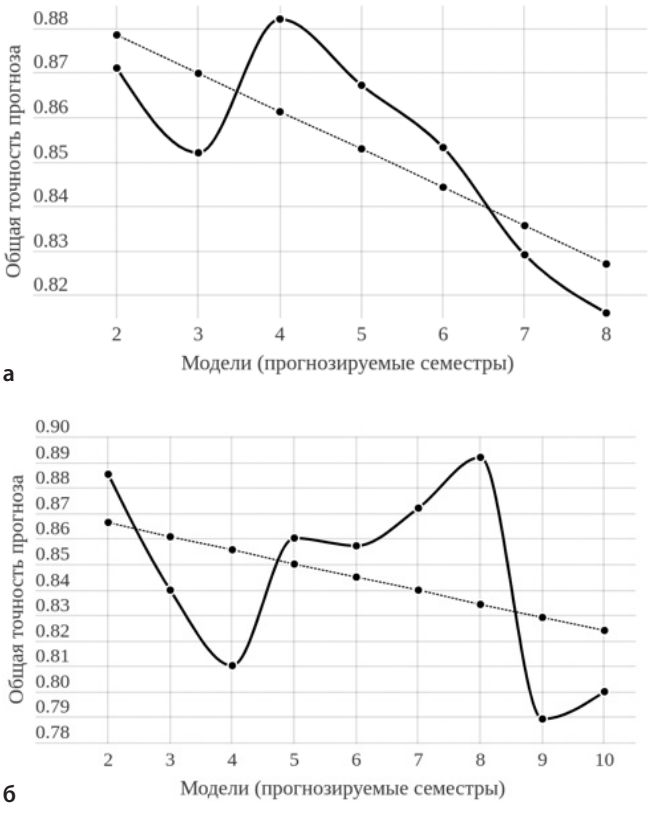
РЕЗУЛЬТАТЫ И ВЫВОДЫ

Для прогнозирования отчисления студентов предложены и обучены достаточно простые модели искусственных нейронных сетей, архитектурно представляющие собой «нейросетевые воронки» с различным размером входного слоя, решающие задачу бинарной классификации студентов в целевой класс «потенциально отчисленный студент». На вход нейросетей подавались векторы, состоящие от 8 до 24 предикторов, в зависимости от прогнозируемого семестра, которые описывали конкретного студента как до начала его обучения, так и в процессе самого обучения. Качество полученных моделей оценено на основе метрик Accuracy, Recall, Precision, Specificity, F1 Score.

В качестве практического использования полученных моделей авторами был получен прогноз для не входящих в экспериментальную выборку студентов набора 2019–2022 гг. для отдельных семестров образовательных программ. Общая точность прогноза для студентов дневной формы получения образования составила 0,84, а для студентов дистанционной формы – 0,839.

Достаточно любопытна тенденция уменьшения общей точности получаемого прогноза с увеличением номера семестра даже с учетом увеличения объема информации об объекте, изображенная на рисунке. Авторы полагают, что это обусловлено влиянием на отчисление студента факторов, не связанных с образовательным процессом, – социальные, трудовые, и иные причины.

Дискуссионным является вопрос универсальности и тиражируемости предлагаемых моделей для получения прогноза успеваемости студентов. Для ответа на него требуются дополнительные исследования зависимости успеваемости студентов от особенностей организационных процессов и педагогических практик учреждений образования, направлений специальностей, методик оценивания и множества других факторов [12].



Зависимость общей точности прогноза от модели (номера семестра):
а – для студентов дневной формы получения образования;
б – для студентов дистанционной формы получения образования

Таким образом, авторами экспериментально показана возможность практического применения простой искусственной нейронной сети для прогнозирования успеваемости отдельных студентов независимо от специальности в контексте отдельного учреждения высшего образования.

