УДК 519.254
DOI: http://dx.doi.org/10.21686/2500-3925-2022-2-4-13

Т.В. Золотова[1], Д.А. Волкова[2]
[1]Финансовый университет при Правительстве РФ, Москва, Россия
[2]Российский университет дружбы народов, Москва, Россия

Методы интеллектуальной обработки данных для коррекции атипичных значений котировок акций

Цель исследования. *Цель исследования состоит в проведении сравнительного анализа различных методов коррекции атипичных значений статистических данных на фондовом рынке и выработке рекомендаций для их использования.*
Материалы и методы. *В статье проведен анализ российской и зарубежной библиографии по проблеме исследования. Предлагается рассмотрение методов машинного обучения для обнаружения и коррекции выбросов во временных рядах. Математическую основу методов машинного обучения составляют метод Z-score, метод изолирующего леса, метод опорных векторов для обнаружения выбросов и методы винсоризации и множественного вменения для коррекции выбросов. Для построения моделей использован программный инструмент Jupyter Notebook, поддерживающий язык программирования Python. Для реализации методов машинного обучения используются данные котировок акций Московской биржи.*
Результаты. *Продемонстрированы результаты работы алгоритмов машинного обучения для наборов реальных статистических данных, представляющих собой цены закрытия акций трех российских компаний «Сбербанк», «Аэрофлот», «Газпром» в период с 01.12.2019 по 30.11.2020, полученные с сайта с сайта Инвестиционной компании «ФИНАМ». Проведен сравнительный анализ методов обнаружения и коррекции выбросов по среднеквадратическому отклонению. Статистический метод Z-score позволяет точно определить расстояние от подозрительного наблюдения до центра распределения, что является преимуществом. Недостаток данного метода – это влияние выбросов на среднее значение и стандартное отклонение, которое может способствовать маскировке выбросов или некорректному их обнаружению. Метод изолирующего леса распознает выбросы различных видов, а при реализации метода отсутствуют параметры, требующие подбора; но недостатком является более низкая скорость обнаружения выбросов по сравнению с другими методами. Метод опорных векторов считается очень быстрым методом и сводится к решению задачи квадратичного программирования, которая всегда имеет единственное решение. Метод винсоризации для коррекции выбросов снижает влияние выбросов на среднее значение и дисперсию, что является преимуществом, но может вносить систематическую ошибку из-за подбора пороговых значений для разделения наблюдений в выборке. Метод множественного вменения создает для каждого пропущенного значения не одно, а множество вменений, что позволяет избежать систематической ошибки, но за счет высоких вычислительных затрат. Для использованных в работе исходных данных лучший результат показала реализация алгоритма множественного вменения по обнаруженным выбросам методом опорных векторов.*
Заключение. *В теории анализа данных нет универсального метода обнаружения и/или устранения выбросов. В общем случае определение выбросов субъективно, и решение принимается индивидуально для каждого конкретного набора данных с учетом его особенностей или имеющегося опыта в данной области. Реализованные в работе методы обнаружения и устранения выбросов могут найти применение при вычислении более точных значений показателей в различных сферах деятельности, например, для построения более точного прогноза цены акции. В перспективе планируется исследование влияния параметров методов обнаружения и коррекции выбросов на результаты работы моделей, а также оптимизация этих параметров.*

Ключевые слова: *обнаружение выбросов, коррекция выбросов, язык программирования Python, среднеквадратическое отклонение.*

Tatiana V. Zolotova[1], Daria A. Volkova[2]
[1]Financial University under the Government of the Russian Federation, Moscow, Russia
[2]Peoples' Friendship University of Russia, Moscow, Russia

Intelligent Data Processing Methods for the Atypical Values Correction of Stock Quotes

