МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра дискретной математики и информационных технологий

ИССЛЕДОВАНИЕ АНОМАЛЬНЫХ ЗНАЧЕНИЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Королева Ивана Дмитриевича

Научный руководитель
профессор, д. э. н. _____ Л.В. Кальянов

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2021

ВВЕДЕНИЕ

Процесс обнаружения аномалий относится к проблеме поиска закономерностей (паттернов) в данных, которые не соответствуют ожидаемому поведению. Эти несоответствующие шаблоны часто называют аномалиями, выбросами, несогласованными наблюдениями, исключениями, особенностями или загрязнителями в различных областях применения. Из них аномалии и выбросы - это два термина, которые наиболее часто используются в контексте обнаружения аномалий; иногда взаимозаменяемо. Обнаружение аномалий находит широкое применение в самых разных приложениях, таких как обнаружение мошенничества с кредитными картами, страхование или здравоохранение, обнаружение вторжений для кибербезопасности, обнаружение сбоев в критически важных системах безопасности и военное наблюдение за действиями противника.

Важность обнаружения аномалий обусловлена тем фактом, что аномалии в данных преобразуются в важную (и часто критическую) полезную информацию в широком спектре областей применения. Например, аномальная схема движения в компьютерной сети может означать, что взломанный компьютер отправляет конфиденциальные данные в неавторизованное место назначения. Аномальное изображение МРТ может указывать на наличие злокачественных опухолей. Аномалии в данных транзакций по кредитной карте могут указывать на кражу кредитной карты или личных данных, или аномальные показания датчика космического корабля могут указывать на неисправность какого-либо компонента космического корабля.

Целью бакалаврской работы является исследование аномальных значений в данных, применяя различные алгоритмы и определяя их эффективность.

Для достижения цели работы необходимо выполнить следующие задачи:

- изучить функционал и использование модулей Scrapy, Pandas, NumPy;
- создать набор данных;
- изучить типы аномалий значений;
- изучить режимы обнаружения аномальных значений;
- изучить алгоритмы для определения аномальных значений;
- в программе RapidMiner построить процессы для выбранных алгоритмов;
- рассмотреть различные подходы для бинарной классификации данных;

— провести сравнительный анализ эффективности процессы с помощью ROC-анализа;
— определить наилучший процесс для обнаружения аномальных данных.

В данной работе исследуемым объектом является статистика продаж автомобилей в странах европейского и СНГ регионов. Данные собраны с сайта carsalesbase.com [1]. Аномальные значения в наборе данных являются точечными, а для их обнаружения применен подход «обнаружения аномалий без обучения».

Бакалаврская работа состоит из введения, трех разделов, заключения, списка использованных источников и трех приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

1. Понятие аномальных значений и методы их обнаружения. На сегодняшний день не существует общепринятой классификации аномальных явлений в наборах данных [2]. Наиболее часто в работах отечественных и зарубежных ученых встречается классификация, в рамках которой выделены три типа аномалий:

- Точечные аномалии;
- Контекстные аномалии;
- Коллективные аномалии.

Для обнаружения аномалий используются три режима, в зависимости от взаимодействия с метками данных. Они используются для явного указания на то, является ли конкретный экземпляр данных нормальным или аномальным. Разметка зачастую производится специалистами вручную, из-за этого процесс разметки данных требует больших ресурсных затрат [3].

В зависимости от доступных меток, методы обнаружения аномалий могут быть использованы в одном из трех режимов:

- обнаружение аномалий с полным обучением;
- обнаружение аномалий с частичным обучением;
- обнаружение аномалий без обучения.

В данной работе было использовано три следующих алгоритма: k-NN Global Anomaly Score [4], Local Outlier Factor (LOF) [5], Histogram-based Outlier Score (HBOS) [6]. Первые два алгоритма относятся к методам, основанным на плотности, последний алгоритм относится к статистическим методам.

