Министерство науки и высшего образования Российской Федерации

ФГАОУ ВО «Волгоградский государственный университет»

Институт математики и информационных технологий

Кафедра математического анализа и теории функций

УТВЕРЖДАЮ:

Зав. Кафедрой МАТФ

_____ А.А. Клячин

«____»_____20___г.

ОТЧЁТ

о прохождении научно-исследовательской работе

2023-2024 учебный год

Студент Ворозова В.К. (ФИО) ПИб-221 (группа)

Руководитель практики от ВолГУ Гермашев И.В. (ФИО) д.т.н., проф., проф каф. МАТФ (должность, учено звание и степень)

Ответственный за организацию практики от кафедры Гермашев И.В. (ФИО) д.т.н., проф., проф каф. МАТФ (должность, учено звание и степень)

Место прохождения практики Волгоградский государственный университет (наименование учреждения, структурного подразделения)

Сроки прохождения практики с «09» февраля 2024г. по «11» июня 2024г.

1. Ход выполнения практики

№ П/П	Этап практики	Дата	Описание выполненной работы	Отметка руководителя о выполнении
1	Подготовительный этап	09.02.24	Встреча с научным руководителем и начало работы	
2	Ориентировочный этап	16.02.24	Составление плана работы. Подготовка источников информации	
3	Основной этап	01.03.24	Начало изучения подготовительных материалов	
		22.03.24	Начало работы с предоставленной базой данных	
		05.04.24	Изучение предоставленного репозитория	
		19.04.24	Нахождение выбросов разными методами	
		03.05.24	Обработка выбросов в данных	
4	Заключительный этап	07.06.24	Составление отчёта о прохождении научно-исследовательской работы	

Студент _____ _____
(подпись) (расшифровка подписи)

Введение

В эпоху больших данных и машинного обучения, качество данных становится все более важным. Однако, в реальном мире данные редко бывают идеальными. Они часто содержат выбросы - значения, которые сильно отличаются от остальных. Эти выбросы могут исказить результаты анализа данных и обучения моделей машинного обучения, поэтому их важно обрабатывать.

Целью данной работы является изучение и применение методов обработки выбросов в данных перед машинным обучением. Мне предстоит рассмотреть различные подходы к определению, обнаружению и обработке выбросов, а также оценить их эффективность на практических примерах.

1. Теоретические основы обработки выбросов

1.1. Определение выбросов и методы их обнаружения

Выбросы - это наблюдения, которые отклоняются от остальных наблюдений в наборе данных. Они могут быть вызваны различными причинами, такими как ошибки измерения, аномалии в данных или естественные отклонения. Выбросы могут быть унимодальными (одиночными) или мультимодальными (группами).

Существует множество методов для обнаружения выбросов. Некоторые из них включают статистические тесты, такие как Z-оценка и тест Граббса, а также методы машинного обучения, такие как кластеризация и изолирующий лес. Выбор метода зависит от природы данных и конкретной задачи. Рассмотрим несколько способов нахождения выбросов:

1. Визуализация данных - это простой и эффективный подход к обнаружению выбросов (рис.1). По мере увеличения размерности набора данных, его сложнее представить визуально. Следовательно, сложнее обнаружить выбросы в многомерном наборе данных с помощью визуализации.

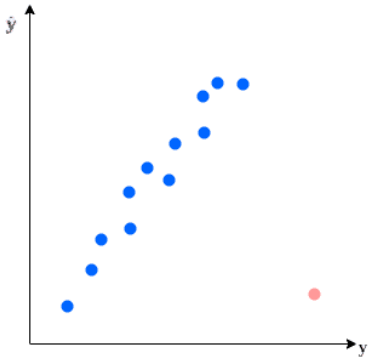

Рисунок 1 – диаграмма рассеяния

2. Квартильный анализ - это метод, который включает в себя разделение данных на четыре равных частях на основе их значений. Квартили - это значения, которые делят набор данных на четыре равных частях. Первый квартиль (Q1) представляет 25-й процентиль, второй квартиль (Q2) представляет 50-й процентиль (также известный как медиана), а третий квартиль (Q3) представляет 75-й процентиль. Четвертый квартиль (Q4) представляет максимальное значение в наборе данных. В методе квартильного анализа мы определяем значения, выходящие за пределы $k \times IQR$ диапазона, как выбросы. Мы можем визуализировать квантильный анализ с помощью прямоугольников. Концы прямоугольника обозначают квартили. Кроме того, на графике отмечена медиана. Усики указывают на $k \times IQR$ диапазон данных (рис. 2).

