Применение методов машинного обучения для оценки абитуриентов

А. А. Тимофеев
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
chelent0ss@gmail.com

***Аннотация*. В докладе рассмотрено применение методов машинного обучения для решения задачи оценки абитуриентов. Описано создание модели по предсказанию успешности обучения абитуриента в ВУЗе. Приведены результаты экспериментов с различными алгоритмами машинного обучения и предобработкой данных. Произведена оценка разработанной модели на основе данных последней сессии.**

Ключевые слова: машинное обучение; абитуриент; сессия; деревья решений

I. Введение

Большой прогресс в сфере искусственного интеллекта внес сильные изменения в нашу повседневную жизнь. Машинное и глубокое обучение нашли себе применение в самых разнообразных областях. С развитием электронного обучения системы с искусственным интеллектом начали использоваться и в сфере образования.

Каждый год в ВУЗы направляет документы большое количество абитуриентов, что создает большую нагрузку на приемную комиссию. Когда желающих поступить на одно направление много, а количество мест ограничено, важно правильно выстраивать приоритеты между поступающими.

В данной работе рассматриваются возможности применения методов анализа данных и машинного обучения для оценки абитуриентов, которая будет заключаться в предсказании того, сдаст ли абитуриент экзамен по одной из будущих дисциплин, или нет. Полученная оценка может быть использована при последующей работе с ними, как при организации приема, так и при дальнейшем ведении их процесса обучения.

II. Существующие решения и аналоги

Существует множество исследований по теме предсказания успехов в обучении студентов различных учебных заведений. Основные различия между ними наблюдаются в данных, используемых при создании модели и алгоритмах обучения модели. Некоторые исследования сосредоточены на предсказании результатов студентов в первом семестре обучения, другие выполняют предсказания результатов на протяжении всего процесса обучения. Так как в нашей работе мы имеем дело с предсказанием успехов именно абитуриентов, то есть результатов в первом семестре обучения, проведем обзор исследований, решающих схожую задачу.

Авторы [1] изучили эффективность различных критериев приема, принятых университетами Пакистана, а в частности Университета информационных технологий (ITU) в Лахоре, построив при этом модель, предсказывающую средний балл диплома. Сперва они провели корреляционный анализ, чтобы выявить, от каких данных успеваемость студента зависит сильнее. Далее авторы использовали такие методы, как обучающее векторное квантование, рекурсивное удаление признаков, а также дерево решений, для определения значимости признаков в существующей выборке. В итоговой модели множественной линейной регрессии использовался набор признаков, состоящий из сертификата о высшем среднем образовании, сертификата о среднем образовании, результатов вступительного теста, а также результатов собеседования. Итоговые результаты регрессии показали, что средний балл диплома может быть адекватно предсказан по сертификату о высшем среднем образовании и результатам вступительного теста.

В исследовании [2] авторы использовали методы машинного обучения для предсказания результатов студентов по курсам английского языка и математики, с целью улучшить систему распределения учеников между коррекционными курсами и курсами переходного уровня. Процесс отбора признаков начался с включения 50 и 180 признаков в обучение моделей для курсов по английскому языку и математике соответственно. Далее авторы итеративно исключали признаки, которые не представляли значимости для модели. В итоге для модели по курсу английского языка остались следующие признаки: общий средний балл (среднее значение всех средних баллов, которые студент получил за все семестры и все курсы в учебном заведении), средний балл по предмету (средний балл всех оценок, полученных учащимся по английскому языку) и средний балл без предмета (средний балл по всем оценкам, полученным учащимся за семестр или четверть за вычетом английского языка). Для модели по курсу математики к указанным выше признакам добавились признаки, характеризующие общую успеваемость учеников с 9-й по 12-й классы, а также успеваемость по математике за этот же промежуток обучения. Среди алгоритмов машинного обучения рассматривались логистическая регрессия, наивный байесовский классификатор, дерево решений, дерево решений с градиентным ускорением, случайный лес, метод опорных векторов и нейронная сеть с двумя скрытыми слоями. Наилучший результат в предсказании того, пройдет ли студент курс, или нет, показала модель,