Purpose of the study. *The purpose of the study is to carry out a comparative analysis of various methods for correcting atypical values of statistical data on the stock market and to develop recommendations for their use.*
Materials and methods. *The article analyzes Russian and foreign bibliography on the research problem. Consideration of machine learning methods for detecting and correcting outliers in time series is proposed. The mathematical basis of machine learning methods is the Z-score method, the isolation forest method, support vector method for outlier detection, and winsorization and multiple imputation methods for outlier correction. To create the models, the Jupyter Notebook software tool, which supports the Python programming language, was used. To implement machine-learning methods, data from stock quotes of the Moscow Exchange are used.*
Results. *The results of machine learning algorithms are demonstrated for sets of real statistical data representing the closing prices of shares of three Russian companies "Sberbank", "Aeroflot", "Gazprom" in the period from 01.12.2019 to 30.11.2020, obtained from the website of the Investment Company "FINAM". A comparative analysis of methods for detecting and correcting outliers by standard deviation has been carried out. The Z-score statistical method allows you to accurately determine the distance from the suspicious observation to the distribution center, which is an advantage. The disadvantage of this method is the influence of outliers on the mean and standard deviation, which can contribute to the masking of outliers or their incorrect detection. The isolation forest method recognizes outliers of various types, and when implementing the method, there are no parameters that require selection; but the disadvantage is the slower detection rate of outliers compared to other methods. The support vector machine is a very fast method and is reduced to solving a quadratic programming problem, which always has a unique solution. The winsorization method for correcting outliers reduces the effect of outliers on the mean and variance, which is an advantage, but may introduce bias due to the selection of thresholds to separate*

observations in the sample. The multiple imputation method creates for each missing value not one, but many imputations, which avoids a systematic error, but at the expense of high computational costs. For the initial data used in the work, the best result was shown by the implementation of the multiple imputation algorithm based on the detected outliers by the support vector method.

Conclusion. *There is no universal method for detecting and/or eliminating outliers in data analysis theory. In general, the determination of outliers is subjective, and the decision is made individually for each specific dataset, considering its characteristics or existing experience in this area. The practical implementation of the methods for detecting and eliminating outliers used in this work can be a tool for calculating more accurate indicators in any area, for example, to improve forecasting the stock price. As part of further work, it is possible to consider the optimization of the parameters used in the methods of detecting and correcting outliers to study their effect on the results of the models.*

Keywords: *outlier detection, outlier correction, Python programming language, standard deviation.*

Введение

В простейшем случае выброс представляет собой наблюдение, которое существенно отличается от остальных наблюдений набора данных. Основная проблема выявления выбросов состоит в определении того, действительно ли наблюдения, несовместимые с остальными данными, являются выбросами. Также является проблемой и то, что в теории анализа данных нет универсального метода обнаружения и/или устранения выбросов. В качестве примера можно привести работу [6], к которой представлены алгоритмы для анализа выбросов, включая сведения о том, когда и почему конкретные алгоритмы работают эффективно; обсуждаются последние идеи в этой области, такие как ансамбли выбросов, матричная факторизация, ядерные методы и нейронные сети.

Проблема обнаружения выбросов на основе статистических методов и методов машинного обучения затрагивалась во многих отечественных и зарубежных работах. Некоторые методы обнаружения выбросов можно найти, например, в работах [2, 9, 10, 14, 15, 17, 18]. Описанные в этих работах подходы к обнаружению выбросов основаны на статистических методах и методах машинного обучения и представляют собой методы обнаружение выбросов с помощью кластеризации с учителем и без учителя, на основе искусственных нейронных сетей, способом построения профилей, представляющих реальное и идеальное поведения, с использованием методов машинного обучения с построением самосовершенствующейся системы на основе предыдущих знаний, гибридные методы.

