

УДК 316.344.3

DOI: 10.17223/1998863X/64/13

В.В. Кашпур, Е.Ю. Петров, В.Л. Гойко, А.В. Фещенко

ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ ЦИФРОВЫХ СЛЕДОВ ДЛЯ ПРОГНОЗИРОВАНИЯ ОБРАЗОВАТЕЛЬНЫХ ДОСТИЖЕНИЙ СТУДЕНТОВ

Исследование выполнено при финансовой поддержке РФФИ в рамках
научного проекта № 19-31-51001.

В статье представлены результаты построения модели прогнозирования образовательных достижений студентов на основе данных из электронной образовательной системы ТГУ и их цифрового следа в ВКонтакте с использованием алгоритмов машинного обучения. Сделан вывод о специфике цифрового следа студентов с высокими образовательными достижениями, проявляющийся в подписках и количестве групп членства, друзей и подписчиков.

Ключевые слова: цифровой след, студенты, образовательные достижения, машинное обучение

Введение

Реализация образовательной политики в направлении выявления и поддержки талантов, профориентации и индивидуализации обучения формирует запрос как на анализ факторов, влияющих на образовательные достижения студентов, так и на разработку прогностических моделей их образовательной успешности, основанной на выявленных факторах. Традиционными источниками данных для решения этой задачи выступают данные, генерируемые в рамках систем электронного обучения образовательных учреждений, данные психологических тестирований и социологических опросов студентов, а также образовательная и социально-экономическая статистика. Чаще всего модели анализа и прогнозирования образовательных достижений используют данные о текущей академической успеваемости студентов [1]. Однако до сих пор мало используются возможности анализа данных, генерируемых студентами в интернете и социальных медиа, – цифровые следы. Под цифровым следом (англ. digital footprint) в данной статье понимается совокупность информации о пользователе и структуре и содержании его активности в он-лайне.

Изучение цифрового следа студента позволяет повысить качество учебной аналитики и прогнозтики за счет следующих преимуществ. Во-первых, цифровой след содержит большое количество открытых пользовательских данных о персональных (когнитивных, мотивационных, психологических) характеристиках студентов. Важно отметить, что эти данные генерируются естественным образом самим респондентом, а не в искусственной ситуации тестирования или опроса. Во-вторых, анализ цифровых следов студентов позволяет оценивать образовательный потенциал «на входе» в университет. В отличие от существующих моделей прогнозирования образовательных до-

стижений, основанных на данных о текущей успеваемости, анализ цифровых следов дает возможность дать такой прогноз относительно студентов, только что поступивших в университет и еще не прошедших первые сессии. В-третьих, цифровые следы дают новую возможность для оценки и анализа образовательных достижений обучающихся – они позволяют оценить неформальные и информальные образовательные достижения студентов, которые не фиксируются в рамках образовательной программы по направлению обучения в университете.

Поэтому субъекты управления университетами уже сейчас и в ближайшем будущем будут генерировать запрос на разработку аналитических инструментов и моделей работы с цифровыми следами, которые позволят повысить качество управленческих решений в сфере управления образовательным процессом в целом и образовательными траекториями студентов в частности.

Данная статья резюмирует один из первых опытов использования анализа цифровых следов студентов для задач учебной аналитики, а именно прогнозирования формальных образовательных достижений обучающихся. В качестве исследовательских в данной статье рассматриваются следующие вопросы:

- какие конкретные компоненты цифровых следов можно использовать для прогнозирования образовательных достижений студентов?
- как можно дифференцировать студентов по уровню их формальных образовательных достижений?
- какие алгоритмы можно использовать для прогнозирования образовательных достижений студентов?

Рассмотрению ответов на эти вопросы и посвящена данная статья.

