ПРОГНОЗИРОВАНИЕ ВЕРОЯТНОСТИ ОТЧИСЛЕНИЯ СТУДЕНТОВ: ПРЕПРОЦЕССИНГ ДАННЫХ

Третьяков К.С.[1], Девин Г.А.[2]
[1]Национальный исследовательский Томский политехнический университет, ИШИТР, 8К12, e-mail: kst12@tpu.ru
[2]Национальный исследовательский Томский политехнический университет, ИШИТР, 8К12, e-mail: gad7@tpu.ru
Научный руководитель: Брагин А.Д., старший преподаватель

Аннотация

Работа посвящена проблеме высокого процента отчислений студентов. Основная цель исследования – разработка метода для прогнозирования вероятности отчисления студентов, что важно для улучшения успеваемости и сохранения студентов, а также для эффективного распределения ресурсов и создания поддерживающих программ. Методология работы включает использование датасетов, содержащих информацию о различных аспектах студенческой активности, и подготовку этих данных.

Ключевые слова: очистка данных от выбросов и ошибок, препроцессинг данных, обработка пропущенных значений, выделение и создание новых признаков

Введение

Проблема высокого процента отчислений студентов ставит перед учебными заведениями серьезные вызовы. Для решения этой проблемы, прогнозирование вероятности отчисления студентов становится важным инструментом для принятия мер по улучшению успеваемости и сохранению студентов, а также эффективного распределения ресурсов и создания программ поддержки.

Основная цель исследования является создание модели, способной предсказывать вероятность отчисления студентов на основе доступных данных об их активности и успеваемости в учебном процессе. После прогнозирования вероятности отчисления модель будет использоваться для выявления студентов с высоким риском отчисления, что позволит учебным заведениям проводить более эффективную работу с такими студентами.

Описание алгоритма

Для прогнозирования отчисления студентов были использованы наборы информации, содержащие следующие данные: студент, дата рождения, пол, группа, дисциплина, семестр, специальность, форма обучения, тип финансирования, пропуски занятий, часы, выделенные на дисциплину.

Для того чтобы данные стали пригодными для анализа и моделирования, мы провели ряд этапов препроцессинга данных. Эти этапы включают в себя очистку данных от выбросов и отсутствующих значений, нормализацию числовых признаков, кодирование категориальных признаков, и создание признаков, которые могут быть полезными для модели.

Анализ данных:

Ниже приведена гистограмма (рис. 1), отображающая распределение числа записей в зависимости от года поступления, для строк, в которых значение столбца оценка отсутствует. Также представлена столбчатая диаграмма (рис. 2), иллюстрирующая процент записей без оценки в зависимости от значения года набора.

При проведении первичного анализа было выявлено значительное количество нулевых значений в наборе данных об успеваемости студентов. В связи с этим было принято решение исключить данные, относящиеся к периоду до 2010 года.

Также, мы не учитывали студентов, которые ушли в академический отпуск по причинам, не связанным с их успеваемостью, такими как призыв в армию, уход за ребенком до 1,5 лет, беременность и роды, а также загранкомандировки. Не учитывали записи об оценках студентов, которые вышли из академа, восстановились, закончили обучение или стали магистрами.

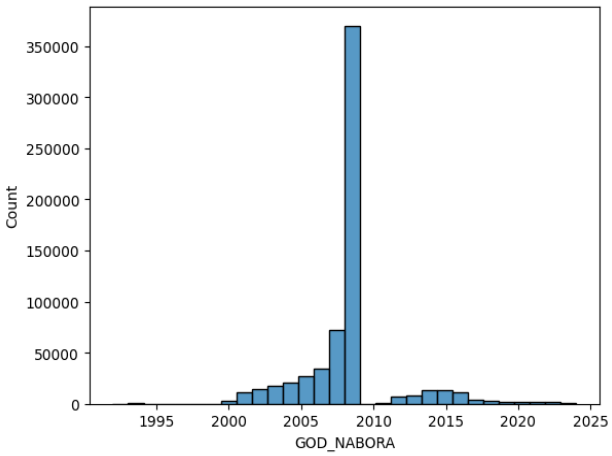

Рис. 1. Гистограмма распределения отсутствующих значений для столбца оценки

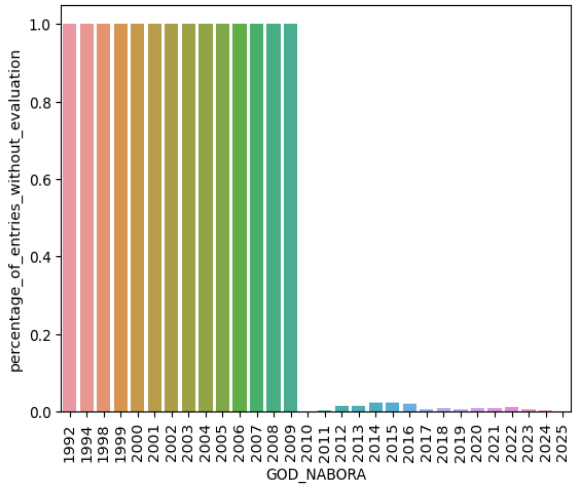

Рис. 2. Процент записей без оценки

Ниже (рис. 3) приведена столбчатая диаграмма, отображающая количество уникальных предметов для каждого года поступления. Эта диаграмма помогла выявить выбросы в данных, так как были обнаружены предметы, относящиеся к 25-му году поступления.