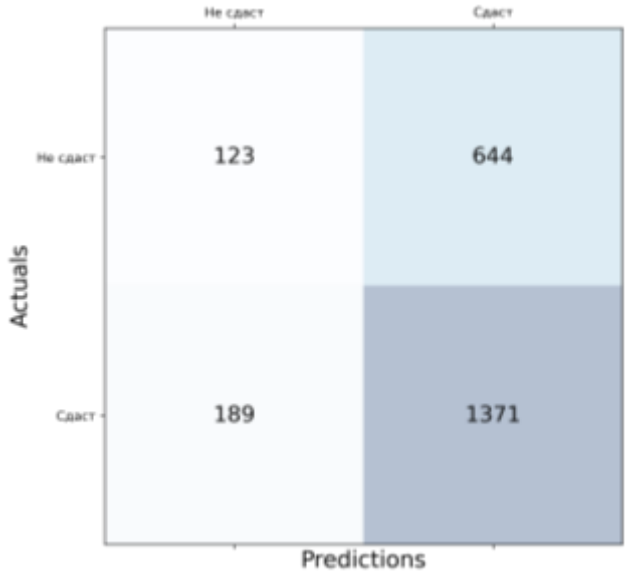

Рис. 1. Матрица ошибок

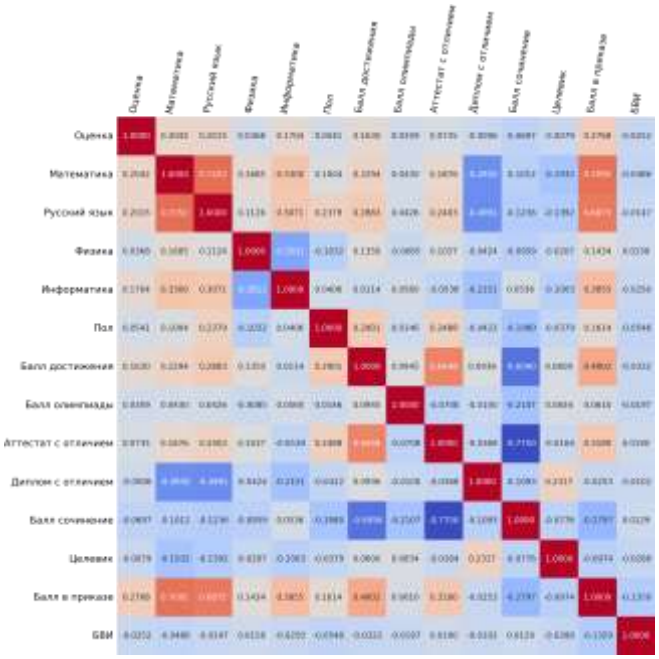

Рис. 2. Корреляционная матрица количественных признаков

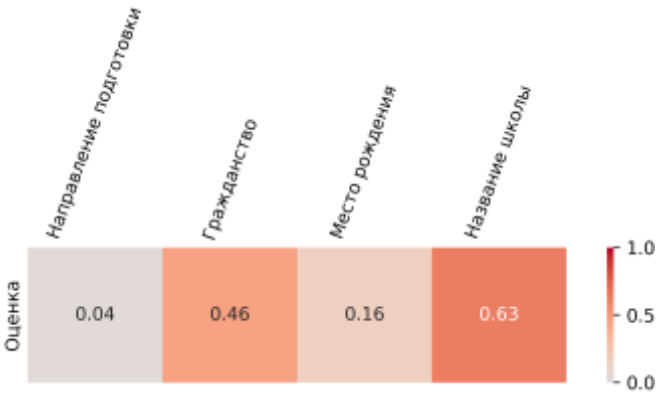

Рис. 3. Корреляция категориальных признаков с оценкой

используящая случайный лес. Общая точность данной модели составила 67.9 % на тестовых данных, неиспользованных во время обучения.