Обзор литературы по тематике взаимосвязи интернет-активности и образовательных достижений обучающихся

Публикации по тематике взаимосвязи интернет-активности в социальных сетях и образовательных достижений обучающихся начали появляться относительно недавно – в начале 2010-х гг. Это связано с новизной самого изучаемого феномена цифровых следов в социальных сетях. Первая социальная сеть Classmates запущена в 1995 г., самая распространенная социальная сеть в мире Facebook была основана в 2004 г. (причем как социальная сеть Гарвардского университета, затем ставшая открытой для любых пользователей). Выход на рынок смартфона Apple в 2007 г. резко ускорил рост численности пользователей социальных сетей благодаря удобным мобильным версиям социальных сетей и мобильному интернету. Поэтому в период с 2008 по 2012 г. наблюдался резкий рост количества пользователей социальных сетей и объема генерируемых ими цифровых следов [2].

Первые публикации по теме фиксировали негативную взаимосвязь между использованием социальных сетей и образовательными достижениями обучающихся. В частности, в ряде исследований было отмечено негативное влияние частоты использования Facebook на вовлеченность студентов в образовательный процесс [3, 4]. Аналогичные выводы были сделаны и в отношении школьников. Так, более низкий результат образовательного тести-

рования и меньшее количество проведенных за учебой часов в неделю для школьников, пользующихся Facebook, были выявлены в исследовании П. Киршнера и А. Карпинского [5]. Однако результаты этих исследований были получены с использованием традиционных методов социальных наук – массового опроса и кейс-стади. Их общим недостатком можно считать выстраивание выводов о влиянии интернет-активности в социальных сетях на образовательные достижения не на основе фактической информации, а на основе репрезентации представлений студентов и школьников.

Начиная с 2015 г. стали появляться публикации, основанные на анализе цифровых следов обучающихся при помощи методов Data Mining и интеллектуального анализа данных с использованием алгоритмов машинного обучения. Практически все подобные публикации демонстрируют позитивную взаимосвязь между цифровым следом и интернет-активностью и образовательными достижениями студентов и школьников. В качестве примера можно привести следующие публикации: статьи И. Смирнова, фиксирующие положительную взаимосвязь между фактом присутствия студента в «ВКонтакте» и его средним баллом [6], а также положительную зависимость между объемом вокабуляра на странице социальной сети, частотой использования английских слов, слов, связанных с чтением, мышлением, запоминанием, длиной слов и постов и академической успеваемостью [7]; статья А. Красильникова и М. Семеновой, демонстрирующая положительную взаимосвязь между временем, проведенным в «ВКонтакте» перед экзаменами, и величиной оценки за экзамены [8]. Ряд исследований фиксируют взаимосвязь подписок, отражающих интересы школьников и студентов и их академическую успеваемость [9, 10]. Взаимосвязь интернет-активности и когнитивных (интеллект, креативность) и мотивационных особенностей личности обучающихся была зафиксирована в исследованиях коллектива под руководством А. Фещенко [11]. Эти исследования нашли значимое практическое применение – с помощью анализа цифрового следа в социальной сети «ВКонтакте» осуществляется рекрутинг абитуриентов в Томском государственном университете. С помощью применения методов искусственного интеллекта распознаются предметные интересы человека на основе его подписок в социальной сети, а также обучается модель машинного обучения для поиска «своего» абитуриента Томского государственного университета на основе результатов психологического тестирования, отражающего когнитивные способности студентов.

Отдельным значимым компонентом цифрового следа, имеющим взаимосвязь с образовательными достижениями, являются сетевые связи обучающихся. Исследования на материале иностранных и российских студентов зафиксировали структуриацию сетевых связей в зависимости от образовательных достижений, показав наличие эффекта гомофилии среди студентов с высокими и низкими образовательными достижениями [12, 13].

Методы и источники данных

Достижение цели прогнозирования образовательных достижений на основе данных цифровых следов обучающихся базировалось на использовании современных методов сбора, обработки и анализа цифровых данных – Data Mining и машинное обучение.

