УДК 004.85

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ВЫЯВЛЕНИЯ ВЗАИМОСВЯЗИ АКАДЕМИЧЕСКОЙ УСПЕВАЕМОСТИ И ДАННЫХ ПРОФИЛЯ СОЦИАЛЬНОЙ СЕТИ

И. Р. Ихсанов[1], И. С. Шахова[2]

Высшая школа информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

[1]ilias.ihsanov@gmail.com, [2]is@it.kfu.ru

Аннотация

Предложена модель машинного обучения для выявления взаимосвязи между данными профиля социальной сети и академической успеваемости учащегося, а также прогнозирования среднего балла успеваемости по данным параметрам.

Ключевые слова*: машинное обучение, социальные сети, психометрия, академическая успеваемость, образование, абитуриент*

ВВЕДЕНИЕ

Исследование, проведенное в Тинто в 1987 году, показало, что примерно 57% студентов выбирают учебное заведение, не обращая внимания на факультет обучения, а 43% студентов вуза бросают учебу, так и не получив диплом о высшем образовании. Особое внимание в исследовании уделялось факторам, влияющим на способность студента успешно закончить высшее учебное заведение. Был изучен ряд академических факторов для выявления студентов, которые с наибольшей вероятностью достигнут успеха. Исследователями была выявлена зависимость, что студенты, обладающие высокой уверенностью в себе, самообладанием, устремленностью в достижении целей связаны с более высокой успеваемостью. Кроме того, студенты, которые являются адаптивными перфекционистами, с большей вероятностью успешно завершают обучение. Таким образом, было выявлено, что личностные параметры пригодны для определения будущей успеваемости и вероятности отчисления студента из вуза [1].

Однако сбор и анализ данных о личностных характеристиках представляют собой трудозатратный процесс, так как включают в себя целый набор задач: от составления вопросов анкетирования до анализа проведенного тестирования для выявления персональных характеристик респондента.

Таким образом, данная работа нацелена на разработку модели машинного обучения для выявления взаимосвязи индивидуальных характеристик учащихся и их академической успеваемости, а также прогноза среднего балла успеваемости по данным характеристикам.

ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

Психометрия – это изучение психологических изменений личности: способностей, взглядов и качеств. В рамках психометрии спроектирована модель личности человека, состоящая из пяти черт: экстраверсии (черта характеризуется склонностью к широким социальным контактам), доброжелательности, добросовестности, невротизма (эмоциональной нестабильности) и открытости к новому опыту [2]. Личностная модель позволяет с научной точки зрения прогнозировать действия человека, делать выводы о его профессиональной пригодности, перспективах профессионального роста, возможности работы в коллективе и многое другое.

В частности, в Израиле существует единый психометрический экзамен для поступающих в вузы. Он официально рассматривается в Израильском центре экзаменов и оценок как средство прогнозирования шансов на успех в занятиях в высших учебных заведениях [3].

Подходы к автоматизации данного тестирования приведены в научной работе доктора Б. Шуотан из Китайской Национальной академии наук, где путем анализа активности в социальной сети вычисляется каждая из пяти диспозиций личности [4]. Основные параметры, собранные из социальных сетей во время исследования: возраст, родной город, частота использования социальной сети, частота загрузки материалов и многие другие. Для решения задачи классификации в данном исследовании было проведено тестирование набора данных на многих алгоритмах классификации, таких, как наивный байесовский классификатор (NB), метод опорных векторов (SVM), дерево решений и так далее. Ученые выяснили, что дерево решений C4.5 может дать лучшие

результаты [5].

Помимо указанных выше параметров, ученые собирали информацию о демографических переменных (пол, этническая принадлежность и уровень образования родителей), частоте использования Facebook (FBTime или FBCheck) и частоте действий в Facebook.

Кроме того, в рамках данного исследования удалось выявить зависимость между затраченным временем на учебную работу и временем, проведенным в социальной сети. Эти результаты согласуются с другими исследованиями, которые обнаружили, что использование интернета и, в частности, Facebook, определенным образом приводит к улучшению психосоциальных результатов, а использование Twitter определенным образом приводит к лучшим академическим результатам [6].

Ученые из Калифорнии показали, что легкодоступные цифровые записи поведения, такие, как Facebook Likes, могут использоваться для автоматического и точного прогнозирования ряда очень чувствительных личных качеств, включая этническую принадлежность, религиозные и политические взгляды, личностные качества, интеллект, счастье, факт развода родителей, возраст и пол [7]. Данный анализ основан на наборе данных о более чем 58 000 добровольцах, которые предоставили свои «лайки» в Facebook, подробные демографические профили и результаты нескольких психометрических тестов. Предложенная модель использует уменьшение размерности для предварительной обработки данных Likes, которые затем вводятся в линейную регрессию для прогнозирования отдельных психо-демографических профилей из оценок «Мне нравится» [8].

Также в рамках текущего исследования была проанализирована работа Национального исследовательского Томского государственного университета «Методы и инструменты выявления перспективных абитуриентов в социальных сетях» [9]. Результаты исследования показывают, что методика предсказания образовательных интересов и признаков одаренности по подпискам пользователей дала лучший результат. По этим параметрам есть возможность конструирования прогностической модели выявления перспективных абитуриентов через проекцию целевой модели выпускника Томского государственного университета.

ВХОДНЫЕ ДАННЫЕ

По данным немецкого аналитического агентства Statista [10], в России проникновение социальных сетей оценивается в 47%, аккаунты в них имеют 67,8 млн россиян. На рисунке 1 видно, что активнее всего в РФ используют YouTube (63% опрошенных), второе место занимает ВКонтакте — 61%. Глобальный лидер Facebook лишь на четвертой строчке с показателем в 35%. Среди мессенджеров доминируют Skype и WhatsApp по 38%.

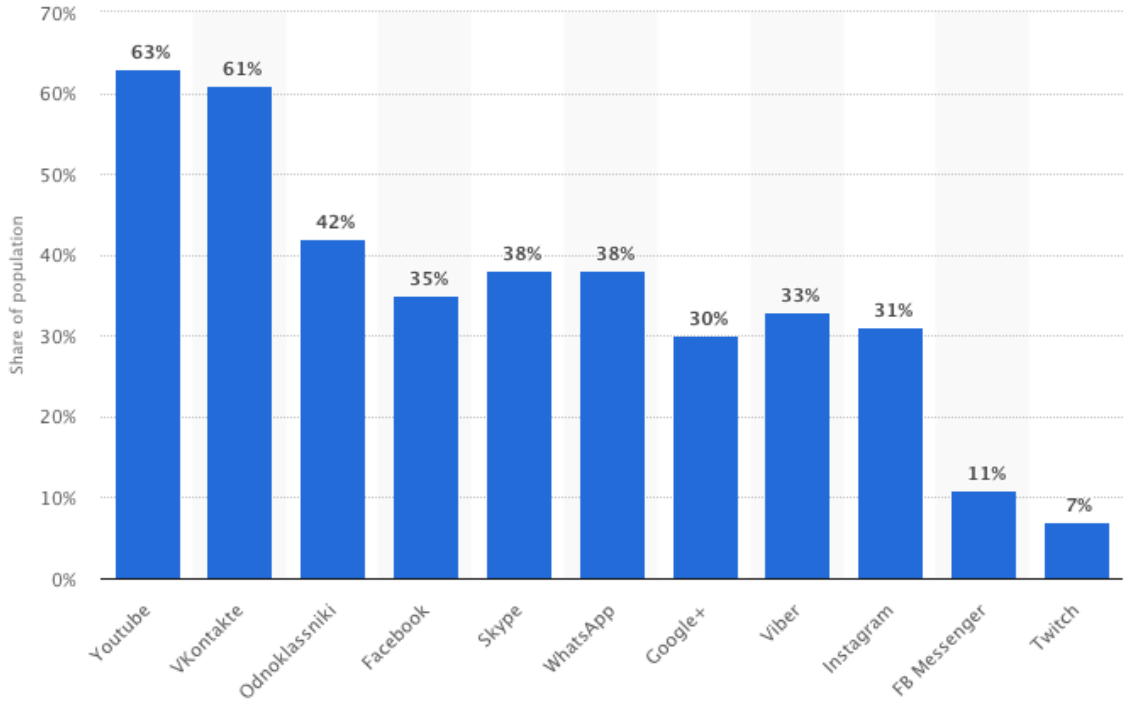

Рис. 1 – Активность использования социальных сетей в России

В исследовании [11] было выявлено, что 97,4% опрошенных студентов зарегистрировано в социальных сетях и являются активными пользователями интернета. Лидирующей социальной сетью была определена ВКонтакте [12], на нее указали 95,5% респондентов. С 94,7%, использующих социальную сеть в возрасте 15–20 лет, и 91,3% в возрасте 18–21 год. Таким образом, в качестве источника данных для текущего исследования была выбрана социальная сеть ВКонтакте.