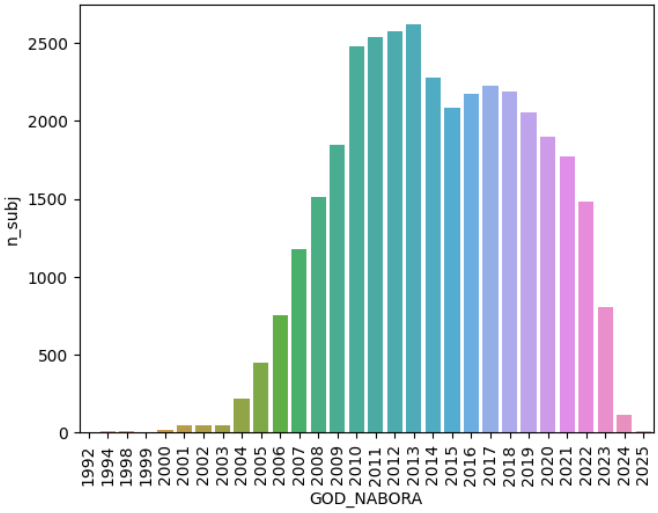

Рис. 3. Количество уникальных предметов

Мы решили использовать информацию только о студентах, которые обучаются на очной форме обучения и являются бакалаврами, и чья дата начала обучения позднее 1 января 2010 года.

Реализовали метод, который заполнил пропущенные оценки студентов в данных. Метод был настроен таким образом, чтобы учитывать наличие символов оценки или балла. Если оценка была указана или балл находился в допустимом диапазоне (от 0 до 100), то метод оставлял значение без изменений. В противном случае, для каждой пропущенной оценки было сгенерировано случайное значение из нормального распределения на основе среднего значения, стандартного отклонения, минимального и максимального значений оценок для соответствующего символа и года поступления. Таким образом, пропущенные оценки были успешно заполнены, а данные стали пригодными для дальнейшего анализа.

Для дальнейшей аналитики также были исключены студенты, которые были отчислены по следующим причинам: болезнь, нарушение правил проживания в общежитии, нарушение правил внутреннего распорядка, окончание обучения по сетевым программам с вузами РФ, завершение программы обучения в рамках академического обмена и призыв в ряды РА.

Был разработан метод для определения того, сдал ли студент дисциплину вовремя. Если студент пересдавал предмет в другом семестре (то есть был должником), запись о его неуспеваемости была стерта из-за особенности хранения данных. Для учета этого случая был разработан метод, который проверяет, сдавал ли студент дисциплину в том семестре, в котором она преподавалась. В случае несвоевременной сдачи в набор данных добавляется запись об оценке «Неудовлетворительно» для изначального семестра, а к оригинальной записи присваивается флаг, обозначающий, что оценка была исправлена в последующем семестре.

Также в процессе анализа данных были обнаружены студенты, у которых информация о годе поступления не совпадала с датами проставления оценок. Такие данные мы удалили, поскольку они являлись выбросами.

Для более полного анализа данных и получения дополнительной информации о каждой группе в наборе данных была разработана функция добавления контекста. Эта функция обогащает данные, добавляя для каждого студента контекст группы и вычисляя статистические характеристики (минимум, максимум, среднее, медиану) для указанного столбца в каждой группе.

Был разработан метод, который добавлял данные о пропусках по виду занятия и контекст пропусков (минимальное, максимальное, среднее и медианное значение для группы). Также было добавлено соотношение пропусков к часам, отведенным на занятия в аудитории, тип финансирования (флаг того, что студент перешел с договора на бюджет). Дополнительно были добавлены данные о часах, отведенных на дисциплину для группы студентов.

В результате финальный набор данных имел такие столбцы: уникальный id, балл за дисциплину, семестр закрытия дисциплины, флаг исправленного предмета, столбцы, в которых отражен контекст баллов группы, столбцы, в которых отражен контекст пропусков группы, и столбцы, в которых указано соотношение часов, выделенных на дисциплину и пропусков по виду занятия.

Заключение

В результате проведенного анализа и предобработки данных была подготовлена информация о студенческой активности и успеваемости для дальнейшего моделирования и аналитики. Проведенные этапы помогли улучшить качество данных, заполнить пропущенные значения оценок, исключить ненужные записи, а также добавить контекст для более полного анализа. Это позволит эффективно провести дальнейшее исследование и выявить закономерности, влияющие на успеваемость студентов и другие аспекты их активности.

Список использованных источников

1. Understanding LSTM Networks // colah`s blog URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (дата обращения: 27.03.2024).

2. Understanding Bidirectional RNN in PyTorch // towardsdatascience.com URL: https://towardsdatascience.com/understanding-bidirectional-rnn-in-pytorch-5bd25a5dd66 (дата обращения: 26.03.2024).

3. Айвазян С.А. Прикладная статистика: Классификация и снижение размерности: справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков. – М.: Финансы и статистика, 1989. – 607 с.

4. Воронцов К.В. Комбинаторные оценки качества обучения по прецедентам // Докл. РАН. – 2004. – Т. 394. – № 2.– С. 175–178.

5. Рекуррентная нейронная сеть (RNN): виды, обучение, примеры // neurohive.io URL: https://neurohive.io/ru/osnovy-data-science/rekurrentnye-nejronnye-seti/ (дата обращения: 26.03.2024).