Методы Data Mining использовались при получении данных цифровых следов студентов. В качестве входных данных использован цифровой след пользователя из социальной сети «ВКонтакте», а также данные об успеваемости студента в системе электронного обучения LMS Moodle.

Сбор данных осуществлялся в несколько этапов. На первом этапе выгружались данные об успеваемости студентов ТГУ из LMS Moodle. Выгружены данные об успеваемости 9 360 студентов. Во избежание возникновения утечки персональных данных они были предварительно обезличены. Помимо среднего балла, были получены данные о факультете, уровне образования, статусе студента, направлении подготовки, а также идентификатор в системе Moodle.

Второй этап сбора данных – идентификация студентов в социальной сети. Данный этап проводился в полуавтоматическом режиме с использованием алгоритмов поиска, использующих открытое API ВКонтакте. С помощью метода VK API-search выполнялся поиск пользователей в социальной сети. При совпадении ФИО и возраста пользователи добавлялись в результирующую выборку, после чего проводилась ручная валидация пользователей. Проверялись их подписки на наличие университетских или региональных сообществ, а также поля профиля, где указано место учебы. Отсеивались неактивные и заблокированные аккаунты. В итоге было идентифицировано 6 488 пользователей.

Заключительный этап сбора данных – выгрузка цифрового следа идентифицированных пользователей. Структура цифрового следа пользователя «ВКонтакте» состоит из следующих элементов:

1) Общая информация – пол, город, количество друзей, подписчиков, постов на стене, фотографий, видеозаписей, аудиозаписей, информация об образовании, карьере. Эти признаки могут давать какую-то общую информацию о человеке и его активности в социальной сети.

2) Личная информация – это информация, указываемая пользователем самостоятельно. Это жизненная позиция, интересы, любимые фильмы, музыка, цитаты, произвольная информация о себе. В теории это должно быть очень хорошим признаком для составления некоторого цифрового портрета человека, однако на практике эти поля заполнены подавляющим меньшинством пользователей.

3) Сообщества, на которые подписан пользователь (далее – «подписки пользователя»). Сообщества «ВКонтакте» упорядочены в соответствии с частотой и характером взаимодействия пользователя с контентом. Очевидно, что пользователь подписан на те сообщества, которые ему интересны в силу тех или иных причин. Таким образом, если получить распределение тематик в сообществах пользователя, то можно узнать его интересы.

Все собранные данные объединялись в единую базу, необходимую для построения модели прогнозирования образовательных достижений студентов.

Методы машинного обучения, которые использовались для построения модели прогнозирования образовательных достижений на основе данных цифрового следа, базировались на применении следующих алгоритмов: градиентного бустинга над решающими деревьями из библиотек CatBoost и LightGBM, а также стохастического градиентного спуска. Их применение

было обусловлено рамочными требованиями модели: масштабируемость – так как количество переменных цифрового следа пользователей довольно велико и необходимо, чтобы алгоритм мог без труда справляться с ними; быстрая обучаемость и минимальные требования к вычислительным ресурсам – алгоритм должен быстро реагировать для оперативного и своевременного получения прогноза.

Дифференциация обучающихся в зависимости от их образовательных достижений

Одним из важных результатов построения модели прогнозирования образовательных достижений на основе цифровых следов явилась разработка такого подхода к дифференциации обучающихся на студентов с высокими и низкими образовательными достижениями, который позволил бы использовать его в рамках применения инструментов машинного обучения.

Существует несколько подходов, которые используются в подобных задачах. В ходе исследования был оценен каждый подход путем подачи фиксированной выборки студентов в модель машинного обучения и по результатам F-меры делать выводы относительно точности и полноты модели.

1. Разбиение по 75-му перцентилю. Это был первый подход к дифференциации студентов. В рамках данного подхода считалось, что студенты, средний балл которых больше этой границы, обладают высокими образовательными достижениями (1), если средний балл меньше или равен границе – низкими (0). При таком подходе возникают сильный дисбаланс классов и слабая разделимость между ними, что видно по результатам тестирования.