ЛИТЕРАТУРА

1. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., Liao, S.N. Predicting academic performance: a systematic literature review // In: Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (IT-iCSE 2018 Companion). Association for Computing Machinery. New York, USA. – 2018. – P. 175–199.
2. Raju, D., Schumacker, R. Exploring student characteristics of retention that lead to graduation in higher education using data mining models / D. Raju, R. Schumacker // Journal of College Student Retention: Research, Theory and Practice. – 2015. – Vol. 16, №. 4. – P. 563–591.
3. Lesinski, G., Corns, S., Dagli, C. Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy / G. Lesinski, S. Corns, C. Dagli // Procedia Computer Science. – 2016. – Vol. 95. – P. 375–382.
4. Lau, E. T., Sun, L., Yang, Q. Modelling, prediction and classification of student academic performance using artificial neural networks // SN Applied Sciences. – 2019. – Vol. 1, art. № 982.
5. Gafarov, F. M., Rudneva, Ya. B., Sharifov, U. Yu (2023). Predictive Modeling in Higher Education: Determining Factors of Academic Performance. Vysshee obrazovanie v Rossii = Higher Education in Russia. – Vol. 32, №. 1. – P. 51–70.
6. Горбачевская, Е. Н. Классификация нейронных сетей / Е. Горбачевская // Вестник ВУиТ. – 2012. – № 2 [Электронный ресурс]. – Режим доступа: http://cyberleninka.ru/article/n/klassifikatsiya-neyronnyh-setey. – Дата доступа: 15.01.2024.
7. Забашта, А. С., Фильченков, А. А. Построения наборов данных для задачи бинарной классификации по их характеристическому описанию / А. С. Забашта, А. А. Фильченков // Научно-технический вестник информационных технологий, механики и оптики. – 2017. – Т. 17. – № 3. – С. 498–505.
8. Воронцов, К. В. Математические методы обучения по прецедентам (теория обучения машин) [Электронный ресурс]. – Режим доступа: http://www.machinelearning. ru/wiki/images/6/6d/voron-ml-1.pdf. – Дата доступа: 10.01.2024.
9. DatRet: Tensorflow implementation for structured tabular data [Электронный ресурс]. – Режим доступа: https://github.com/AbdualimovTP/datret. – Дата доступа: 19.01.2024.
10. Alyahyan, E., Düştegör, D. Predicting academic success in higher education: literature review and best practices / E. Alyahyan, D. Düştegör // International Journal of Educational Technology in Higher Education. – 2020. – Vol. 17, art. № 3. – P. 1–21.
11. Михайличенко, А. А. Аналитический обзор методов оценки качества алгоритмов классификации в задачах машинного обучения / А. А. Михайличенко // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки, 2022. – № 4. – С. 52–59.
12. Шмелева, Е. Д., Фрумин, И. Д. Факторы отсева студентов инженерно-технического профиля в российских вузах / Е. Д. Шмелева, И. Д. Фрумин // Вопросы образования. – 2020. – № 3. – С. 110–136.

In the article, a predictive model of students' academic performance was obtained using two types of predictors. The peculiarity of the approach is that the predictors are data on students' intermediate academic performance, allowing to obtain a prediction of academic performance in relation to the academic semester.

The main task of the study is binary classification of students into two target classes: «potentially expelled student» and «student who potentially passed the session». To solve this problem, a multilayer artificial neural network architecture with automatic generation of layers and subsequent reduction of the number of neurons is applied. Two independent samples were used as a data set: full-time and distance education students, which were studied independently of each other. The quality of the predictive models was assessed using standard metrics – Accuracy, Recall, Precision, Specificity and F1 Score calculated using the inconsistency matrix. The results of the analysis indicate high stability of the model as confirmed by cross-validation.

The research confirms the possibility of successful application of simple artificial neural network models for predicting students' academic performance within a particular institution, regardless of their specialty.

Keywords: predictive model, student performance, educational analytics, multilayer neural network, artificial neural network.

Получено 25.01.2024.