Политика предоставления данных социальной сетью ВКонтакте дает возможность получить только те данные, которые пользователь разрешил показывать остальным пользователям ВКонтакте. Также политика

предоставления данных ВКонтакте ограничивает предоставление информации о пользователе без регистрации в системе [13]. Информация, предоставляемая ВКонтакте, которую можно использовать для первичного анализа в рамках текущего исследования, включает в себя следующие данные: пол, дата рождения, количество видео, количество альбомов, количество аудио, количество заметок, количество фотографий на странице, количество групп, количество друзей, общее количество доступных фотографий, количество подписчиков, количество интересных страниц, семейное положение, количество подписок, количество фотографий профиля, интересные страницы с тематикой страниц в порядке популярности, подарки, количество подписок на популярных личностей.

СБОР И ОБРАБОТКА ДАННЫХ

Чтобы получить сочетание ФИО, рейтинга и идентификатора профиля в ВКонтакте, вручную были проведены поиск пользователей в социальной сети и сопоставление их идентификатора с местом в рейтинге. В результате входные данные представляют собой модель, состоящую из набора: name – имя пользователя, id – уникальный идентификатор пользователя в социальной сети ВКонтакте, range – средний балл студента за год по сданным предметам в первый и второй семестры.

Таким образом, из 394 человек 354 человека были со страницами в открытом доступе и 40 человек – с приватными профилями и минимально возможными данными (количество друзей, пол, дата рождения).

Следующим этапом обработки данных стала их очистка. В полученных данных были преобразованы поля приватных аккаунтов, изменены признаки, имеющие тип объекта, в числовой тип, а также пустые значения были заменены на соответствующий тип.

На этом этапе была проведена очистка от выбросов (результат измерения, отличающийся от общей выборки). Удаление значений признаков было произведено, руководствуясь правилом экстремальных аномалий [16] по формулам:

$$IQ = (Q3 – Q1), \quad (1)$$

$$Q_н = Q1 – 3 * IQ, \quad (2)$$

$$Q_в = Q3 + 3 * IQ, \quad (3)$$

где $Q_н$ – внешний нижний забор; $Q_в$ – внешний верхний забор; IQ – интерквартильный размах; Q1 – первый квартиль, определяемый как 25-й процентиль данных; Q3 – третий квартиль, определяемый как 75-й процентиль данных.

Значения ниже нижнего внешнего забора и выше внешнего верхнего в признаках «подписки» и «количество фотографий профиля» были удалены.

После обработки аномальных значений была удалена одна колонка – количество групп, в которых состоит пользователь, так как в нем отсутствовало больше половины значений.

Следующим этапом стал разведочный анализ данных. В процессе анализа были выявлены аномальные значения – подписки студента на популярные личности. Также в процессе анализа были выявлены сильные зависимости баллов студента и его подписок на сообщества. Например, была выявлена сильная по сравнению с другими параметрами корреляция с тематикой групп, на которые подписаны студенты, и их средний балл. То есть, исходя из корреляций на рисунке 2, можно сказать, что порядковый номер страницы в списке интересных страниц имеет связь с итоговым баллом студента. Для уточнения были построены графики плотности, представляющие собой сглаженные гистограммы для демонстрации взаимосвязи тематики групп и баллов студента [17]. Фактическое влияние на итоговый средний балл студента изображено на рисунках 3–7.

Если посмотреть на графики плотности баллов студента и группы с конкретной тематикой в приоритете страниц пользователя с 1 по 5, то заметно, что студенты с интересной страницей на первом месте по приоритету с тематикой «программирование» и «соседи» учатся в диапазоне с 80 до 95

баллов. Студенты, подписанные на страницы с тематикой «образование», «видеоигры» и «креативная работа» учатся с 60 до 80 баллов. Таким образом, этот признак был в первую очередь включен в модель.