2. Выбор крайних границ с использованием полярных квартилей (1-й и 4-й квартили) [14]. Считаем, что высокими образовательными достижениями обладают студенты, средний балл которых выше 75-го перцентиля, низкими – ниже 25-го перцентиля. Такой подход обеспечивает сбалансированность классов и хорошую разделимость между ними, однако теряется половина данных.

3. Система ECTS [15]. Европейская система оценивания учета работы студентов в рамках образовательной программы, которая имеет следующую градацию: А – лучшие 10%, В – следующие 25%, С – следующие 30%, D – следующие 25%, Е – следующие 10%. В рамках системы ECTS считается, что высокими образовательными достижениями обладают студенты, имеющие оценки «А» и «В», остальные относятся к низкому уровню. При таком методе дифференциации имеем более сбалансированные классы, чем в 1-м методе, к тому же система ECTS предусматривает метод конвертирования оценок между университетами в разных странах, что расширяет границы применения предложенной методики. Однако при апробации данного подхода получается слабая разделимость между классами.

4. На основании анализа преимуществ и недостатков описанных методов в рамках данного исследования был предложен другой подход, заключающийся в разбиении по крайним границам системы ECTS. Было предложено считать, что высокими образовательными достижениями обладают студенты, чей средний балл выше границы 65-го перцентиля, а низкими – ниже границы 35-го перцентиля. Тем самым получена сбалансированность классов, хо-

рошая разделимость между ними и меньшая потеря данных, чем в методе полярных квартилей.

Таким образом, резюмируя полученные результаты (табл. 1), получаем, что лучше всего по каждому из направлений обучения (гуманитарное, естественное, техническое) себя показал метод, при котором берутся крайние границ по ECTS.

Таблица 1. Результаты тестирования методов дифференцирования обучающихся по академической успеваемости (значение метрики F-score)

Направление подготовки	Метод 1. Разбиение по 75-му перцентилю	Метод 2. Разбиение по полярным квартилям. 1-й квартиль – низкий, 4-й – высокий	Метод 3. Система ECTS	Метод 4. Крайние границы по ECTS (высокий $\geq 65\%$, низкий $\leq 35\%$)
Гуманитарное	0,43	0,62	0,60	0,65
Техническое	0,46	0,66	0,62	0,69
Естественное	0,43	0,63	0,58	0,72

В дальнейшем этот метод использован для построения модели прогнозирования образовательных достижений студентов на основе цифровых следов.

Структура модели прогнозирования образовательных достижений на основе цифровых следов

На первом этапе построения модели прогнозирования образовательных достижений данные цифровых следов разделялись на три части: обучающая выборка (70% студентов), валидационная (10% студентов) и тестовая (20% студентов). Для построения модели были отобраны наиболее значимые для определения образовательных достижений компоненты цифровых следов обучающихся. Наиболее значимыми переменными оказались: пол, количество друзей, направление обучения, интересы пользователей – подписки.

Подбор подходящих параметров модели осуществлялся с помощью метода GridSearchCV из библиотеки Scikit-Learn. Обучение модели происходило на самом большом массиве данных – обучающей выборке, после чего выполнялось тестирование на данных, которые алгоритм еще не видел, – тестовой выборке. По результатам тестирования лучше всего себя показал алгоритм градиентного бустинга над решающими деревьями реализации Catboost (табл. 2). Он был использован в качестве основного алгоритма модели прогнозирования образовательных достижений студентов на основе цифровых следов.

Таблица 2. Результат тестирования алгоритмов (значение метрики F-score)

Направление	SGD	LGBM	Catboost
Гуманитарное	0,41	0,62	0,65
Техническое	0,43	0,66	0,69
Естественное	0,41	0,63	0,72

Применение модели прогнозирования с учетом подписок пользователей осложнялось чрезмерно большим количеством признаков – при первой итерации в базе данных было 55 469 подписок, поэтому была выполнена процедура отбора наиболее значимых признаков, известная как «feature selection». Для каждого признака была выделена его значимость. На первой итерации использовались все параметры, на них обучалась модель и делалось предска-

зание на тестовых данных. Далее последний по значению важности параметр отбрасывался. Эти операции повторялись, пока не был достигнут наилучший результат. Таким способом удалось повысить значение F-меры модели на 2% для студентов всех направлений обучения.