После определения выбросов, следующим шагом является их устранение, т. е. либо коррекция точек данных до их правильных значений, либо удаление таких наблюдений из набора данных. Для сглаживания атипичных данных предлагаются различные эвристические методы, вплоть до исключения атипичных точек (см., например, [3, 5, 7, 12, 15, 19]): исправление выброса до правильного значения; удаление выброса из данных; подход, требующий продолжать признавать наличие выброса, но ничего не делать с его значением; метод анализа результатов с выбросами и без них; винсоризация, предполагающая преобразование крайних значений в заданный процентиль данных; усечение, при котором происходит установка наблюдаемых значений в пределах правдоподобного диапазона и исключение других значений из набора данных; преобразование, предполагающим применение детерминированной математической функции (например, логарифмической функции) к каждому значению, чтобы не только сохранить выброс данных в анализе, но также уменьшить дисперсию ошибок; модификация, т.е. изменение значения выброса вручную на другое, менее экстремальное значение; двухэтапная робастная процедура; прямой робастный метод с использованием итеративно взвешенных наименьших квадратов и некоторые другие.

Как известно, проблема выбросов возникает в задачах построения экономических трендов. В регрессионном анализе работа над выбросами традиционно состоит из трех этапов: обнаружение выбросов, устранение выбросов, построение регрессионной зависимости. В работе [13] предложен комбинированный подход к преобразованию данных в регрессионном анализе, объединяющий (после обнаружения выбросов) процедуру коррекции выбросов и построение регрессионной зависимости. При этом коррекция выбросов строится на основе линейных преобразований исходных данных.

Обнаружение и коррекция выбросов, т.е. значений данных, которые не соответствуют общему их распределению, является важным вопросом в том числе во многих экономических областях, т. к. выбросы могут информировать о критически важных событиях. Так, например, процедура обнаружения частого использования кредитной карты в финансовых транзакциях для противодействия мошенничеству в банке относится к работе с выбросами в данных; об этом идет речь, например, в работах [1, 4]. При этом в большинстве случаев злоумышленники приспосабливаются таким образом, чтобы наблюдения, являющиеся выбросами, казались нормальными, что затрудняет задачу определения аномального поведения. Таким образом, актуальность исследований, связанных с ана-

лизом выбросов, не вызывает сомнения.

Возникает вопрос считать ли выбросами те данные, которые на первый взгляд существенно не отличаются от остальных наблюдений набора данных. Обсуждение этого вопроса в статье проводится с использованием методов машинного обучения. В общем случае определение выбросов субъективно, и решение принимается индивидуально для каждого конкретного набора данных с учетом его особенностей или имеющегося опыта в данной области.

Задача исследования состоит в применении методов машинного обучения для анализа котировок акций на финансовых рынках и нахождения наилучшего подхода к решению вопроса обнаружения и коррекции выбросов по критерию среднего квадратического отклонения. В данной работе для обнаружения и коррекции выбросов рассматриваются методы машинного обучения, математическую основу которых составляют метод Z-score, метод изолирующего леса, метод опорных векторов для обнаружения выбросов и методы винсоризации и множественного вменения для коррекции выбросов (см., например, [8, 11, 12, 14, 15]).

Статистический метод Z-score проверяет каждый предполагаемый выброс и позволяет точно определить расстояние от подозрительного наблюдения до центра распределения, что является преимуществом. Но метод Z-score имеет недостаток, который заключается в том, что среднее значение и стандартное отклонение, которые используются в вычислении, подвержены влиянию выбросов, что может способствовать маскировке выбросов или некорректному их обнаружению.

Метод изолирующего леса распознает выбросы различных видов: как изолированные точки с низкой локальной плотностью, так и кластеры выбросов малых размеров. Также при реализации метода отсутствуют параметры, требующие подбора. Но недостатком является скорость обнаружения выбросов, которая не является его сильной стороной по сравнению с другими методами.

Метод опорных векторов считается очень быстрым методом и сводится к решению задачи квадратичного программирования, которая всегда имеет единственное решение. Недостатком метода является отсутствие устойчивости к шуму.

Использование метода винсоризации для коррекции выбросов снижает влияние выбросов на среднее значение и дисперсию, которые очень чувствительны к наличию выбросов, что является преимуществом, но в то же время может вносить систематическую ошибку из-за подбора пороговых значений для разделения наблюдений в выборке.