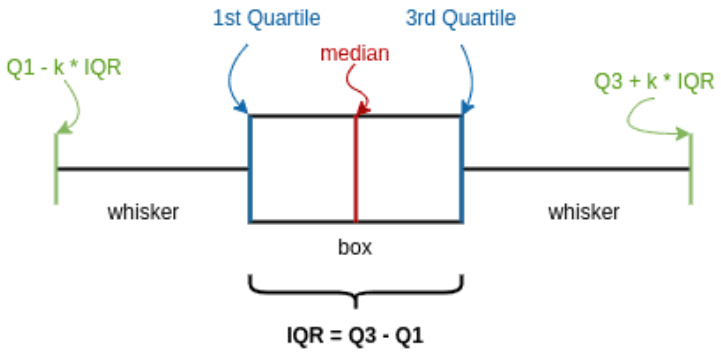

Рисунок 2 – Boxplot (ящик с усами)

3. Z-оценка, также известная как стандартное значение или z-score, - это статистический показатель, который измеряет относительное отклонение наблюдаемого или измеренного значения от среднего значения. Это безразмерный показатель, который используется для сравнения значений разной размерности или шкалы измерений. Если Z-оценка равна 0, это означает, что оценка точки данных идентична средней оценке. Если Z-оценка меньше нуля, то значение ниже среднего, если Z-оценка больше нуля, то значение выше среднего. Z-оценка рассчитывается по следующей формуле: $Z = \frac{x-\mu}{\sigma}$ где: x - это значение точки данных, μ - это среднее значение набора данных, σ - это стандартное отклонение набора данных.

4. Тест Граббса основан на предположении о нормальном распределении. Критерий Граббса определяет один выброс за одну итерацию. Этот выброс исключается из набора данных и тест повторяется до тех пор, пока не будут обнаружены все выбросы. Тест Граббса определён для гипотез:

- H0: В наборе данных нет выбросов
- H1: В наборе данных присутствует как минимум один выброс

Критерий Граббса рассчитывается как: $G = \frac{|X - \bar{X}|}{S}$ где x - это максимальное или минимальное значение в наборе данных, $\bar{x}$ - это выборочное среднее, s - это выборочное стандартное отклонение. Значение критерия Граббса показывает максимальное абсолютное отклонение от выборочного среднего в единицах среднеквадратичного отклонения. Если тестовая статистика больше критического значения, это означает, что значение является выбросом в этом наборе данных.

5. Изолирующий лес: создает случайные разделы на основе объекта. Древовидная структура визуализирует, как мы формируем разделы. Итак, количество ребер от корня до выборки представляет собой количество разбиений, необходимых для выделения этого конкретного наблюдения. Средняя длина пути таких деревьев служит функцией принятия решения. Всем наблюдениям мы присваиваем оценку аномалии. В результате, когда мы формируем лес таких случайных деревьев, они в совокупности дают длину пути для данного наблюдения. Более короткие средние значения, скорее всего, будут выбросами (рис.3).