Еще одним инструментом повышения точности и полноты прогнозной модели стала тематическая классификация подписок пользователей. Все подписки пользователей были классифицированы по следующим укрупненным категориям:

1. Духовная жизнь, эзотерика.
2. Образование, наука.
3. Развлечения, юмор.
4. Искусство.
5. Здоровый образ жизни, спорт.
6. Общественно-политические группы.
7. Бизнес, работа.
8. Хозяйство, техника.
9. Мусорные подписки.

После добавления тематической классификации подписок в модель машинного обучения для прогнозирования образовательных достижений ее качество повысилось на 4%. Таким образом, F-мера финальной модели стала равной 78% для студентов естественных, 73% для технических и 69% для гуманитарных направлений обучения. Такие значения точности и полноты прогнозной модели считаются достаточными для решения задачи прогнозирования образовательных достижений студентов.

Таким образом, была разработана конечная структура модели представленная в виде схемы на рис. 1.

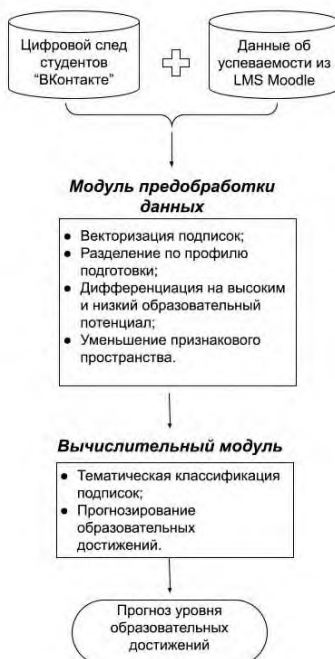


Рис. 1. Модель прогнозирования образовательных достижений студентов

После построения окончательной структуры модели с использованием метрики прироста информации (Information gain) была оценена степень влияния конкретных переменных на прогнозное значение (в нашем случае прогноз наличия высоких образовательных достижений у студента). Для каждой функции метрика прироста информации показывает, насколько в среднем прогноз изменяется при изменении значения функции. Чем больше значение важности, тем в среднем больше будет изменение значения прогноза.

Таблица 3. Наиболее значимые переменные модели прогнозирования образовательных достижений студентов на основе цифрового следа

Значение метрики прироста информации («Information gain»)	Название переменной
24,46969292	Пол
3,272388456	Доля в подписках сообществ из категории «образование, наука»
2,922822125	Количество друзей
1,761577834	Количество подписчиков
0,607760814	Количество групп
0,130184106	Доля в подписках сообществ из категории «бизнес, работа»

Заключение

Разработка модели прогнозирования образовательных достижений студентов на основе цифрового следа имеет как теоретическую, так и практическую значимость. Теоретическая значимость определяется тем, что апробация модели прогнозирования образовательных достижений студентов позволяет сделать вывод о специфике цифрового следа студентов с высокими формальными образовательными достижениями и его отличии от цифровых следов других обучающихся. Эти отличия связаны с личными интересами, отражающимися в их цифровом следе в виде подписок. Но также на них влияют и структурные характеристики интернет-активности студента, такие как количество групп, в которых он состоит, количество друзей и подписчиков.