2. Автоматизация процессов сбора данных и их предварительной обработки. Перед построением процессов с использованием выбранных алгоритмов в программе RapidMiner, с помощью инструментов Python был создан и предварительно обработан набор данных. В ходе разработки были :

- с помощью модулей Python разработать краулер для извлечения данных с сайта, содержащего необходимые данные, с целью сформировать набор данных, который будет использован в дальнейшем;
- рассмотреть пример работы процесса, определяющий аномальные значения в программе RapidMiner;
- построить процессы с использованием выбранных алгоритмов;
- произвести анализ эффективности процессов с помощью ROC-анализа.

Для создания краулера использовался фреймворк `Scrapy` [7], [8] для Python. Установка производилась с помощью `Anaconda` [9]. Внутри краулера определены две функции:

— извлечение ссылок, введущих на каждую из представленных на сайте стран с помощью заданного регулярного выражения [10], и рекурсивный переход по полученным ссылкам;

— извлечение необходимых данных из таблиц, содержащихся на странице каждой страны и запись их в отдельный файл.

Предварительная обработка данных, включающая в себя подсчет изменений уровня продаж в процентном соотношении, нормализацию данных и объединение данных о продажах для каждой страны в единый .csv файл производилось с помощью модулей `Pandas` [11], `os` [12], `glob` [14].

Подсчет изменений уровня продаж для каждой страны с помощью Python. Такой подход обусловлен тем, что сразу в общем файле подсчет напрямую (например в Excel) производится некорректно: значения берутся из предыдущей ячейки, так, в запись не в начале набора данных, являющуюся начальной для очередной страны, попадает значение, вычисленное по последней записи предыдущей страны, что приводит к искажению набора данных.

Вручную убирать значения из подобных ячеек некорректно из-за нерациональности, так как при большем объеме данных этот процесс займет намного больше времени.

Целью нормализации является преобразование объектов выборки к единому масштабу. Данная операция повышает производительность и устойчивость модели к обучению [13]. В программе Rapidminer [15], в которой будет построен процесс для анализа, есть оператор нормализации данных. Однако, если применить его на общий набор данных, то данные не будут равномерно распределены. В связи с этим, нормализация производилась отдельно для данных в рамках каждой страны с помощью Python.

На примере набора данных для России в таблице 1 показаны ненормализованные и нормализованные данные (выделено normalized).

Таблица 1 – выполненная нормализация данных

	year	sales	sales (normalized)	change	change (normalized)
1	2005	1806625.0	0.251144	NaN	NaN
2	2006	1886824.0	0.304358	0.044392	0.619376
3	2007	2582682.0	0.766075	0.368799	0.991285
4	2008	2907857.0	0.981835	0.125906	0.712827
5	2009	1465922.0	0.025080	-0.495875	0.000000
6	2010	1914323.0	0.322604	0.305883	0.919157
7	2011	2634875.0	0.800706	0.376400	1.000000
8	2012	2935233.0	1.000000	0.113993	0.699170
9	2013	2777547.0	0.895372	-0.053722	0.506897
10	2014	2491394.0	0.705503	-0.103024	0.450376
11	2015	1603253.0	0.116203	-0.356484	0.159803
12	2016	1428123.0	0.000000	-0.109234	0.443256
13	2017	1599718.0	0.113857	0.120154	0.706233
14	2018	1800351.0	0.246981	0.125418	0.712267
15	2019	1754297.0	0.216423	-0.025581	0.539158
16	2020	1598369.0	0.112962	-0.088883	0.466586

3. Разработка информационной технологии анализа аномальных значений средствами RapidMiner. В ходе построения процессов с использованием выбранных алгоритмов, в программе RapidMiner выполнялись операции:

— генерации идентификаторов для повышения качества анализа полученных результатов;

— отсеивания пустых экземпляров данных;

— записи обработанного алгоритмами набора данных в новые файлы формата .csv для последующей классификации и анализа эффективности каждого из алгоритмов.