Метод множественного вменения является надежным и информативным методом, т. к. для каждого пропущенного значения создается не одно, а множество вменений и позволяет избежать систематической ошибки, но за счет высоких вычислительных затрат. Когда имеется большой объем данных множественное вменение сложно реализовать.

Для построения моделей использован программный инструмент Jupyter Notebook, поддерживающий язык программирования Python. Вообще, машинное обучение не привязано к конкретному языку программирования, но существуют языки, традиционно применяющиеся в машинном обучении и имеющие развитую систему библиотек, фреймворков и других инструментов. Одним из распространенных языков для данных целей является Python, в экосистеме которого присутствуют все необходимые библиотеки для использования практически всех существующих алгоритмов и методов машинного обучения и анализа данных. В данной статье результаты работы алгоритмов машинного обучения продемонстрированы для наборов данных, представляющих собой котировки акций трех российских компаний. Проведен сравнительный анализ полученных результатов.

Реализация методов обнаружения выбросов

Проведем сначала описание и предварительную обработку данных, а затем применим метод Z-score, метод изолирующего леса, метод опорных векторов для обнаружения выбросов во временных рядах. Исходными данными являются ежедневные цены закрытия акций компаний «Сбербанк», «Аэрофлот», «Газпром» в период с 01.12.2019 по 30.11.2020, полученные с сайта www.finam.ru [20]. Проиллюстрируем более подробно проведенные исследования для данных компании «Сбербанк». На рис. 1 представлен график изменения цены закрытия акций компаний «Сбербанк» за указанный период.

Для построения моделей будем применять программный инструмент Jupyter Notebook, поддерживающий язык программирования Python. Первый шаг предполагает загрузку необходимых библиотек: numpy, pandas, matplotlib.pyplot и из sklearn – preprocessing, IsolationForest, svm, а также исходных данных через функцию pd.read_csv.

Для визуализации исходных данных будем использовать функцию scatterplot, аргументы которой есть x_data – значения по оси *x*, y_data – значения по оси *y*, x_label – название оси *x*, y_label – название оси *y*, title – заголовок графика.

Для визуализации выбросов мы определим функцию

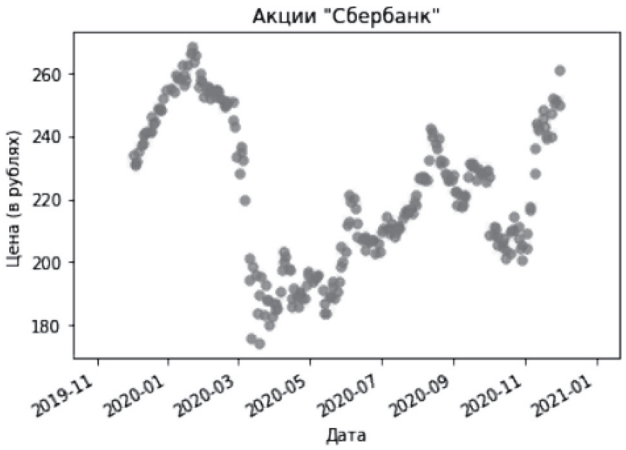

Рис. 1. График цен закрытия акций «Сбербанк»