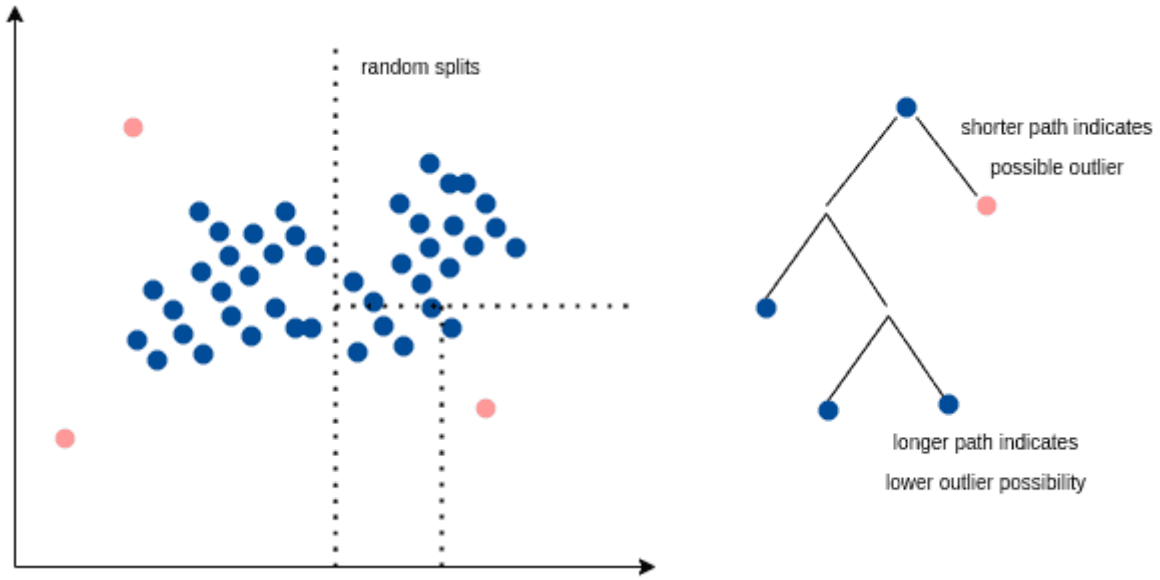

Рисунок 3 – изолирующий лес

1.2. Способы обработки выбросов

Удаление выбросов: Это самый простой и прямолинейный метод обработки выбросов. Если выбросы являются результатом ошибок в данных или не отражают общую тенденцию данных, их можно просто удалить. Однако этот метод имеет свои недостатки. Удаление выбросов может привести к потере важной информации, особенно если выбросы являются значимыми аномалиями, которые необходимо учесть.

Замена выбросов: Вместо удаления выбросов их можно заменить на другие значения. Например, выбросы могут быть заменены на среднее или медианное значение. Этот метод может помочь снизить влияние выбросов на анализ данных, но он также может исказить распределение данных и ввести искажение в анализ.

Использование робастных методов: Робастные методы - это методы, которые устойчивы к выбросам. Например, вместо использования среднего значения можно использовать медиану, которая менее чувствительна к выбросам. Робастные методы могут быть особенно полезны при работе с данными, которые содержат много выбросов.

Преобразование данных: В некоторых случаях выбросы могут быть обработаны путем преобразования данных. Например, логарифмическое преобразование может помочь уменьшить влияние выбросов.

Биннинг данных: Биннинг данных - это процесс преобразования непрерывных числовых переменных в категориальные ‘бины’ или ‘корзины’. Это может быть полезно для управления выбросами, поскольку они могут быть включены в верхний или нижний ‘бин’.

2. Практическая часть

2.1. Описание используемых данных

Данные в предоставленной мне таблице представляют собой результаты измерений внешних и внутренних температур легких, полученных с помощью метода микроволновой радиотермометрии. Измерения были проведены в 14 различных точках, чтобы обеспечить максимально точное и детализированное представление о температурных паттернах.

В каждой строке таблицы представлены данные об одном пациенте, всего 192 человека. А каждый столбец – это точка измерения. Также есть метка cls означающая болен пациент или нет. Стоит обратить внимание, что все температурные значения представлены в градусах Цельсия.

2.2. Применение методов обнаружения выбросов и их обработки

При первичном анализе базы данных было найдено пропущенное значение среди внешних температур левого лёгкого одного из пациентов, которое было заменено серединным значением.

Были построены графики boxplot. Где по оси абсцисс метка cls (0 – здоров, 1 – covid-19 или пневмония), а по оси ординат медиана внутренних (рис. 4) и внешних (рис. 5) температур. Оба boxplot показывают медиану (середину) данных, первый и третий квартили (25-й и 75-й процентили), а также выбросы. Выбросы представлены отдельными точками вне усов. Категория ‘0’ не имеет выбросов, в то время как категория ‘1’ имеет несколько выбросов.

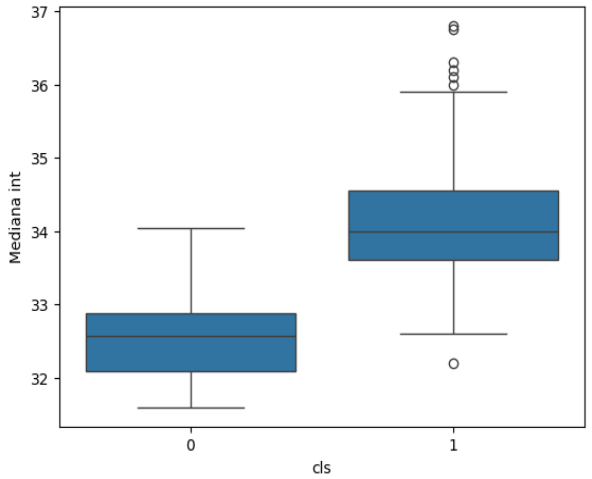

Рисунок 4 – boxplot по Mediana int и cls

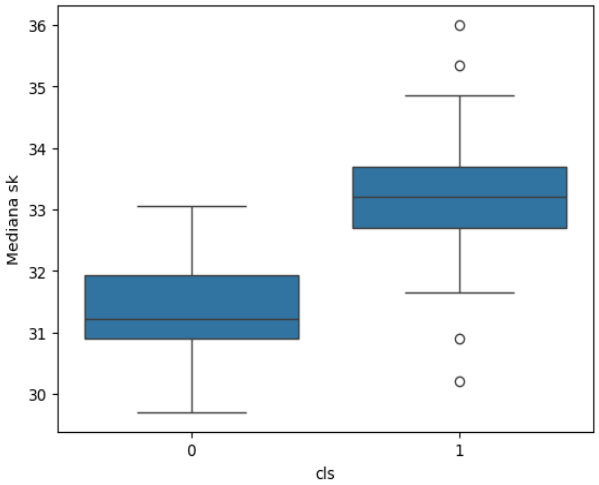

Рисунок 5 – boxplot по Mediana sk и cls

Был предложен подход к обнаружению и обработке выбросов, который сочетает в себе статистические методы и методы машинного обучения.