В качестве демонстрации работы в среде RapidMiner, на рисунке 1 показан построенный процесс с примененным алгоритмом k-NN Global Anomaly Score. Дальнейшая работа сопряжена с изучением данных, полученных в результате выполнения процесса. Все алгоритмы, использованные в работе, были установлены с расширением `Anomaly Detection` [16]

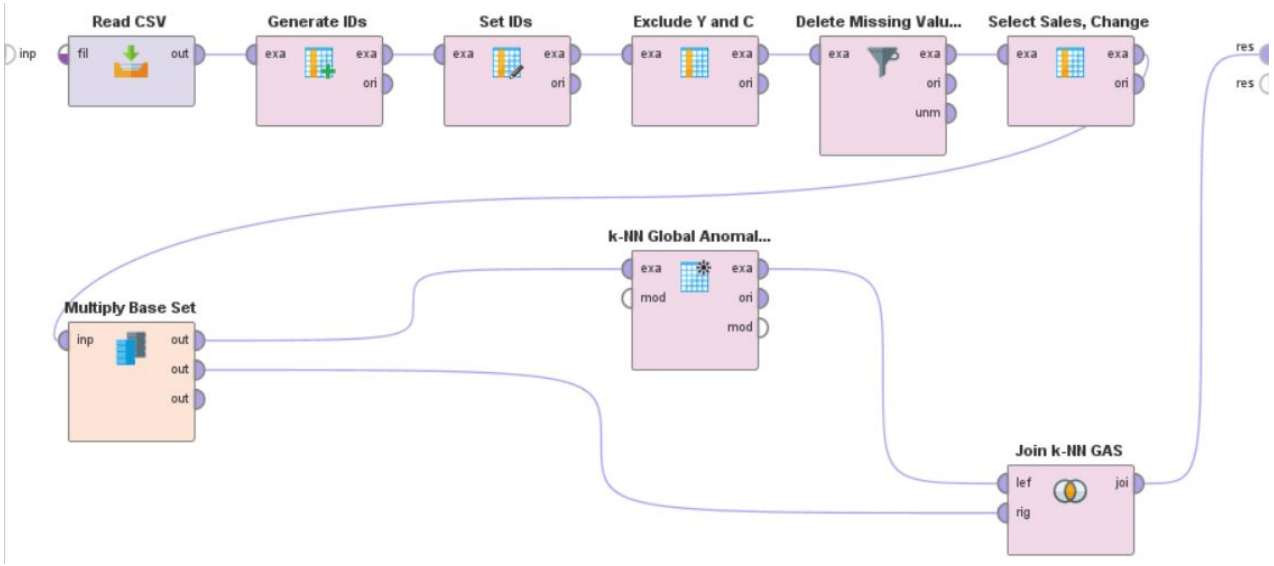

Рисунок 1 – общий вид построенного процесса

Вкладка данных, полученных в результате работы алгоритма k-NN Global Anomaly Score показана на рисунке 2.

Result History | ExampleSet (Join LOF)

Open in | Turbo Prep | Auto Model

Data | Statistics | Visualizations | Annotations

Row No.	id	outlier	sales	change
817	sweden_2020	0.184	0.657	0.025
630	norway_2018	0.176	0.897	0.141
711	serbia_2008	0.175	0.687	0.068
211	denmark_2020	0.164	0.811	0.137
457	kazakhstan_...	0.161	0.984	0.208
527	luxembourg_...	0.154	0.636	0.072
15	albania_2020	0.149	0.846	0.167
620	norway_2008	0.145	0.540	0
649	poland_2020	0.144	0.602	0.063
153	cyprus_2009	0.139	0.524	0.003
92	bosnia_2007	0.133	1	1
272	georgia_2006	0.133	1	1
560	moldova_2008	0.133	1	1
710	serbia_2007	0.133	1	1
631	norway_2019	0.132	0.844	0.196
610	norway_1998	0.129	0.610	0.124
680	romania_2007	0.129	1	0.988
208	denmark_2017	0.128	0.974	0.259

Рисунок 2 – данные, полученные при работе алгоритма k-NN Global Anomaly Score

Вкладка статистики, полученной в результате работы алгоритма k-NN Global Anomaly Score показана на рисунке 3.