Fig. 1. Schedule of closing prices for Sberbank shares

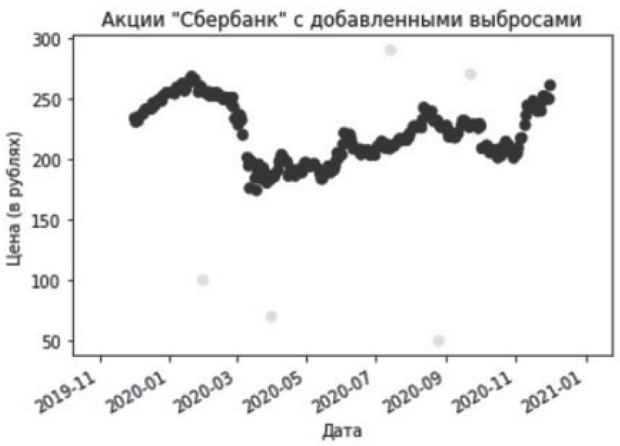

Рис. 2. График цен закрытия акций «Сбербанк» с добавленными выбросами

Fig. 2. Schedule of closing prices of Sberbank shares with added outliers

scatterplot_with_color_coding, часть аргументов которой аналогична аргументам функции scatterplot, и добавлен аргумент «color_code_column» для изображения нормальных данных и выбросов по-разному.

Рассмотрим статистический *метод Z-score* [14]. Как известно, стандартизованная оценка значения x рассчитывается по формуле

$$z = \frac{x - \overline{x}}{\sigma_x},$$

где $\overline{x}$ – математическое ожидание (или среднее значение), σ_x – среднее квадратическое отклонение (СКО), вычисленные для множества данных $\{x_i\}$.

Выбросами будут считаться те значения данных, которые имеют расстояние от $\overline{x}$ более заданного, обычно 2 или 3 СКО. Для реализации метода в Jupyter Notebook воспользуемся функцией z_anomaly_detection, аргументами которой являются: *df* – таблица с данными, column_name – название столбца, в котором отыскиваются выбросы, number_of_stdevs_away_from_mean – количество СКО от среднего значения. Возьмем $z = 2$. Тогда все значения данных, отличающиеся от среднего значения более чем на 2 СКО, помечаются как выбросы.

Проверим работу метода, добавив во временной ряд пять аномальных значений (рис. 2), ожидая, что эти точки будут определены алгоритмом как выбросы. Для этого создадим словарь anomaly_dictionary, в котором ключом является индекс, а значением – цена.

Проверка алгоритма обнаружения с использованием функции z_anomaly_detection показала, что метод Z-score обнаружил четыре из пяти искусственно созданных выбросов. Обнаруженные выбросы в исходных данных представлены на графике (рис. 3), а также в табл. 1.

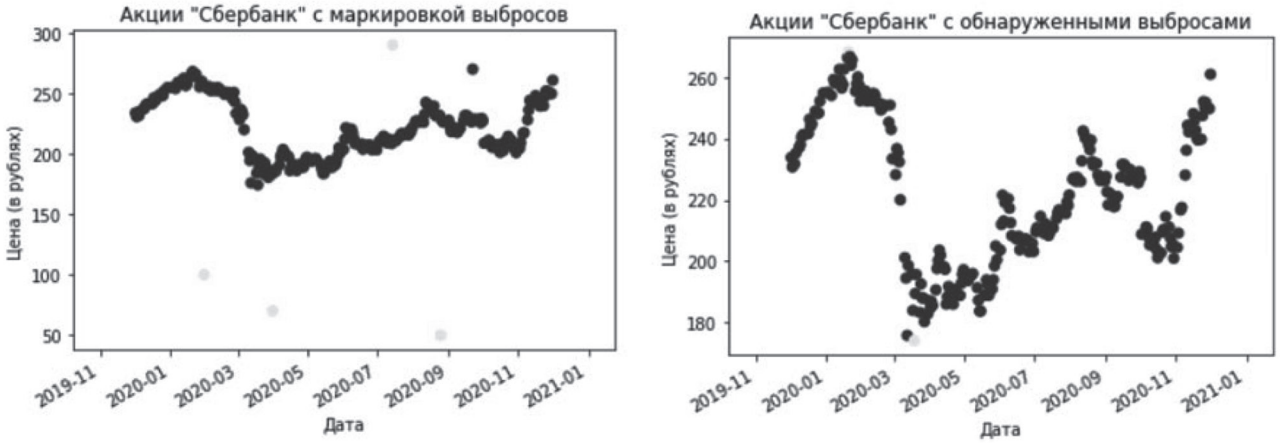

Рис. 3. Графики исходных данных с обнаруженными методом Z-score добавленными выбросами (слева) и выбросами в исходных данных (справа)

Fig. 3. Graphs of the original data with added outliers detected by the Z-score method (left) and outliers in the original data (right)

Выведем таблицу с обнаруженными выбросами в исходных данных (таблица 1).

