***Романов Алексей Алексеевич***
бакалавр, магистрант
ФГАОУ ВО «Российский университет
дружбы народов»
г. Москва
***Кожевникова Елена Алексеевна***
бакалавр, учитель
ГБОУ г. Москвы «Школа №1285»
г. Москва

*DOI 10.31483/r-97403*

# ОПРЕДЕЛЕНИЕ ОБРАЗОВАТЕЛЬНОГО ПОТЕНЦИАЛА УЧАЩЕГОСЯ НА ОСНОВЕ ЕГО СОЦИАЛЬНЫХ ХАРАКТЕРИСТИК ПРИ ПОМОЩИ МАШИННОГО ОБУЧЕНИЯ

***Аннотация****: работа посвящена исследованию эффективности машинного обучения как ресурса повышения качества образования. Машинное обучение основано на разработке таких компьютерных программ и алгоритмов, которые способны самостоятельно обучаться и адаптироваться при подаче новых данных. В исследовании было проведено практическое использование описанных технологий и сделан вывод, что технологии машинного обучения способны корректно оценить образовательный потенциал учащегося.*

***Ключевые слова****: машинное обучение, дифференцированный подход, градиентный бустинг, прогнозирование результата обучения, математическая статистика, повышение качества образования.*

В распоряжении школы имеются большие базы данных об учениках и их родителях. Сотрудники образовательного учреждения анализируют информацию о семьях учащихся, об условиях, в которых живет и воспитывается ребенок, внимательно изучают данные психолого-педагогических мониторингов, чтобы спрогнозировать уровень усвоения ребенком учебной информации. Прогнозирование необходимо в первую очередь для дифференцированного подхода к

обучению каждого ученика с учетом его индивидуальных особенностей. Как показывает практика, когда родители и педагоги ознакомлены с прогнозами будущей успеваемости ученика, реальная успеваемость выше прогнозируемой.

Одной из важнейших задач школы является определение потенциала каждого ученика. Необходимо создать условия для воспитания, становления и формирования личности обучающегося, для развития его склонностей, интересов и способности к социальному самоопределению. В современной школе образование должно способствовать всестороннему развитию каждого ребенка. В силу своих индивидуальных психофизических особенностей, каждый ребенок нуждается в дифференцированном подходе в обучении. А для достижения наилучших результатов необходим такой инструмент, как прогнозирование успешности обучения. Современные исследования в данной области показали эффективность данного метода. Так, согласно исследованиям кандидата психологических наук Н.Е. Подгайского [1], если педагогам и родителям предоставлялась информация о развитии ребенка и прогноз его будущей успеваемости, то реальная успеваемость была выше. Для прогнозирования успешности учеников ученым были взяты следующие критерии: визуальное линейное мышление, визуальное структурное мышление, работоспособность, социальное положение матери, образование матери, жилищные условия, тяготение к отцу, потребность в общении, креативность, практичность. В результате проведенных исследований, Н.Е. Подгайский объединил детей в группы по схожим признакам. Каждой группе учащихся были составлены индивидуальные рекомендации. По истечении двух месяцев анализ результатов показал рост реальной успеваемости.

Интенсивная разработка вычислительной техники и онлайн-технологий привела к появлению нового направления работы с большими объёмами данных – машинному обучению (англ. machine learning). Данная технология способна существенно сократить время психологических и социальных исследований. Машинное обучение – это инструмент для решения множества подобных вычислительных задач. В его основе лежат методы математической статистики, численных методов, теории вероятности, теории графов и других

математических дисциплин, которые позволяют анализировать большие объёмы данных и создавать математические модели базирующиеся на них.

Целью данной работы является проверка эффективности прогнозирования образовательного потенциала учащегося на основе его социальных характеристик при помощи машинного обучения. На основе машинного обучения можно разработать прогнозирующую модель. Для обучения соответствующей модели специалисту необходим развёрнутый набор данных. Набор данных состоит из множества независимых признаков объектов и искомых целевых признаков. Между представленными признаками и искомой величиной существует некоторая зависимость, характеристики которой неизвестны. На основе этих данных алгоритмы машинного обучения способны установить неявные связи и построить модель, которая предскажет искомую величину исходя из его характеристик. Данный раздел машинного обучения можно отнести к категории «обучение с учителем».