Фиксация полученного результата позволяет сделать заявку на дополнение ключевой теории образовательных исследований, объясняющую студенческую успеваемость, – геометрической модели студенческой устойчивости и достижений У. Свейла [16. Р. 12–15]. Данная теория относительно ситуации двадцатилетней давности указывает на три типа факторов, влияющих на академическую успешность и вероятность отчисления студента: институциональные (связанные с условиями обучения в университете), когнитивные (связанные с мотивационными и личностными компетенциями и способностями) и социальные (связанные со статусом, социальным окружением и культурными нормами). Анализ влияния цифровых следов на образовательные достижения показал, что в настоящее время на стыке когнитивных и культурных факторов формируется новый тип факторов, влияющих на образовательные достижения – коннективные, связанные с особенностями онлайн-среды существования студента.

Практическая значимость разработки модели прогнозирования потенциальных образовательных достижений студентов на основе анализа данных цифрового следа заключается в возможности ее использования университе-

тами как инструмента поиска талантливых студентов, а также для тьюторской работы по профилактике студенческих задолженностей и отчислений.

Литература

1. Горбунова Е.В. Выбытия студентов из вузов: исследования в России и США // Вопросы образования. 2018. № 1. С. 110–131. DOI: 10.17323/1814-9545-2018-1-110-131
2. Kitchin R. Big Data, New Epistemologies and Paradigm Shifts // Big Data & Society. 2014. № 1 (1). P. 1–12. DOI: 10.1177/2053951714528481
3. Junco R. The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement // Computers and Education. 2012. № 58 (1). P. 162–171.
4. Paul J., Baker H., Cochran J. Effect of Online Social Networking on Student Academic Performance // Computers in Human Behavior. 2012. № 28. P. 2117–2127. DOI: 10.1016/j.chb.2012.06.016
5. Kirschner P., Karpinski A. Facebook® and academic performance // Computers in Human Behavior. 2010. № 26 (6). P. 1237–1245. DOI: 10.1016/j.chb.2010.03.024
6. Смирнов И.Б., Сивак Е.В., Козьмина Я.Я. В поисках утерянных профилей: достоверность данных ВКонтакте и их значение в образовательных исследованиях // Вопросы образования. 2016. № 4. С. 106–122. DOI: 10.17323/1814-9545-2016-4-106-122
7. Smirnov I. Estimating educational outcomes from students' short texts on social media // EPJ Data Sci. 2020. № 9 (27). DOI: 10.1140/epjds/s13688-020-00245-8
8. Krasilnikov A., Semenova M. Do social networks help to improve student academic performance? The case of Vk.com and Russian students // Economics Bulletin. 2014. № 34. P. 718–733.
9. Поливанова К.Н., Смирнов И.Б. Что в профиле тебе моем данные «ВКонтакте» как инструмент изучения интересов современных подростков // Вопросы образования. 2017. № 2. С. 134–152. DOI: 10.17323/1814-9545-2017-2-134-152
10. Ихсанов И., Шахова И. Применение методов машинного обучения для выявления взаимосвязи академической успеваемости и данных профиля социальной сети // Russian Digital Libraries Journal. 2019. № 2. С. 95–118. DOI: 10.26907/1562-5419-2019-22-2-95-118
11. Гойко В.Л., Киселев П.Б., Мацута В.В., Суханова Е.А., Степаненко А.А., Фещенко А.В. Методы и инструменты выявления перспективных абитуриентов в социальных сетях // Открытое и дистанционное образование. 2017. № 4 (68). С. 45–52.
12. Flashman J. Academic Achievement and Its Impact on Friend Dynamics // Sociology of Education. 2012. № 85. P. 61–80. DOI: 10.1177/0038040711417014
13. Smirnov I., Thurner S. Formation of homophily in academic performance: Students change their friends rather than performance // PLoS ONE. 2017. № 12 (8):e0183473. URL: <https://doi.org/10.1371/journal.pone.0183473> (accessed: 15.09.2021).
14. Lu O.H.T., Huang A.Y.Q., Lin A. J.Q., Ogata H., Yang S.J.H. Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning // Educational Technology & Society. 2018. № 21 (2). P. 220–232.
15. Sousa A., Oliveira C., Borges J. Using Academic Performance to Predict College Students Dropout: a case study // Educação e Pesquisa. 2018. № 44. URL: <https://doi.org/10.1590/s1678-4634201844180590> (accessed: 18.09.2021).
16. Swail W.S. The Art of Student Retention. A handbook for practitioners and administrators // Educational Policy Institute. 2014. URL: https://secureservercdn.net/50.62.198.97/68g.645.myftpupload.com/wp-content/uploads/2020/01/Artofstudentretention_2008.pdf (accessed: 20.09.2021).