Таблица 1 (Table 1)

Обнаруженные выбросы в исходных данных методом Z-score

Detected outliers in the original data by the Z-score method

	<DATE>	<CLOSE>	<CLOSE>_Z_Anomaly
32	2020-01-21	268.06	True
71	2020-03-18	174.27	True

Применим теперь *метод изолирующего леса* [11], идея которого основывается на случайном разбиении пространства признаков, таком что в среднем изолированные точки отсекаются от нормальных, кластеризованных данных. Окончательный результат усредняется по нескольким запускам стохастического алгоритма. Алгоритм изолирующего дерева (Isolation Tree) заключается в построении случайного бинарного решающего дерева. Корнем дерева является всё пространство признаков; в очередном узле выбирается случайный признак и случайный порог разбиения, сэмплированный из равномерного распределения на отрезке от минимального до максимального значения выбранного признака. Критерием останова является тождественное совпадение всех объектов в узле, то есть решающее дерево строится полностью. Ответом в листе, который также соответствует anomaly_score алгоритма, объявляется глубина листа в построенном дереве.

Для реализации данного метода используем модель IsolationForest из пакета scikit-learn, которую протестируем на нашем наборе данных. Для этого будем использовать функцию isolation_forest_anomaly_detection, аргументами которой являются df – таблица с данными, column_name – название столбца, в котором ищутся выбросы, outliers_fraction – доля допустимых выбросов.

Как и раньше проверим работу метода. Добавим к исходным данным пять аномальных значений, ожидая, что эти значения будут определены алгоритмом изолирующего леса как выбросы. Для этого создадим словарь anomaly_dictionary, в котором ключом является индекс, а значением – цена. Применяя функцию isolation_forest_anomaly_detection на исходных данных с добавленными значениями, получаем следующий результат: алгоритм изолирующего леса обнаруживает все искусственно созданные выбросы, но при этом метод классифицирует как выбросы не только добавленные аномальные значения. Возможно, что данное явление относится к ложным срабатываниям, а, возможно, точки, не относящиеся к пяти добавленным значениям, действительно являются выбросами. Визуализируем исходные данные с обнаруженными выбросами (рис. 4).

Выведем таблицу с обнаруженными выбросами (табл. 2).

Таблица 2 (Table 2)

Обнаруженные выбросы в исходных данных методом изолирующего леса

Detected outliers in the original data by the isolating forest method

	<DATE>	<CLOSE>	<CLOSE>_Isolation_Forest
26	2020-01-13	262.40	True
30	2020-01-17	262.50	True
31	2020-01-20	266.28	True
32	2020-01-21	268.06	True
33	2020-01-22	266.54	True
34	2020-01-23	263.73	True
35	2020-01-24	265.49	True
38	2020-01-29	259.94	True
67	2020-03-12	175.91	True
71	2020-03-18	174.27	True
78	2020-03-27	180.38	True
79	2020-03-30	183.00	True
249	2020-12-01	260.81	True

Обнаружим выбросы с использованием одноклассного алгоритма *метода опорных векторов* (SVM). В основе метода SVM лежит идея перевода исходных векторов в пространство большей размерности и