III. Начальные эксперименты

A. *Описание датасета*

Для проведения исследования был сформирован набор данных, состоящий из различной информации об абитуриентах СПбГЭТУ «ЛЭТИ» за 2016 – 2019 года, а также успехах их обучения. Формирование датасета потребовало приведения данных за разные года к одному виду. Информация об абитуриентах включала в себя признаки, характеризующие успехи обучения абитуриентов в средней школе, такие как баллы абитуриентов за Единый Государственный Экзамен (ЕГЭ) и уровень аттестата об окончании среднего образования, а также различные демографические признаки, например, пол, место рождения, местоположение школы, и множество других. В качестве валидационного датасета были использованы данные об абитуриентах за 2021 год, а также данные об их результатах обучения в первом семестре по дисциплинам «Математический анализ», «Физика», «Информатика» и «Программирование».

B. *Классификация*

Целью задачи классификации было поставлено определение того, сдаст ли абитуриент, будущий первокурсник, экзамен по той или иной дисциплине, или не сдаст. Соответственно задача является задачей бинарной классификации. Так в обучающем и валидационном датасетах оценки «3» и выше были помечены как «сдаст», а оценки ниже и недопуски к сдаче экзамена или зачета как «не сдаст». В качестве алгоритма было выбрано Дерево решений [3]. В качестве признаков использовались сумма баллов за ЕГЭ и данные о различных достижениях абитуриента, таких как олимпиады, уровень его аттестата и т. п.

Полученные метрики классификации представлены в табл. 1, матрица ошибок представлена на рис. 1.

ТАБЛИЦА I Метрики классификации в начальных экспериментах

Класс	Точность	Полнота	F1
«Не сдаст»	0.39	0.16	0.23
«Сдаст»	0.68	0.88	0.77

Как видно из таблицы и матрицы ошибок, модель плохо классифицировала экземпляры класса «Не сдаст».

```
Data: raw_city_string
Result: clean_city_string

// Очищаем строку
cleared_string ← clear_string(raw_city_string);

best_score ← 0;
best_word ← None;

for word ∈ cleared_string.split('  ') do
    for city ∈ clean_cities do
        score ← levenshtein_distance(word, city);
        if score > best_score ∧ score > 80 then
            best_score ← score;
            best_word ← city;
        end
    end
end

if best_word then
    return best_word
end

return cleared_string
```

Рис. 4. Алгоритм по нормализации признака «Место рождения»

IV. Улучшение модели

A. *Анализ данных*

В ходе анализа данных была построена корреляционная матрица количественных признаков для валидационного датасета. В качестве коэффициента корреляции использовался коэффициент корреляции Пирсона. Наиболее коррелирующие с оценкой количественные признаки представлены на рис. 2 (БВИ – без вступительных испытаний). Также была построена корреляционная матрица для категориальных признаков, которая показала связь оценки с такими признаками как «Направление подготовки», «Гражданство», «Место рождения» и «Название школы». Матрица представлена на рис. 3. Стоит учесть, что столь высокие значения корреляции могли быть вызваны малым количеством экземпляров для отдельно взятой категории.

B. *Дополнительная обработка данных*

Такие категориальные признаки, как «Место рождения» и «Название школы», представляют собой вручную введенные данные, содержащие множества вариаций ввода одного и того же значения признака, а также опечатки.

Для нормализации признака «Место рождения», представляющий из себя названия городов, был реализован специальный алгоритм, часть которого для одного экземпляра представлена на рис. 4. В ходе работы алгоритма данные, полученные в результате ручного ввода, очищаются от лишних слов, таких как различные сокращения, а также слова, не несущие какого-либо смысла в контексте решаемой задачи. Далее строки разделяются на слова, для каждого из которых в базе данных городов ищется город, ближайший по расстоянию Левенштейна. Результаты выполнения алгоритма для различных экземпляров кэшируются с целью оптимизации. Применение данного алгоритма помогло уменьшить количество уникальных экземпляров признака «Место рождения» в 3 раза.