На данный момент для решения подобных задач используются методы решающего древа, случайного леса, линейной регрессии, градиентного бустинга и другие алгоритмы. Градиентный бустинг основан на итеративном алгоритме поиска минимума функции потерь с последовательным построением набора моделей через улучшение предсказаний.

В рамках текущей работы не стоит цели минимизации абсолютной ошибки определения результата, а только проверка адекватности и возможности применения технологии. Для проверки гипотезы будет использована технология CatBoost (от англ. categorial boosting – бустинг для категориальных признаков). В качестве рабочего материала воспользуемся базой данных из открытого источника [2]. В нашем распоряжении данные о итоговой успеваемости 649 учащихся старших классов двух португальских школ (таблица 1).

Распределение итоговых оценок учеников после обработки данных отображено на рисунке 1. Стандартное отклонение в выборке составляет 2,66.

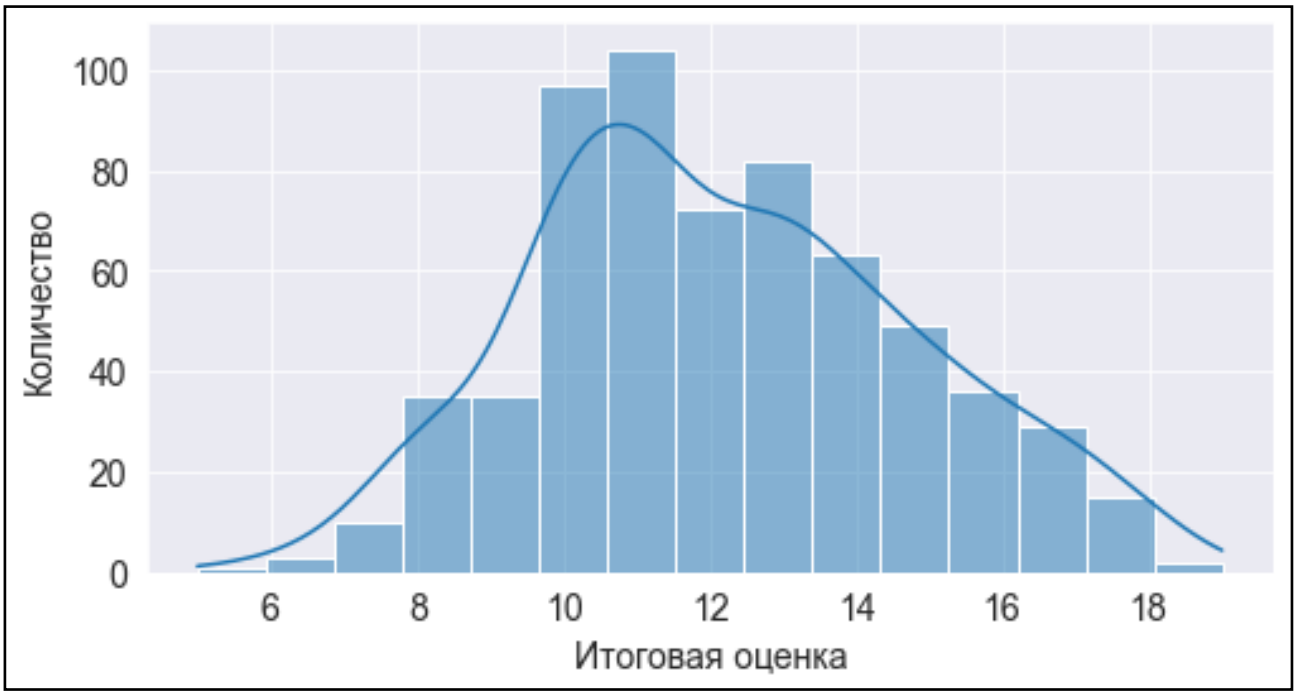


Рис. 1. Столбчатая диаграмма распределения итоговых оценок в наборе данных

Все описанные признаки были использованы при разработке модели машинного обучения. Вычислительное устройство воспринимает только числовую информацию. Всем категориальным признакам был присвоен номер категории с применением технологии Ordinal Encoding. При обучении модели вычислительная машина по умолчанию не может понять величину значимости того или иного признака, поэтому все числовые значения были стандартизированы с применением технологии Standard Scaler. Стандартизация производится по следующей формуле:

$$\text{Новое значение} = \frac{\text{Старое значние-Среднее значение}}{\sqrt{\text{Дисперсия}}} \quad (1)$$