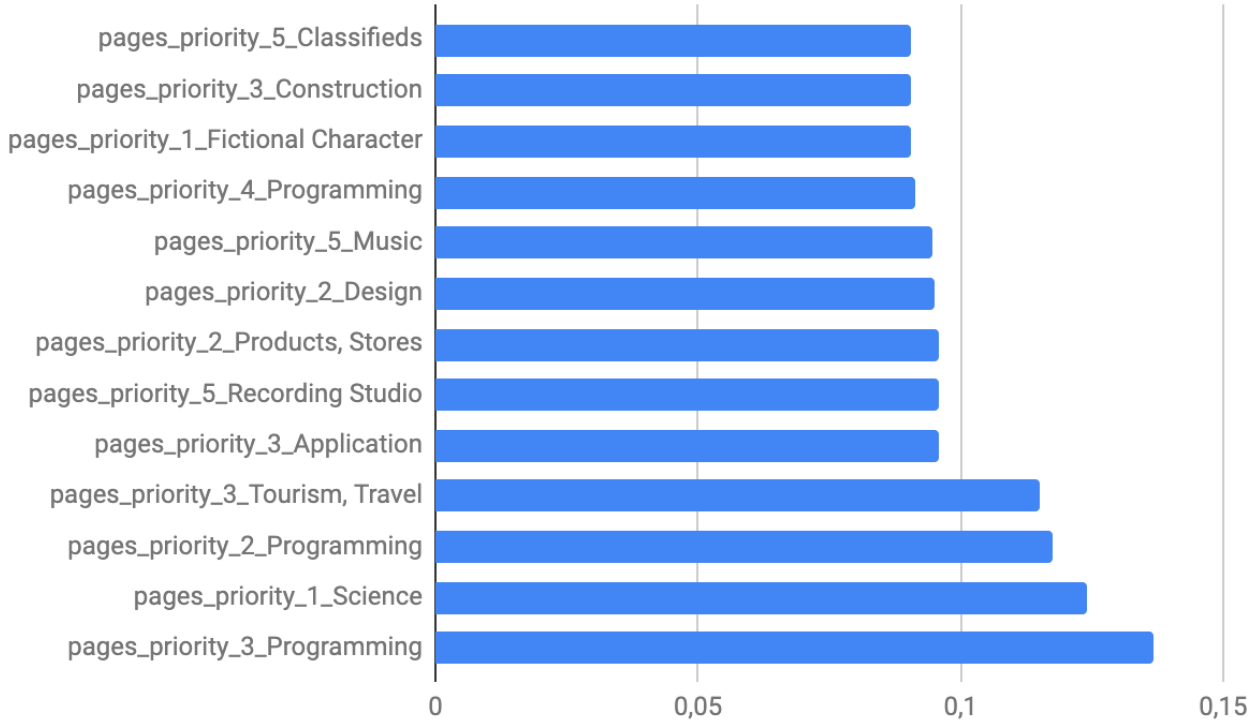

Рис. 2 – Корреляция баллов и тематик групп студента

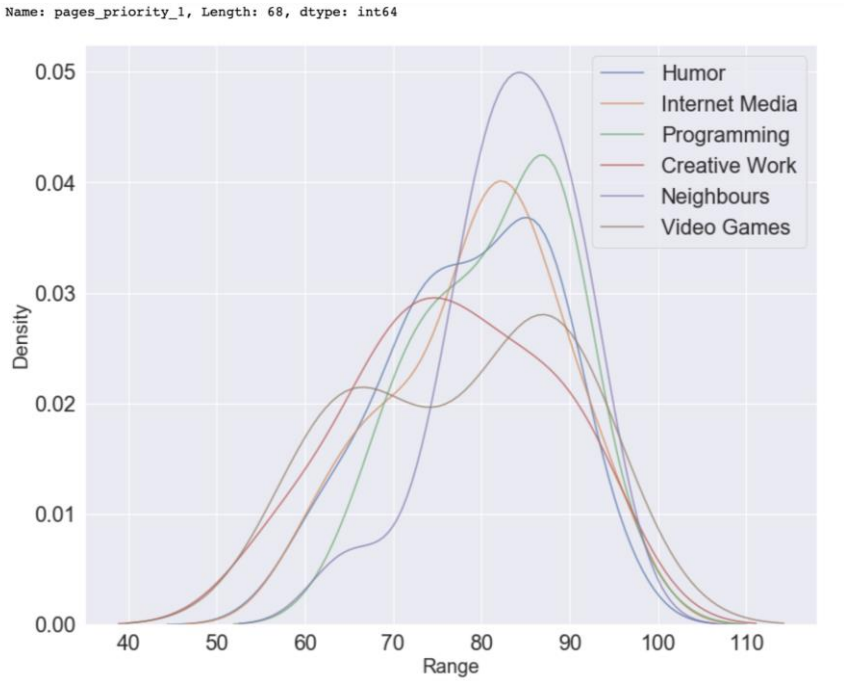

Рис. 3 – Плотность баллов студента и интересных страниц на первом месте по приоритету

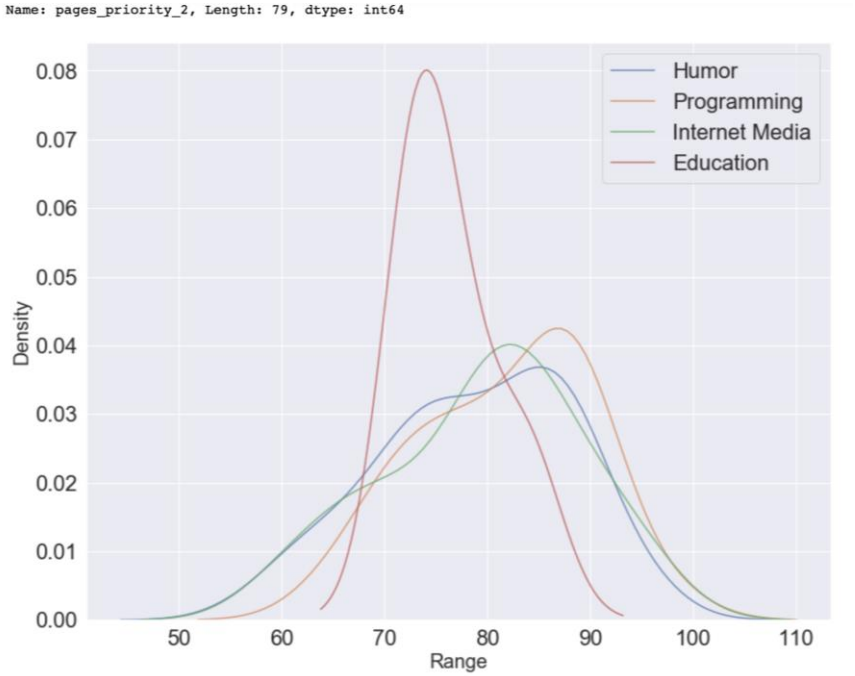

Рис. 4 – Плотность баллов студента и интересных страниц на втором месте по приоритету

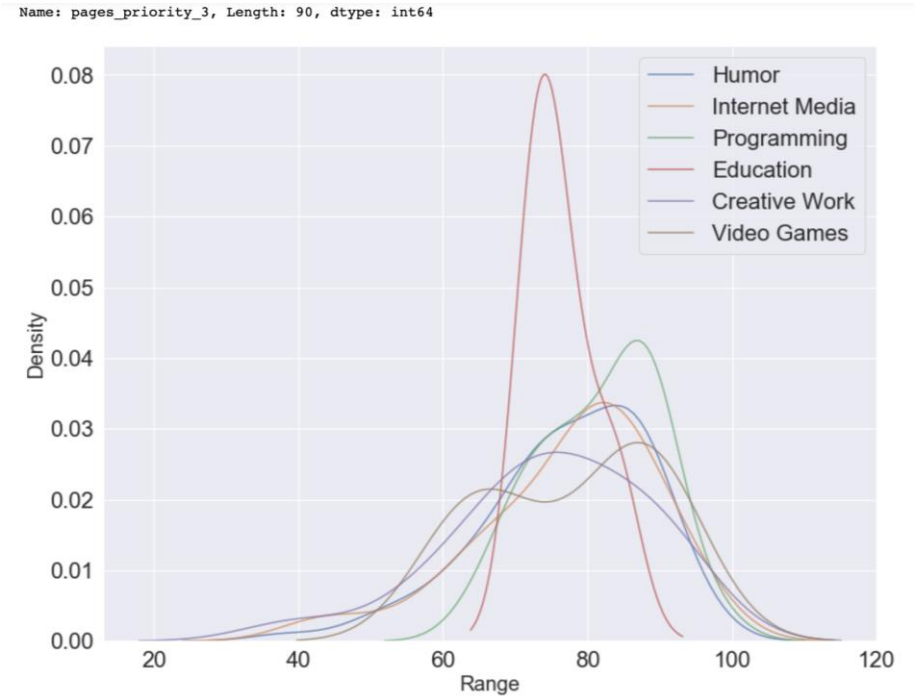

Рис. 5 – Плотность баллов студента и интересных страниц на третьем месте по приоритету

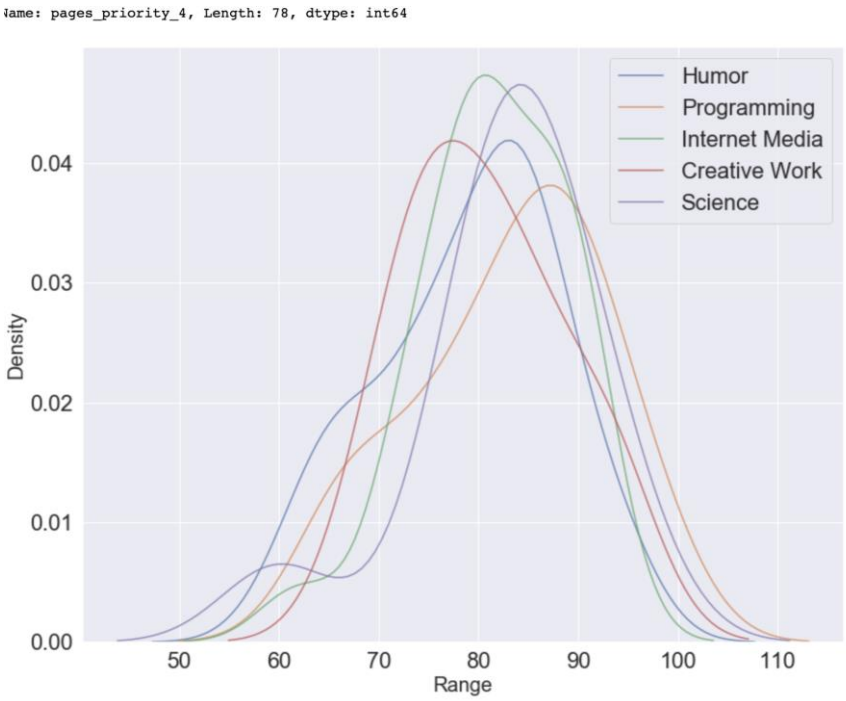

Рис. 6 – Плотность баллов студента и интересных страниц на четвертом месте по приоритету