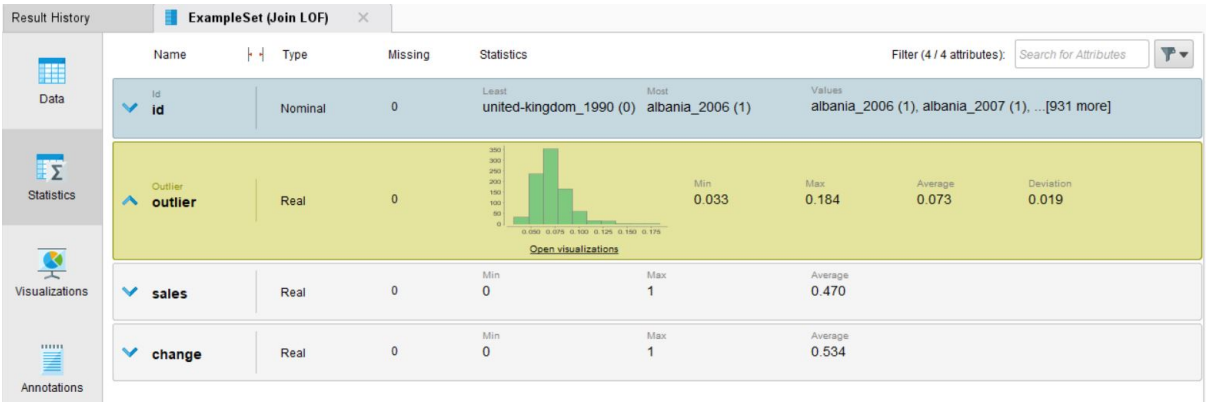

Рисунок 3 – статистика для процесса

Визуализация представлена на рисунке 4.