Таблица 1

Описание рабочего материала

| *Признаки, искомый признак* | *Описание* | *Значения* |
|---|---|---|
| Школа | Одна из двух школ | 1, 2 |
| Пол | Бинарное | М, Ж |
| Возраст | | 15 – 22 |
| Место проживания | Тип адреса (город / сел. мес.) | Г / С |
| Размер семьи | Бинарное | Меньше или равно 3, больше 3 |
| Совместное проживание родителей | | Совместно, раздельно |

| | | |
|---|---|---|
| Образование матери | | отсутствует, начальное, среднее, высшее |
| Образование отца | | Аналогично матери |
| Работа матери | | «сфера образовательная», «сфера здравоохранения», «гражданские услуги», «государственная служба», «безработная», «иная сфера» |
| Работа отца | | Аналогично матери |
| Причина выбора школы | Оценка учащегося | («близость к дому», «репутация школы», «предпочтение конкретного предмета», «иное» |
| Представитель ребёнка | При поступлении | «мать», «отец», «иное» |
| Время пути до школы | | «менее 15 минут», «15–30 минут», «30–60 минут», «более часа» |
| Время самостоятельного обучения | Оценка учащегося | «менее 2 часов», «2–5 часов», «5–10 часов», «более 10 часов» |
| Количество проваленных срезов (за всё время) | | 0 – 3, 4 и более |
| Доп. обр. поддержка | Бинарное | Да / Нет |
| Семейная обр. поддержка | Бинарное | Да / Нет |
| Платные занятия | Бинарное | Да / Нет |
| Секции/Кружки | Бинарное | Да / Нет |
| Посещение детского сада | Бинарное | Да / Нет |
| Желание высшего образования | Бинарное | Да / Нет |
| Наличие интернета | Бинарное | Да / Нет |
| Наличие романтических отношений | Бинарное | Да / Нет |
| Качество семейных отношений | Оценка учащегося | 1 – 5 |
| Время прогулок с друзьями | Оценка учащегося | 1 – 5 |
| Алкоголь в будние дни | Оценка учащегося | 1 – 5 |
| Алкоголь в выходные дни | Оценка учащегося | 1 – 5 |
| Состояние здоровья | Оценка учащегося | 1 – 5 |
| Количество пропусков | | 0 – 93 |
| | | |
| Итоговая оценка | Искомая величина | 0 – 20 |

Для оценки качества модели воспользуемся величиной средней абсолютной ошибкой (англ. Mean Absolute Error):

$$\text{MAE} = \frac{1}{\text{Кол.объектов}} \sum_{i=1}^{\text{Кол.объектов}} |\text{значение}_i\text{-прогноз}_i| \quad (2)$$

После обучения модели методом CatBostRegressor с iterations 1000, learning_rate 0,01, random_state 25, subsample 0,8 значение величины средней абсолютной ошибки составило 1,71 единицы. Если данную величину перевести к российскому образцу, то ошибка составила бы 0,34 единицы. При этом константная медианная модель показала бы точность вычислений равной 2,11 единицы. Полученные результаты адекватны.

По итогам разработки модели также была исследована величина влияния признаков на прогноз (рисунок 2). Следует отметить, что влияние данных признаков не указывает направление корреляционной связи.

Полученные значения величины влияния стоит рассматривать исключительно в рамках текущей задачи. Каждый обучающий набор данных должен подбираться исходя из региона исследования, возрастной группы учащихся и цели исследования. Также при наборе данных следует учитывать принцип мультиколлериальности признаков для понижения сложности интерпретации параметров и их надёжности.

Из изложенного можно сделать вывод, что применение машинного обучения в определении образовательного потенциала учащегося обоснованно и может быть использовано в рабочем процессе. Для разработки соответствующей модели в рамках учреждения или географического района необходимо набрать достаточный набор информационных данных о успеваемости учеников и их характеристиках.