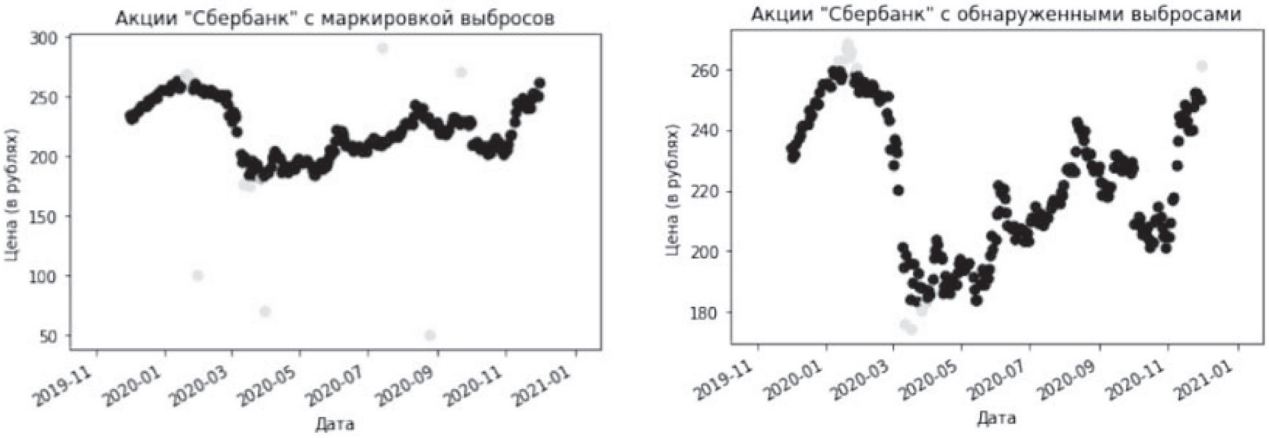

Рис. 4. Графики исходных данных с обнаруженными методом изолирующего леса добавленными выбросами (слева) и выбросами в исходных данных (справа)

Fig. 4. Graphs of the original data with added outliers detected by the isolation forest method (left) and outliers in the original data (right)

построения разделяющей гиперплоскости. После этого переходят к построению двух гиперплоскостей, параллельных разделяющей гиперплоскости так, чтобы опорные векторы (относящиеся к разным классам, ближайшие к разделяющей гиперплоскости) принадлежали им. При таком построении гиперплоскостей образуется полоса, свободная от данных точек. Получение оптимальной разделяющей гиперплоскости, т.е. такой разделяющей гиперплоскости, для которой ширина полосы максимальна, сводится к решению задаче квадратичного программирования [8].

Для реализации метода SVM используем модель OneClassSVM(), из пакета scikit-learn. Построим модель и протестируем ее на нашем наборе данных, используя функцию one_class_SVM_anomaly_detection, аргументами которой являются dataframe – таблица с данными, columns_to_filter_by – название столбца, в котором ищутся выбросы, outliers_fraction – доля допустимых выбросов.

Применяя функцию one_class_SVM_anomaly_detection, получаем график исходных данных с обнаруженными добавленными выбросами (рис. 5). Видно, что метод опорных векторов обнаруживает не все искусственно созданные выбросы, однако также, как и изолирующий лес, он классифицирует и ложные срабатывания, которые возможно являются выбросами на самом деле. Для определения выбросов в исходных данных получим график исходных данных с обнаруженными выбросами.

Выведем таблицу с обнаруженными выбросами (табл. 3).

Таблица 3 (Table 3)

Обнаруженные выбросы в исходных данных методом опорных векторов

Detected outliers in the original data by the support vector machines

	<DATE>	<CLOSE>	One_Class_SVM_Anomaly
14	2019-12-20	244.71	True
15	2019-12-23	248.80	True
16	2019-12-24	248.67	True
20	2019-12-30	254.75	True
21	2020-01-03	255.00	True
31	2020-01-20	266.28	True
35	2020-01-24	265.49	True
67	2020-03-12	175.91	True
71	2020-03-18	174.27	True
98	2020-04-24	188.91	True
99	2020-04-27	188.90	True
211	2020-10-07	210.60	True
249	2020-12-01	260.81	True

Проводились аналогичные расчеты с использованием ежедневных цен закрытия акций компаний «Аэрофлот» и «Газпром». Для данных этих компаний метод Z-score обнаруживает все искусственно созданные выбросы; методы изолирующего леса и опорных векторов обнаруживают все искусственно созданные выбросы, но классифицируют и ложные срабатывания.