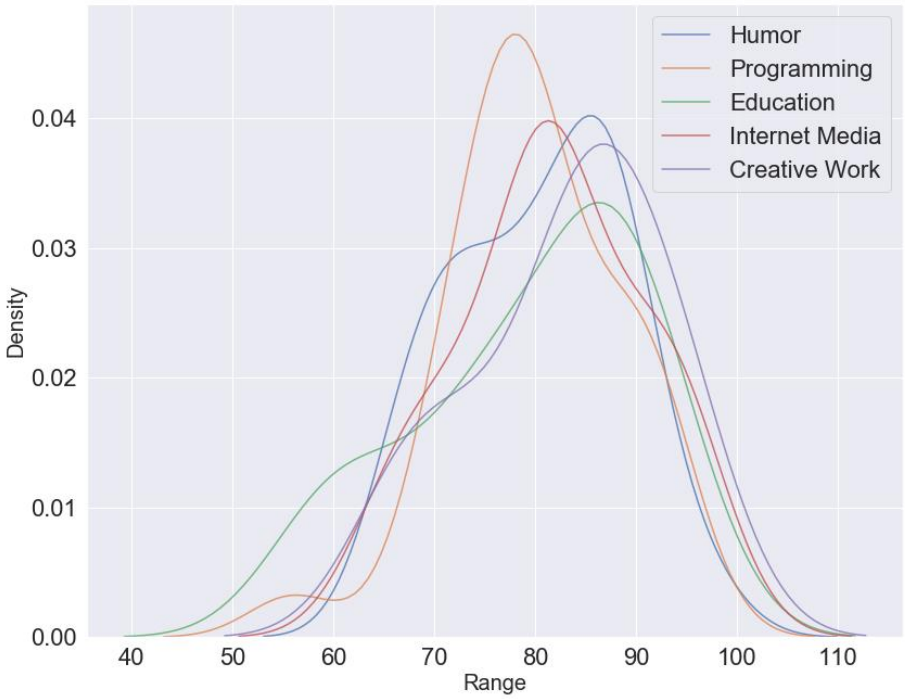

Рис. 7 – Плотность баллов студента и интересных страниц на пятом месте по приоритету

За счет конструирования (создания новых признаков из имеющихся данных) и выбора признаков (удаление лишних признаков и сохранение

коррелирующих) уменьшаются временные затраты на машинное обучение. Конструирование представляет собой получение и создание новых признаков из полученных данных. К операциям конструирования относятся извлечение натурального логарифма, применение кодирования к категориальным переменным (алгоритмы машинного обучения не работают с строковыми типами напрямую), извлечение квадратного корня. Выбор признаков заключается в процессе выбора из данных подходящих признаков. В данном процессе удаляется часть несущественных признаков, и остаются те, которые оказывают влияние на модель машинного обучения.

В процессе преобразований были выполнены кодирование категориальных переменных и извлечение натурального логарифма от числовых переменных.

В процессе обучения было замечено, что пользователи, имеющие менее 56 баллов (отчисленные), ухудшают процесс обучения модели, в результате чего средняя абсолютная ошибка для модели достигает 10,3. Было принято решение исключить этих пользователей из выборки. На начало преобразований было 15 признаков. В результате проведенных операций осталось 368 пользователей и 27 признаков: пол, количество видео, количество альбомов, количество аудио, количество заметок, количество фото, количество друзей, количество подписчиков, количество популярных личностей, количество интересных страниц, семейное положение, интересные страницы в приоритете от 1 до 5 по категориям, а также признаки, полученные при преобразовании: извлечение квадратного корня и логарифма.

Было проверено, нет ли избыточных признаков (признаки, коррелирующие друг с другом), так как удаление одного из коллинеарных признаков помогает модели быть более интерпретируемой.

Коллинеарность признаков была вычислена популярным методом фактора увеличения дисперсии. Для удаления и поиска коллинеарных признаков был использован коэффициент B-корреляции, если коэффициент корреляции между признаками был больше 0,6, то один из признаков был исключен.

Последним шагом в преобразовании данных стало масштабирование количественных признаков. Необходимость этого шага обусловлена тем, что

алгоритм градиентного бустинга чувствителен к масштабированию данных, а также значения признаков находятся в разных диапазонах. Количественные признаки: количество видео, фото, аудио, подписок, групп, подписчиков, друзей, интересных страниц, популярных личностей были нормализованы, то есть диапазон значений был приведен к формату от 0 до 1. Больше всего нормализация требуется для алгоритмов k-ближайших соседей и метода опорных векторов, так как они в своих алгоритмах учитывают евклидово расстояние между наблюдениями.

Существует два способа масштабирования объектов – это стандартизация и нормализация. Для решения данной задачи был выбран способ нормализации, потому что было замечено, что при обучении с разными способами преобразования нормализация показала лучший результат. Нормализация заключается в том, что выбирается минимальное значение признака и делится на разницу между максимальным и минимальным. Таким образом гарантируется диапазон значений от 0 до 1. На выходе были получены количественные признаки с нормальным форматом данных для модели с диапазоном от 0 до 1.

ВЫБОР И НАСТРОЙКА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

На рисунке 8 представлена связь между точностью и интерпретируемостью нескольких алгоритмов [19]. Интерпретируемость означает, что мы можем понять причины, по которым алгоритм дал конкретный ответ. Также интерпретируемость гарантирует, что модель является правильной и неправильной по определенным причинам. Понимание модели помогает улучшить ее и укрепить уверенность в том, что выбранная модель будет работать с меньшей ошибкой или с большим процентом точности.

В качестве методов машинного обучения были выбраны:

- Линейная регрессия,
- Метод k-ближайших соседей,
- «Случайный лес»,
- Градиентный бустинг,
- Метод опорных векторов.

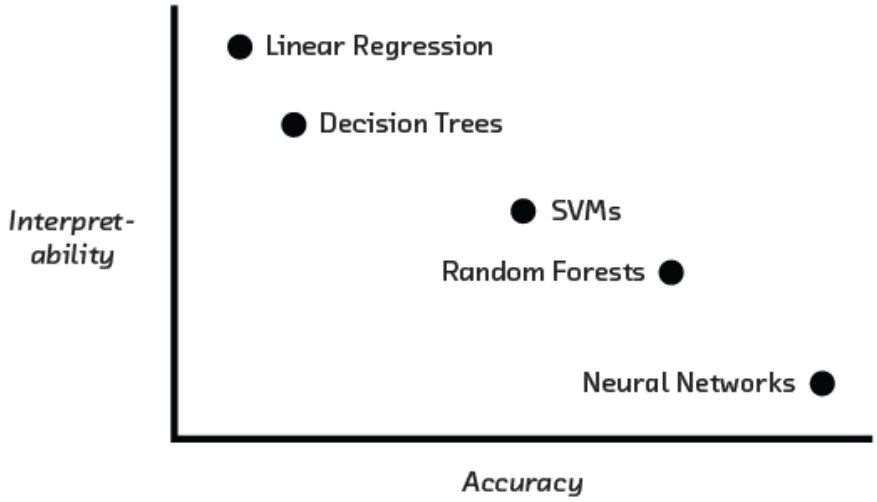

Рис. 8 – График интерпретируемости и точности алгоритмов

При обработке данных были отброшены признаки, в которых не хватает больше половины значений, однако остались признаки с меньшим количеством отсутствующих значений. Каждое пустое значение было заполнено методом медианного заполнения, который заменяет пустые значения средним значением соответствующего признака.

В результате обработанные данные были использованы для обучения модели. С помощью метрики средней абсолютной ошибки была произведена оценка качества каждой модели, а затем выбрана наиболее эффективная для дальнейшей оптимизации. Самый простой алгоритм линейной регрессии был исключен из-за значительно большего показателя средней абсолютной ошибки. Результат сравнения алгоритмов изображен на рисунке 9.

Так как базовый уровень ошибки составил 8.8829, а полученный результат оказался лучше, то можно сказать, что поставленная задача решается с помощью машинного обучения.

Лучшими моделями в сравнении оказались:

- Градиентный бустинг – 8.2285;
- К-ближайших соседей – 8.2318.

Хотя два алгоритма имеют небольшую разницу в результатах, был выбран алгоритм градиентного бустинга для его дальнейшей оптимизации.