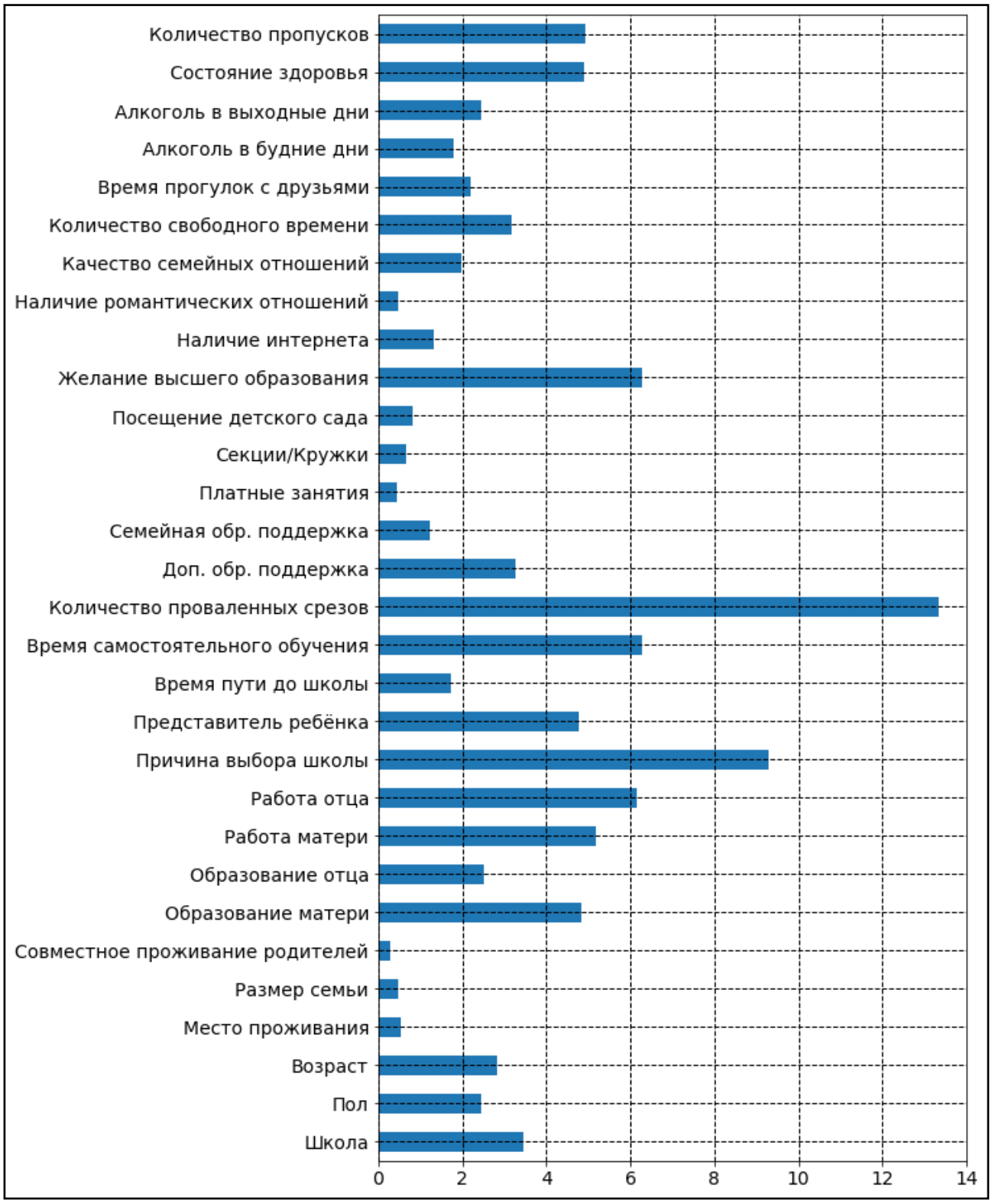


Рис. 2. Процентное влияние исследуемых признаков на модель машинного обучения

***Список литературы***

1. Подгайский Н.Е. Психолого-педагогическое прогнозирование успешности обучения первоклассников: дис. … канд. психол. наук / Н.Е. Подгайский; ГОУ ВПО «Нижегородский педагогический университет».

2. Student Performance Data Set [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/larsen0966/student-performance-data-set (дата обращения: 12.12.2020).

3. Bishop C.M. Pattern Recognition and Machine Learning. – Springer, 2006.