Vitaliy V. Kashpur, Tomsk State University (Tomsk, Russian Federation); Sirius University of Science and Technology (Sochi, Russian Federation).

E-mail: vitkashpur@mail.ru

Evgeniy Y. Petrov, Tomsk State University (Tomsk, Russian Federation); Sirius University of Science and Technology (Sochi, Russian Federation).

E-mail: petrov@data.tsu.ru

Viacheslav L. Goiko, Tomsk State University (Tomsk, Russian Federation); Sirius University of Science and Technology (Sochi, Russian Federation).

E-mail: goiko@data.tsu.ru

Artem V. Feshchenko, Tomsk State University (Tomsk, Russian Federation); Sirius University of Science and Technology (Sochi, Russian Federation).

E-mail: goiko@data.tsu.ru

Vestnik Tomskogo gosudarstvennogo universiteta. Filosofiya. Sotsiologiya. Politologiya – Tomsk State University Journal of Philosophy, Sociology and Political Science. 2021. 64. pp. 140–150.

DOI: 10.17223/1998863X/64/13

POSSIBILITIES OF USING DIGITAL FOOTPRINTS TO PREDICT EDUCATIONAL ACHIEVEMENTS OF STUDENTS

Keywords: digital footprint; students; educational achievement; machine learning

The study is supported by the Russian Foundation for Basic Research, Project No. 19-31-51001.

The article summarizes one of the first experiences of using the analysis of students' digital footprints for educational analytics tasks, namely, predicting the formal educational achievements of students. Prediction of educational achievement based on digital footprint data was based on the use of Data Mining and Machine Learning methods. Data Mining methods were used to obtain student digital footprint data. As input data, the user's digital footprint from the social network Vkontakte, as well as data on student academic performance in the LMS Moodle of Tomsk State University, was used. Machine Learning methods were based on the application of the gradient boosting algorithm on decision trees from the CatBoost library. Data were obtained on 6,488 students of Tomsk State University. The sample set was divided into three parts: a training sample (70% of students), a validation sample (10% of students), and a test sample (20% of students). To build the model, the most significant components of the digital footprints of students were selected for determining educational achievements: gender, number of friends, direction of study, user interests – subscriptions. As part of building a predicting model, the following procedures were performed: a) differentiation of students into students with high (average score above the 65th percentile border) and low (average score below the 35th percentile border) educational achievements was implemented; b) the procedure for selecting the most significant features, known as "feature selection", was performed; c) a thematic classification of user subscriptions was carried out according to the following enlarged categories: spiritual life, esotericism; education, science; entertainment, humor; art; healthy lifestyle, sports; socio-political groups; business, work; economy, technology; junk subscriptions. The final quality metric (F-measure) of the final model for predicting high educational achievements became 78% for natural, 73% for technical and 69% for humanitarian areas of study. The conducted research has fixed the specifics of the digital footprint of students with high formal educational achievements. It manifests itself in personal interests, reflected in their subscriptions, as well as in some structural characteristics of the student's Internet activity: the number of membership groups, the number of friends and subscribers. As a result, the analysis of the influence of digital footprints on educational achievements showed that now, at the junction of cognitive and cultural factors, a new type of factors is being formed that affects educational achievements – connective, associated with the peculiarities of the student's online environment.