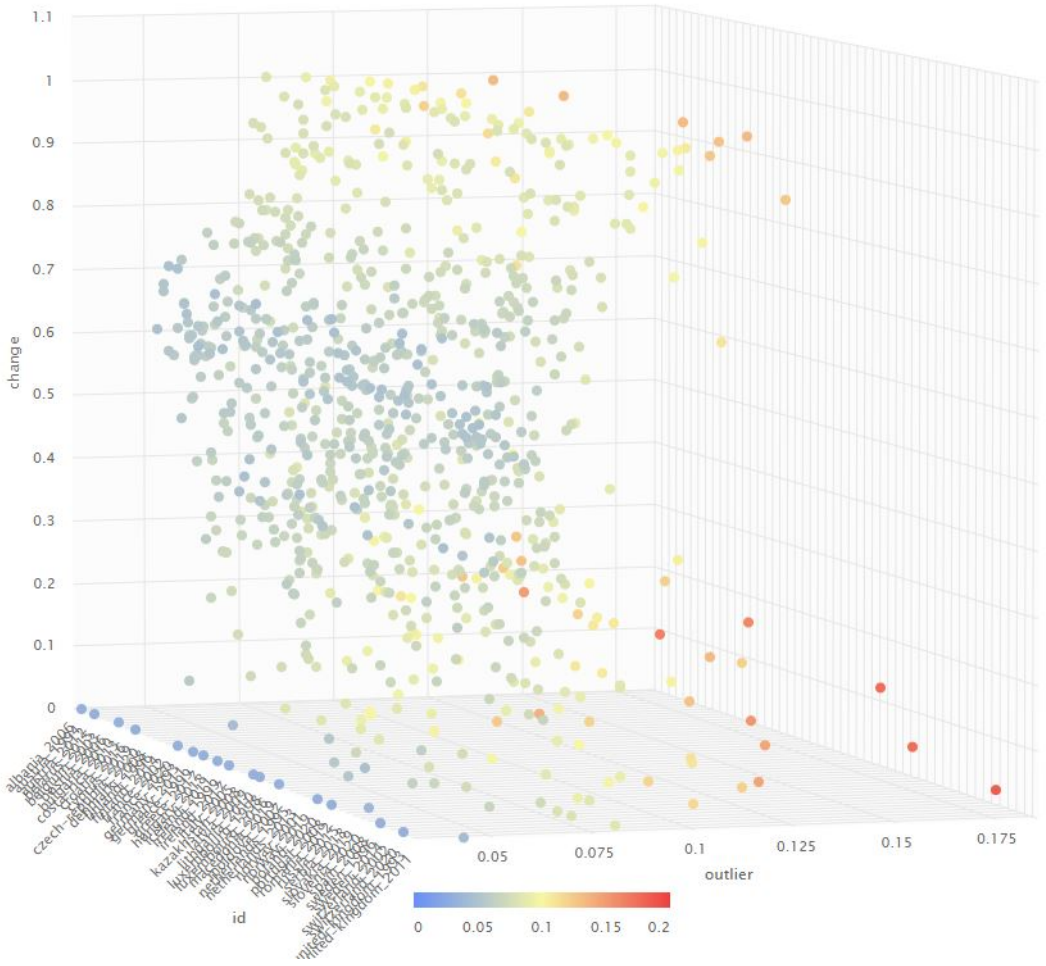

Рисунок 4 – визуализация данных

Для применения остальных алгоритмов требуется только замена оператора, отвечающего за алгоритм. Настройки визуализации также идентичны.

В таблице 2 показано сравнение результатов работы алгоритмов. Полужирным шрифтом выделены значения, совпадающие между алгоритмами.

Таблица 2 – сопоставление аномальных данных, выявленных алгоритмами алгоритмами

k-NN	LOF	HBOS
Швеция, 2020	Великобритания, 2016	**Люксембург, 2020**
Норвегия, 2018	Португалия, 1992	Бельгия, 2006
Сербия, 2008	Хорватия, 2008	Испания, 2008
Дания, 2020	Германия, 1991	Швеция, 2001
Казахстан, 2014	**Норвегия, 2018**	Австрия, 1992
Люксембург, 2020	Ирландия, 2000	Хорватия, 2020
Албания, 2020	**Сербия, 2008**	**Сербия, 2008**
Норвегия, 2008	**Швеция, 2020**	Финляндия, 2011
Польша, 2020	Литва, 2019	Исландия, 2003
Кипр, 2009	Венгрия, 2005	Великобритания, 2012

Бинарная классификация (разметка данных) необходима для того, чтобы определить эффективность той или иной модели с помощью ROC-анализа. Разметка данных осуществлялась с помощью визуальной оценки полученных результатов, а также с помощью межквартильного размаха [17], реализованного с помощью модуля `numpy` [18] для Python.

ROC-анализ представляет собой графический метод оценки качества работы бинарного классификатора. Два класса содержат показания с положительными и отрицательными исходами. В основе метода лежит построение ROC-кривой (ROC – receiver operating characteristic — рабочая характеристика приёмника) [19]. Ключевым параметром для оценки в ROC-анализе является значение площади под ROC-кривой [20].

Оценка эффективности, произведенная только на созданном с помощью краулера наборе данных, показала, что наилучшими являются процессы с использованием алгоритмов k-NN Global Anomaly Score и Local Outlier Factor (LOF) и визуальной оценки для решения задачи бинарной классификации, так как подход, примененный для процесса с алгоритмом Histogram-Based Outlier Score требовал дополнительных операция вне программы RapidMiner.

ЗАКЛЮЧЕНИЕ

В рамках бакалаврской работы была поставлена цель построения процессов для определения аномальных данных с помощью различных алгоритмов. Для достижения цели данной работы был проведен анализ предметной области, и выполнены следующие задачи:

- получены навыки по использованию функционала модулей языка Python: Scrapy, Pandas, NumPy и др.;
- создан набор данных;
- изучены типы аномалий значений;
- изучены режимы обнаружения аномальных значений;
- изучены алгоритмы для определения аномальных значений;
- в программе RapidMiner построены процессы для выбранных алгоритмов;
- рассмотрены различные подходы для бинарной классификации данных;
- проведен сравнительный анализ эффективности процессов с помощью ROC-анализа;
- на доступных данных был определен наилучший из исследуемых процессов для обнаружения аномальных данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Total Market Car Sales by Country [Электронный ресурс]. URL: //https://carsalesbase.com/total-market-sales-country/europe-car-sales-data/ (дата обращения 03.05.2021) Загл. с экрана Яз. англ.