Метод обнаружения выбросов на основе пороговых значений: выбросы определяются по следующим критериям:

1. Если абсолютное значение разности между текущим элементом и его соседними элементами превышает 0.03 (при переводе в диапазон от 0 до 1), то текущий элемент считается выбросом.
2. Если абсолютное значение разности между текущим элементом и средним значением строки превышает 0.02, то текущий элемент также считается выбросом.

После обнаружения выбросов используется обученная на этих данных модель машинного обучени**я** для предсказания корректного значения для каждого выброса. Для этого удаляется выброс из строки данных, полученная строка передается в модель для предсказания, где выброс заменяется предсказанным значением.

В ходе проведённой работы было найдено 10 точек среди внешних температур и 51 точка среди внутренних, которые выходили за пороговые значения и были заменены.

Однако, избавится от выбросов по столбцам медианных температур каждого пациента не удалось (рис.6 и рис.7). Изучив каждый выброс отдельно, удалось установить, что у данных пациентов все измерения либо сильно ниже чем у других, либо сильно выше. А так как при поиске выброса сравниваются температуры одного пациента и в рамках этого пациента пороговые значения не были превышены, то значения тех, что обозначены как выбросы, не изменились.

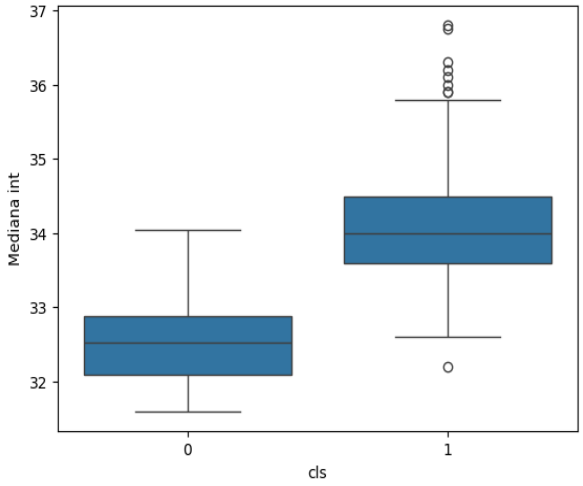

Рисунок 6 – boxplot по Mediana int и cls

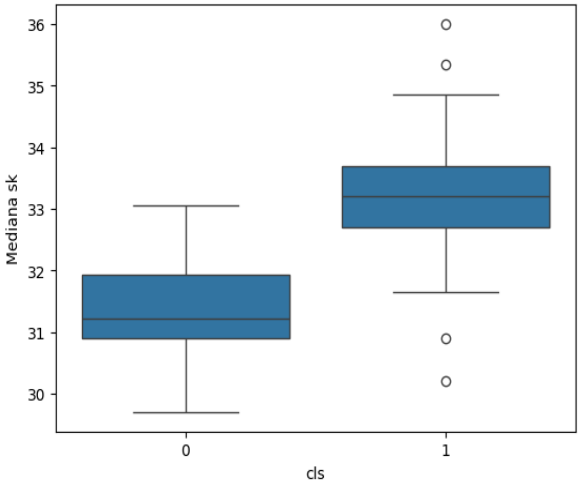

Рисунок 7 – boxplot по Mediana sk и cls

Заключение

В ходе исследования была подробно изучена проблема обработки выбросов в данных перед машинным обучением. Было обнаружено, что эта проблема имеет значительное влияние на качество и точность прогнозирования моделей машинного обучения.

Был проверен подход к обнаружению и обработке выбросов, который сочетает в себе статистические методы и методы машинного обучения. Однако результаты не оправдали ожиданий. Это подчеркивает сложность задачи обработки выбросов и важность правильного выбора метода.

Список использованной температуры

1. Книга: “Python Machine Learning” by Sebastian Raschka.
2. Книга: «Введение в машинное обучение с помощью Python» А.Мюллер, С.Гвидо.
3. Статья: “Outlier Detection and Handling” на сайте baeldung.com (https://www.baeldung.com/cs/ml-outlier-detection-handling).
4. Репозиторий на GitHub: https://github.com/Nekekys/neuralNetworkPY