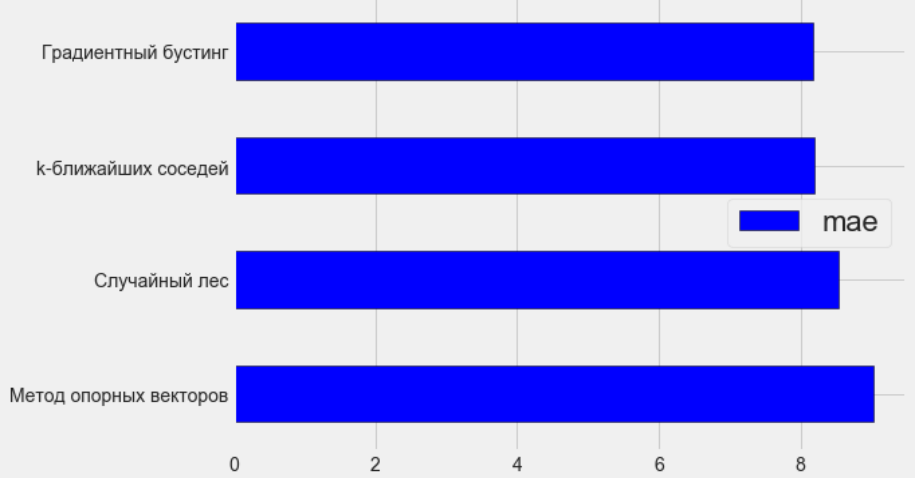

Рис. 9 – Сравнение моделей по средней абсолютной ошибке

Гиперпараметры модели могут быть рассмотрены как настройки алгоритма машинного обучения, настраиваемые перед обучением. Примерами могут служить число деревьев в случайном лесу или число соседей, используемых в регрессии k-ближайших соседей. Изменяя параметры, изменяем баланс между переобучением и недообучением. На рисунке 10 видно, как недообучение и переобучение модели влияют на прогнозы [20].

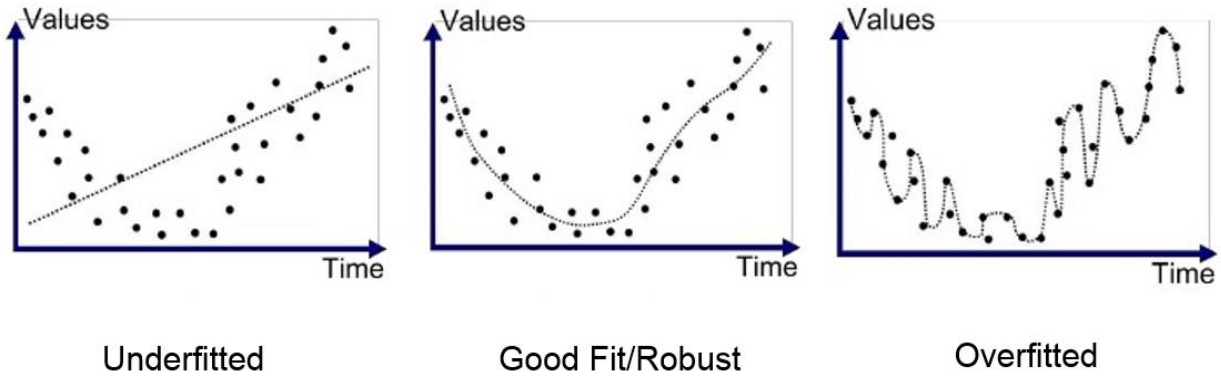

Рис. 10 – График недообучения (слева) и график переобучения (справа)

Метод настройки гиперпараметров был реализован с помощью случайного поиска с перекрестной проверкой. Перекрестная проверка – это метод оценки модели и её поведения на независимых данных. При оценке модели имеющиеся в наличии данные разбиваются на k частей. Затем на k−1 частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется k раз. Таким образом, каждая из k частей данных используется для тестирования [21].

Весь процесс перекрестной проверки был выполнен следующим образом:

- задана сетка гиперпараметров (внешняя конфигурация по отношению к модели, значения которой невозможно оценить по данным) [22];
- случайно была выбрана комбинация гиперпараметров;
- создана модель с использованием этой комбинации;
- оценивается полученная средняя ошибка работы модели с помощью перекрестной проверки;
- принято решение, какие гиперпараметры дают лучший результат.

В данном случае используется регрессионная модель на основе градиентного бустинга. Метод является сборным (состоит из множества учеников), в данном случае из отдельных деревьев решений. Ученики в градиентном бустинге обучаются последовательно, то есть каждый последующий ученик учитывает ошибки предыдущего.

В выбранной модели были настроены: способ минимизации функции потерь, количество используемых слабых деревьев решений, минимальное количество примеров в узле дерева решений, минимальное количество для разделения узла дерева решений и максимальное количество признаков для разделения узлов. Для определения лучшей настройки были использованы 29 различных гиперпараметров.

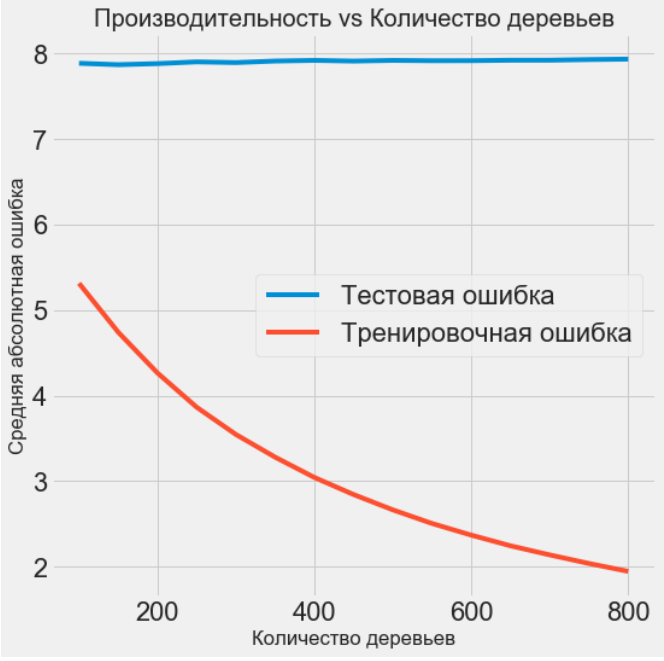

Рис. 11 – Зависимость количества деревьев и MAE

На рисунке 11 показано, как изменение количества деревьев (оценщиков) с сохранением параметров других настроек влияет на абсолютную ошибку.

Из рисунка 11 видно, что ошибка обучения значительно ниже, чем ошибка тестирования, что показывает применимость данной модели для решения поставленной задачи. Более того, по мере увеличения количества деревьев количество подгонки увеличивается. Как ошибка на тестовой выборке, так и ошибка на обучающей выборке уменьшаются по мере увеличения числа деревьев, но ошибка на обучающей выборке уменьшается быстрее.

Основываясь на результатах перекрестной проверки, лучшая модель использует 100 деревьев и достигает ошибки перекрестной проверки под 8. Это указывает на то, что средняя оценка перекрестной проверки оценки студента находится в пределах 8 баллов от истинного ответа.

Ошибка базовой модели на тестовой выборке: MAE=8.2285, ошибка финальной модели на тестовой выборке: MAE=7.9807. Гиперпараметрическая настройка помогла улучшить точность модели на 3%.

РЕЗУЛЬТАТЫ РАБОТЫ

На рисунках 12 и 13 показано, как были спрогнозированы баллы студентов по сравнению с реальными значениями. Как видно, спрогнозированные данные на финальной модели были более приближены к реальным значениям в отличие от базовой (ненастроенной модели). Также на рисунках 14 и 15 можно заметить, что количество предсказаний с большой ошибкой стало меньше по сравнению с базовой моделью. На рисунках 16 и 17 продемонстрирована плотность прогнозируемых и реальных значений. По рисункам видно, что плотность баллов в районе 80 баллов стала меньше.