Следующим этапом стала обработка экземпляров признака «Название школы». Среди данных экземпляров наблюдалось большое количество уникальных, что могло негативно сказаться на качестве модели. Для обобщения и нормализации данного признака был реализован алгоритм, сопоставляющий указанному названию школы вероятности того, что у школы тот, или иной тип учебного заведения: общеобразовательная школа, гимназия, лицей и т. п. Для реализации алгоритма из датасета были собраны уникальные типы учебных учреждений в различных формах их написания: краткие и полные. В ходе работы алгоритма в строке с названием школы сначала выполняется поиск сокращения типа учебного заведения, затем, если сокращение не было найдено, производится поиск полного типа. Результатом поиска полного типа является расстояние Левенштейна между указанным названием школы и типом учебного заведения. Описание работы алгоритма для одного экземпляра приведено на рис. 5.

```
Data: raw_school_string
Result: school_type_map

// Очищаем строку
cleared_string ← clear_string(raw_school_string);

result ← {};

for type ∈ full_matches do
    if type ∈ cleared_string then
        result[type] ← 1;
    end
    else
        result[type] ← 0;
    end
end

for (type, patterns) ∈ near_matches do
    for pattern ∈ patterns do
        match = get_near_match(cleared_string, pattern);
        if match then
            result[type] ← max(result[type], match);
        end
        else
            result[type] ← 0;
        end
    end
end

return result
```

Рис. 5. Алгоритм по определению типа учебного заведения

C. *Результаты*

Итоговый набор признаков, использовавшихся при обучении: Физика (ЕГЭ), Математика (ЕГЭ), Русский язык (ЕГЭ), Информатика (ЕГЭ), Балл за сочинение, Место рождения, Направление подготовки, Гражданство, Балл за достижения, Является сиротой, Является инвалидом, Есть аттестат с отличием, Пол, Является целевиком, Является иностранцем, Балл за олимпиады и конкурсы, Есть диплом с отличием, Тип школы.

Были проведены эксперименты с различными алгоритмами машинного обучения. Результаты экспериментов представлены в табл. 2.

ТАБЛИЦА II МЕТРИКИ КЛАССИФИКАЦИИ В ЭКСПИРИМЕНТАХ С ОБНОВЛЕННЫМИ ДАННЫМИ

Алгоритм	**Класс**	**Точность**	**Полнота**	**F1**
Дерево решений	«Не сдаст»	0.53	0.11	0.18
	«Сдаст»	0.68	0.95	0.80
Случайный лес	«Не сдаст»	0.0	0.0	0.0
	«Сдаст»	0.67	1.00	0.80
Градиентный бустинг	«Не сдаст»	0.54	0.58	0.56
	«Сдаст»	0.79	0.76	0.77

Как видно из таблицы модель со Случайным лесом [5] не смогла выявить ни одного случая, когда студент не сдал экзамен, а результаты модели с Градиентным бустингом [6] оказались лучше, чем модели с Деревом решений. Общая точность данной модели составила 70 %.

V. ЗАКЛЮЧЕНИЕ

В данной статье рассмотрена задача оценки абитуриента. Задача решена при помощи методов машинного обучения, а именно алгоритма Градиентного бустинга. Признаки для обучения модели были отобраны в ходе корреляционного анализа доступных данных.

Полученная модель была протестирована на данных последней сессии на начало 2022 года. Модель показала точность в 70 %.

Тем не менее, модель не смогла достаточно хорошо определить случаи, когда студенты не сдали экзамен по той или иной дисциплине. Для улучшения качество оценки абитуриентов необходимо провести дальнейшую работу по нормализации данных в датасете, поиску новых признаков, а также применению более сложных методов машинного обучения.

СПИСОК ЛИТЕРАТУРЫ

[1] Z. Iqbal, J. Qadir and A. N. Mian. Admission Criteria in Pakistani Universities: A Case Study // International Conference on Frontiers of Information Technology (FIT). 2016. С. 69-74.

[2] Dalton, Anthony & Beer, Justin & Kommanapalli, Sriharshasai & Lanich, James. Machine Learning to Predict College Course Success. 2018.

[3] Quinlan, J. R. Induction of Decision Trees // Machine Learning. — Kluwer Academic Publishers. 1986. Т. 1. С. 81-106.

[4] В. И. Левенштейн. Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий Наук СССР. 1965. С. 845-848.

[5] Breiman, Leo. Random Forests // Machine Learning 2001. Ч. 45, Т. 1. С. 5-32.

[6] Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. 1999.