References

1. Gorbunova, E.V. (2018) Elaboration of Research on Student Withdrawal from Universities in Russia and the United States. *Voprosy obrazovaniya – Educational Studies*. 1. pp. 110–131. (In Russian). DOI: 10.17323/1814-9545-2018-1-110-131
2. Kitchin, R. (2014) Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society*. 1(1). pp. 1–12. DOI: 10.1177/2053951714528481
3. Junco, R. (2012) The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement. *Computers and Education*. 58(1). pp. 162–171.
4. Paul, J., Baker, H. & Cochran, J. (2012) Effect of Online Social Networking on Student Academic Performance. *Computers in Human Behavior*. 28. pp. 2117–2127. DOI: 10.1016/j.chb.2012.06.016
5. Kirschner, P. & Karpinski, A. (2010) Facebook® and academic performance. *Computers in Human Behavior*. 26(6). pp. 1237–1245. DOI: 10.1016/j.chb.2010.03.024
6. Smirnov, I.B., Sivak, E.V. & Kozmina, Ya.Ya. (2016) In Search of Lost Profiles: The Reliability of VKontakte Data and Its Importance for Educational Research. *Voprosy obrazovaniya*. 4. pp. 106–122. DOI: 10.17323/1814-9545-2016-4-106-122
7. Smirnov, I. (2020) Estimating educational outcomes from students' short texts on social media. *EPJ Data Sci.* 9(27). DOI: 10.1140/epjds/s13688-020-00245-8
8. Krasilnikov, A. & Semenova, M. (2014) Do social networks help to improve student academic performance? The case of Vk.com and Russian students. *Economics Bulletin*. 34. pp. 718–733.

9. Polivanova, K.N. & Smirnov, I.B. (2017) What's in My Profile: VKontakte Data as a Tool for Studying the Interests of Modern Teenagers. *Voprosy obrazovaniya – Educational Studies*. 2. pp. 134–152. (In Russian). DOI: 10.17323/1814-9545-2017-2-134-152
10. Ikhsanov, I. & Shakhova, I. (2019) Primenenie metodov mashinnogo obucheniya dlya vyavleniya vzaimosvyazi akademicheskoy uspevaemosti i dannykh profilya sotsial'noy seti [Application of machine learning methods to identify the relationship between academic performance and social network profile data]. *Russian Digital Libraries Journal*. 2. pp. 95–118. DOI: 10.26907/1562-5419-2019-22-2-95-118
11. Goyko, V.L., Kiselev, P.B., Matsuta, V.V., Sukhanova, E.A., Stepanenko, A.A. & Feshchenko, A.V. (2017) Metody i instrumenty vyyavleniya perspektivnykh abiturientov v sotsial'nykh setyakh [Methods and tools for identifying promising applicants in social networks]. *Otkrytoe i distantsionnoe obrazovanie*. 4(68). pp. 45–52.
12. Flashman, J. (2012) Academic Achievement and Its Impact on Friend Dynamics. *Sociology of Education*. 85. pp. 61–80. DOI: 10.1177/0038040711417014
13. Smirnov, I. & Thurner, S. (2017) Formation of homophily in academic performance: Students change their friends rather than performance. *PLoS ONE*. 12(8):e0183473. DOI: 10.1371/journal.pone.0183473
14. Lu, O.H.T., Huang, A.Y.Q., Lin, A.J.Q., Ogata, H. & Yang, S.J.H. (2018) Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning. *Educational Technology & Society*. 21(2). pp. 220–232.
15. Sousa, A., Oliveira, C. & Borges, J. (2018) Using Academic Performance to Predict College Students Dropout: a case study. *Educação e Pesquisa*. 44. DOI: 10.1590/s1678-4634201844180590
16. Swail, W.S. (2014) *The Art of Student Retention. A handbook for practitioners and administrators*. [Online] Available from: https://secureservercdn.net/50.62.198.97/68g.645.myftpupload.com/wp-content/uploads/2020/01/Artofstudentretention_2008.pdf (Accessed: 20th September 2021).