Самым значимым признаком при определении оценки были: количество друзей, интересных страниц, фотографий профиля, подписчиков. На рисунке 18 видно, что среди интересных страниц большее влияние произвели группы с тематикой: программирование, юмор, креативная работа, «соседство».

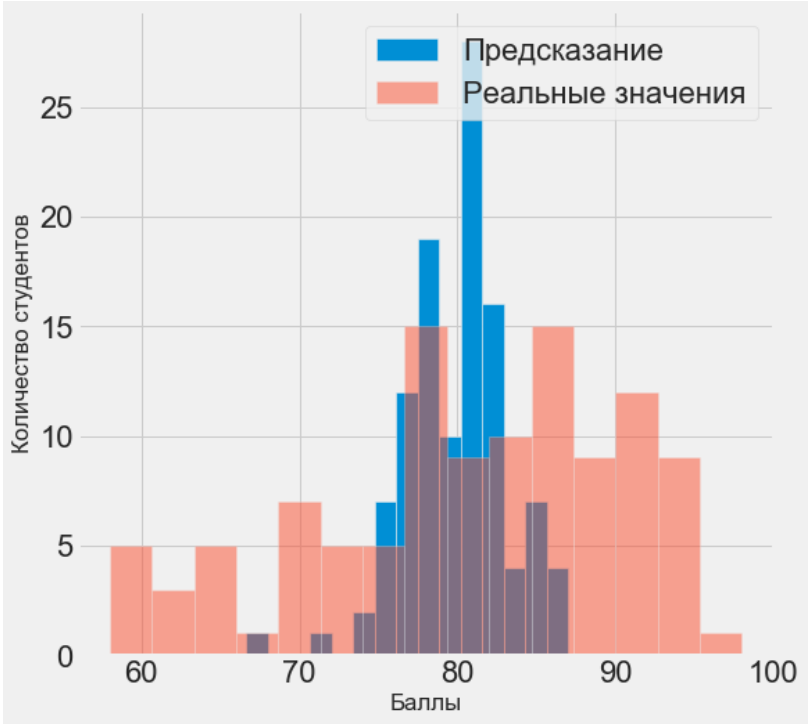

Рис. 12 – Распределение баллов на базовой модели

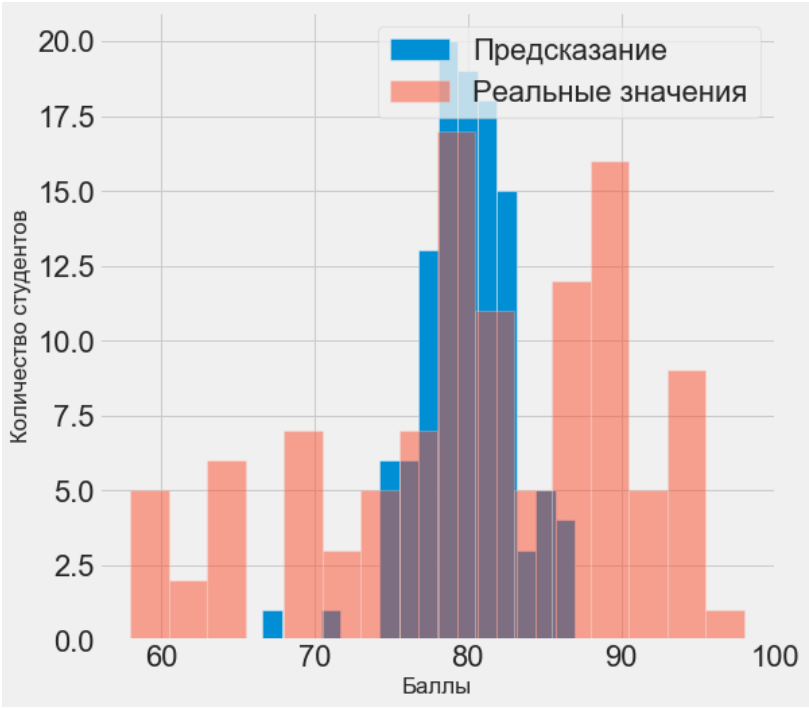

Рис. 13 – Распределение баллов финальной модели

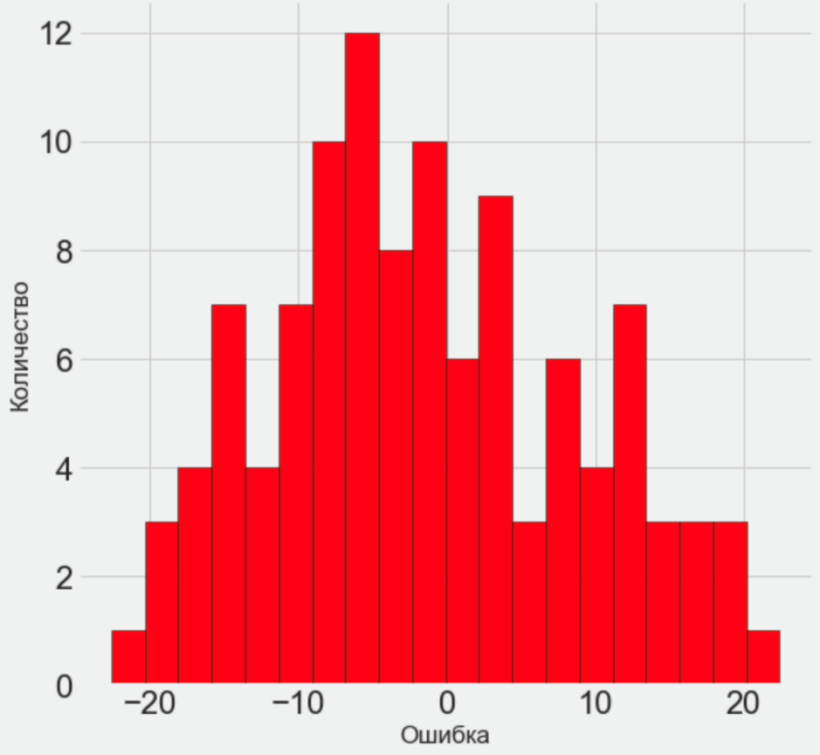

Рис. 14 – Гистограмма погрешности для базовой модели

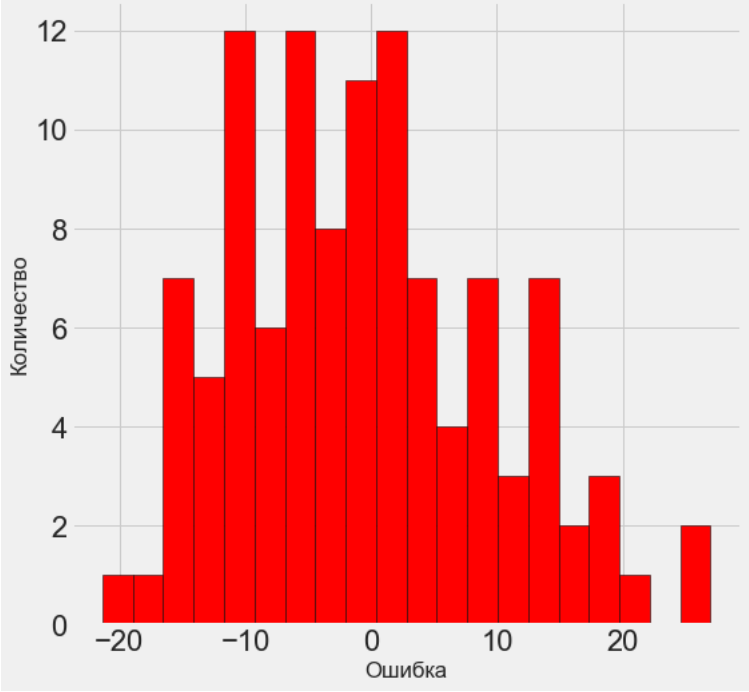

Рис. 15 – Гистограмма погрешности для финальной модели

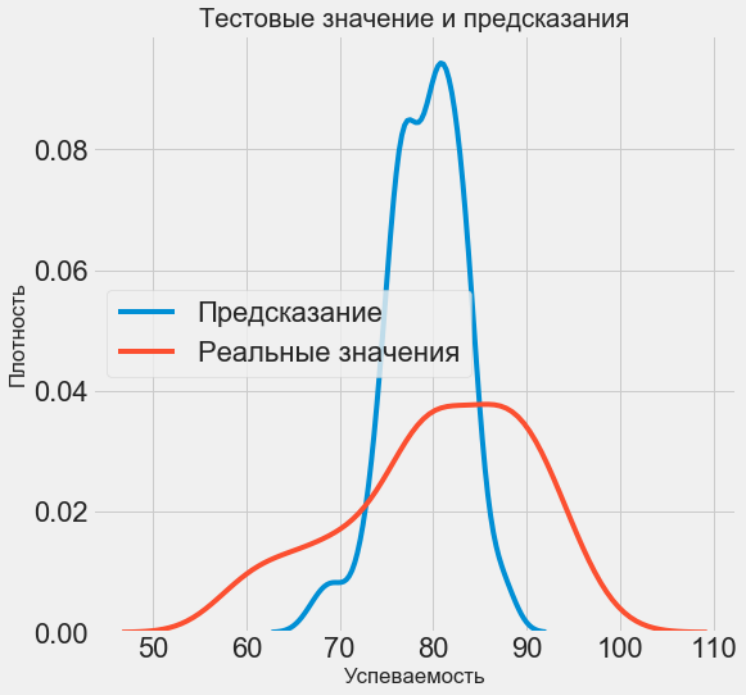

Рис. 16 – Плотность прогнозных и реальных значений базовой модели

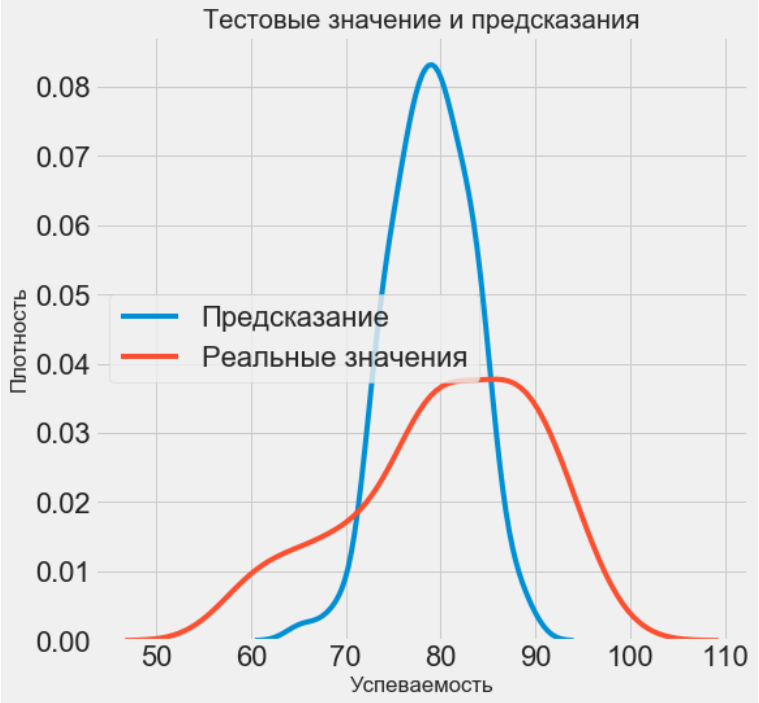

Рис. 17 – Плотность прогнозных и реальных значений финальной модели

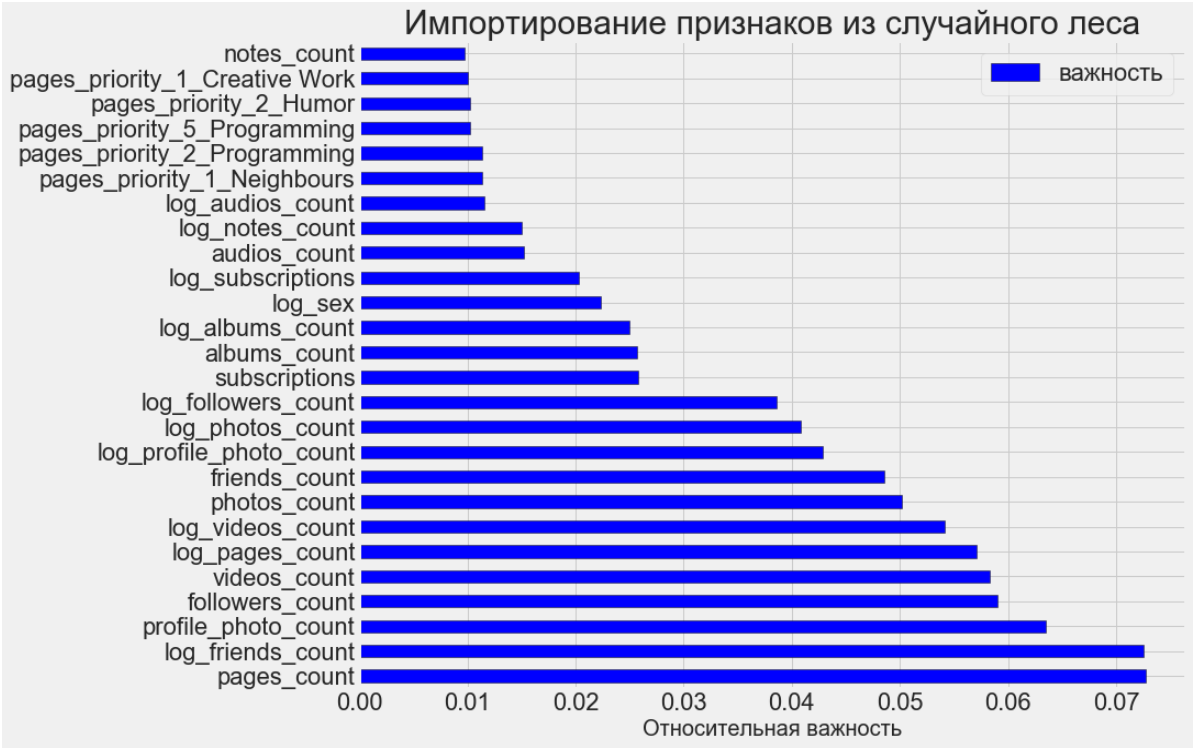

Рис. 18 – Относительная важность признаков

ЗАКЛЮЧЕНИЕ

В статье рассмотрена разработанная модель машинного обучения, которая выявляет взаимосвязь индивидуальных характеристик учащихся и их академической успеваемости, а также прогнозирует средний балл успеваемости по данным характеристикам.