2 RapidMiner. Data Mining Use Cases and Business Analytics Applications – University of Minnesota, 2014, Яз. англ.

3 Anomaly Detection: A Survey, Varun Chandola, Arindam Banerjee, Vipin Kumar – ACM Computing Surveys, 2007, Загл. с экрана Яз. англ

4 Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner, Mennatallah Amer and Markus Goldstein – University of Minnesota, German University in Cairo, Egypt, Яз. англ.

5 LOF: Identifying Density-Based Local Outliers – Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jorg Sander, – Institute for Computer Science, University of Munich, 2000, Яз. англ.

6 Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm – Markus Goldstein, Andreas Dengel, – German Research Center for Artificial Intelligence (DFKI), 2012, Яз. англ.

7 Scrapy | A Fast and Powerful Scraping and Web Crawling Framework [Электронный ресурс]. URL: //https://scrapy.org/ (дата обращения 03.05.2021) Загл. с экрана Яз. англ

8 parser — Access Python parse trees [Электронный ресурс]. URL: //https://docs.python.org/3/library/parser.html (дата обращения 03.05.2021) Загл. с экрана Яз. англ

9 Anaconda Framework [Электронный ресурс]. URL: //https://www.anaconda.com/ (дата обращения 03.05.2021) Загл. с экрана Яз. англ

10 Regular Expression HOWTO [Электронный ресурс] URL: //https://docs.python.org/3/howto/regex.html (дата обращения 04.05.2021) Загл. с экрана Яз. рус

11 pandas - Python Data Analysis Library [Электронный ресурс]. URL: //https://pandas.pydata.org/ (дата обращения 06.05.2021) Загл. с экрана Яз. англ

12 os — Miscellaneous operating system interfaces [Электронный ресурс]. URL: //https://docs.python.org/3/library/os.html (дата обращения 06.05.2021) Загл. с экрана Яз. рус

13 Normalization [Электронный ресурс]. URL: //https://developers.google.com/machine-learning/data-prep/transform/normalization (дата обращения 04.05.2021) Загл. с экрана Яз. англ

14 Glob() function [Электронный ресурс]. URL: //https://www.geeksforgeeks.org/how-to-use-glob-function-to-find-files-recursively-in-python/ (дата обращения 01.05.2021) Загл. с экрана Яз. англ

15 RapidMiner [Электронный ресурс]. URL: //https://rapidminer.com/ (дата обращения 01.05.2021) Загл. с экрана Яз. англ

16 Anomaly Detection [Электронный ресурс]. URL: //https://marketplace.rapidminer.com/UpdateServer/faces/product-details.xhtml?productId=rmx-anomalydetection (дата обращения 06.05.2021) Загл. с экрана Яз. англ

17 Outliers: Finding Them in Data, Formula, Examples [Электронный ресурс]. URL: //https://www.statisticshowto.com/statistics-basics/find-outliers/ (дата обращения 06.05.2021) Загл. с экрана Яз. англ

18 numpy.percentile [Электронный ресурс]. URL: //https://numpy.org/doc/stable/reference/generated/numpy.percentile.html/ (дата обращения 06.05.2021) Загл. с экрана Яз. англ.

19 Unsupervised anomaly detection and access control on network traffic – Dimakogiannis, M.M, – Eindhoven University of Technology, 2017, Яз. англ.

20 Classification: ROC Curve and AUC [Электронный ресурс]. URL: //https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc (дата обращения 06.05.2021) Загл. с экрана Яз. англ.