В результате обучения модели были выявлены признаки, больше всего повлиявшие на составление вывода по полученным данным. К этим признакам относятся количественные признаки: подписки на популярные личности, друзья, фото профиля, а также категориальные признаки: подписки на интересные страницы.

Для улучшения качества модели в дальнейшем могут учитываться дополнительные параметры, такие, как активность в группах, время, проводимое в сети, и др. Кроме того, для получения дополнительной информации могут быть рассмотрены другие социальные сети.

Благодарности

Авторы выражают глубокую признательность младшему научному сотруднику НИЛ «Медицинская информатика» Высшей школы информационных технологий и интеллектуальных систем Алимовой Ильсеяр Салимовне за оказанную помощь в проведении данного исследования.

СПИСОК ЛИТЕРАТУРЫ

1. *Pritchard M.* Using Emotional and Social Factors to Predict Student Success // Journal of College Student Development. 2003. V. 44, No 1. P. 18–28.

2. *What your Facebook likes say about you*. URL: https://www.cbc.ca/ news/technology/facebook-likes-like-a-gift-1.3893298.

3. *Психометрический вступительный экзамен в Израиле* // Официальный сайт путеводителя по Израилю. URL: https://guide-israel.ru/country/ 37376-psixometricheskij- vstupitelnyj- ekzamen/.

4. *Shuotian B., Tingshao Z., Li C.* Big-Five Personality Prediction Based on User Behaviors at Social Network Sites // Cornell University, Tech. Rep. 2012.

5. *Friedrichsen M., Mühl-Benninghaus W.* Handbook of Social Media Managment Value Chain and Business Models in changing media marketing. Springer-Verlag Berlin Heidelberg, 2013. 880 p.

6. *Junco R.* Too much face and not enough books: The relationship between multiple indices of Facebook use and academic performance // Computers in Human Behavior. 2012. V. 28, No 1. P. 187–198.

7. *Junco R.* The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement Received // Magazine Computers & Education. 2012. V. 58, No 1. P. 162–171.

8. *Kosinski M., Stillwell D., Graepel T.* Private traits and attributes are predictable from digital records of human behavior // Magazine PNAS. 2013. V. 110, No 15. P. 5802–5805.

9. *Мацута В.В., Киселев П.Б., Фещенко А.Б., Гойко В.Л., Сузанова Е.А., Степаненко А.А.* Методы и инструменты выявления перспективных абитуриентов в социальных сетях // Открытое и дистанционное образование. 2017. № 4. С. 45–52.

10. *Penetration of leading social networks in Russia as of 4th quarter 2017* //

Statistica. URL: https://www.statista.com/statistics/284447/russia-social-network-penetration/.

11. *Мотивы проявления студентами колледжей социальной активности в социальных сетях: регионального аспекта* // Электронный научный архив УрФУ. URL: http://elar.urfu.ru/bitstream/ 10995/59123/1/978-5-91256-403-1_2018_053.pdf.

12. *Социальная сеть Вконтакте.* URL: https://vk.com/.

13. *Политика конфиденциальности VK.com // Социальная сеть Вконтакте.* URL: https://vk.com/privacy.

14. *VK.com python API wrapper* // GitHub. URL: https://github.com/ voronind/vk.

15. *Kaggle.* URL: https://www.kaggle.com/.

16. *What are outliers in the data* // Engineering statistics handbook. URL: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm.

17. *Histograms and density plots in python* // Towards data science. URL: https://towardsdatascience.com/histograms-and-density-plots-in-python-f6bda88f5ac0.

18. *How to normalize and standardize your machine learning data in weka* // Machine learning mastery. URL: https://machinelearningmastery.com/normalize-standardize-machine-learning-data-weka/.

19. *Generalized Linear Models* // Scikit-learn. URL: https://scikit-learn.org/ stable/supervised_learning.html.

20. *Overfitting vs underfitting: a conceptual explanation* // Towards data science. URL: https://towardsdatascience.com/overfitting-vs-underfitting-a-conceptual-explanation-d94ee20ca7f9.

21. *Что такое кросс-валидация* // Data Science. URL: http://datascientist. one/cross-validation/.

22. *What is the difference between a parameter and a Hyperparameter?* // Machine Learning Mastery. URL: https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/.

MACHINE LEARNING METHODS FOR DETERMINING THE RELATIONSHIP BETWEEN ACADEMIC SUCCESS AND DATA OF SOCIAL NETWORK PROFILE

I. R. Ikhsanov[1], I. S. Shakhova[2]

Higher School of Information Technologies and Intelligent Systems, Kazan (Volga region) Federal University

[1]ilias.ihsanov@gmail.com, [2]is@it.kfu.ru

Abstract

The paper is aimed to propose the machine learning model for determining the relationship between data of social network profile and academic success of students and predicting the success using the data.

Keywords: *machine learning, social networks, psychometrics, academic success, education, abiturient*

REFERENCES

1. *Pritchard M.* Using Emotional and Social Factors to Predict Student Success // Journal of College Student Development. 2003. V. 44, No 1. P. 18–28.

2. *What your Facebook likes say about you*. URL: https://www.cbc.ca/news/technology/facebook-likes-like-a-gift-1.3893298.

3. *Psihometricheskij vstupitel'nyj ekzamen v Izraile //* Oficial'nyj sajt putevoditelya po Izrailyu. URL: https://guide-israel.ru/country/37376-psixometricheskij-vstupitelnyj-ekzamen/.

4. *Shuotian B., Tingshao Z., Li C.* Big-Five Personality Prediction Based on User Behaviors at Social Network Sites // Cornell University, Tech. Rep. 2012.

5. *Friedrichsen M., Mühl-Benninghaus W.* Handbook of Social Media Managment Value Chain and Business Models in changing media marketing. Springer-Verlag Berlin Heidelberg, 2013. 880 p.

6. *Junco R.* Too much face and not enough books: The relationship between multiple indices of Facebook use and academic performance // Computers in Human Behavior. 2012. V. 28, No 1. P. 187–198.

7. *Junco R.* The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement Received // Magazine

Computers & Education. 2012. V. 58, No 1. P. 162–171.

8. *Kosinski M., Stillwell D., Graepel T.* Private traits and attributes are predictable from digital records of human behavior // Magazine PNAS. 2013. V. 110, No 15. P. 5802–5805.

9. *Macuta V.V., Kiselev P.B., Feshchenko A.B., Gojko V.L., Suzanova E.A., Stepanenko A.A.* Metody i instrumenty vyyavleniya perspektivnyh abiturientov v social'nyh setyah // Otkrytoe i distancionnoe obrazovanie. 2017. No 4. S. 45–52.

10. *Penetration of leading social networks in Russia as of 4th quarter 2017* // Statistica. URL: https://www.statista.com/statistics/284447/russia-social-network-penetration/.

11. *Motivy proyavleniya studentami kolledzhej social'noj aktivnosti v social'nyh setyah: regional'nogo aspekta* // Elektronnyj nauchnyj arhiv UrFU. URL: http://elar.urfu.ru/bitstream/ 10995/59123/1/978-5-91256-403-1_2018_053.pdf.

12. *Vkontakte.* URL: https://vk.com/.

13. *Politika konfidencial'nosti VK.com //* Vkontakte. URL: https://vk.com/privacy.

14. *VK.com python API wrapper* // GitHub. URL: https://github.com/voronind/vk.

15. *Kaggle.* URL: https://www.kaggle.com/.

16. *What are outliers in the data* // Engineering statistics handbook. URL: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm.

17. *Histograms and density plots in python* // Towards data science. URL: https://towardsdatascience.com/histograms-and-density-plots-in-python-f6bda88f5ac0.

18. *How to normalize and standardize your machine learning data in weka* // Machine learning mastery. URL: https://machinelearningmastery.com/normalize-standardize-machine-learning-data-weka/.

19. *Generalized Linear Models* // Scikit-learn. URL: https://scikit-learn.org/stable/supervised_learning.html.

20. *Overfitting vs underfitting: a conceptual explanation* // Towards data science. URL: https://towardsdatascience.com/overfitting-vs-underfitting-a-conceptual-explanation-d94ee20ca7f9.

21. *Chto takoe cross-validatsiya* // Data Science. URL: http://datascientist.one/cross-validation/.

22. *What is the difference between a parameter and a Hyperparameter?* // Machine Learning Mastery. URL: https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/.

СВЕДЕНИЯ ОБ АВТОРАХ

ИХСАНОВ Ильяс Раисович – бакалавр Высшей школы информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета по направлению «Прикладная информатика».

Ilias Raisovich IKHSANOV, Bachelor of Science in Applied Informatics from the Higher School of Information Technologies and Intelligent Systems, Kazan (Volga region) Federal University.

email: ilias.ihsanov@gmail.com

ШАХОВА Ирина Сергеевна – ассистент кафедры программной инженерии Высшей школы информационных технологий и интеллектуальных систем Казанского федерального университета. Сфера научных интересов – мобильные приложения, цифровые образовательные системы, индивидуализация образования, мобильное обучение.

Irina Sergeevna SHAKHOVA – teacher of the Higher School of Information Technologies and Intelligent Systems, Kazan Federal University. Research interests include mobile applications, digital educational systems, individualization in education, mobile learning.

email: is@it.kfu.ru

Материал поступил в редакцию 20 июня 2019 года