Предварительно удалив выбросы, проведем сравнение методов обнаружения выбросов в исходных данных по среднеквадратическому отклонению: $\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$, где x_i – i-й элемент множества данных, $\overline{x}$ – среднее значение, n – количество элементов множества данных.

Таблица 4 (Table 4)

Значения СКО для методов обнаружения выбросов

Standard deviation values for outlier detection methods

	Аэрофлот	Газпром	Сбербанк
Z-score	15,06	24,65	22,87
Isolation Forest	15,78	26,8	21,56
SVM	15,35	26,55	22,29

Согласно таблице 4 получаем, что для данных компаний «Аэрофлот» и «Газпром» метод Z-score приводит к меньшим значениям СКО и, тем самым лучше выявляет выбросы, а для данных компании «Сбербанк» лучшим оказался метод изолирующего леса. Таким

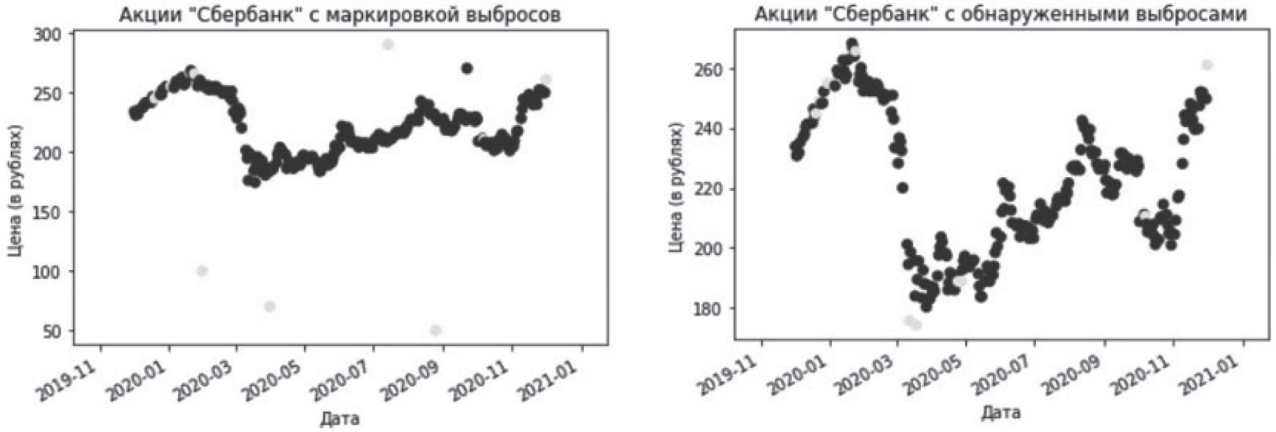

Рис. 5. Графики исходных данных с обнаруженными методом опорных векторов добавленными выбросами (слева) и выбросами в исходных данных (справа)

Fig. 5. Graphs of original data with added outliers detected by the support vector machines (left) and outliers in the original data (right)

образом, на данных выборках ни один из методов не является наилучшим, т. к. каждый метод может давать неточные результаты обнаружения выбросов. Поэтому, чтобы точно говорить являются ли определенные экземпляры данных выбросами, целесообразно на каждом наборе данных проверять несколько методов.

Методы винсоризации и множественного вменения для коррекции выбросов

После обнаружения выбросов в данных перейдем к следующему этапу работы с выбросами – устранению.

Основной проблемой устранения выбросов является то, что в большинстве методов применяется удаление обнаруженных выбросов, что на самом деле не всегда является верным. Например, удаление выбросов может привести к важным изменениям в результатах эксперимента. Поэтому далее будут рассмотрены методы устранения выбросов, которые не удаляют обнаруженные выбросы, а корректируют их значения. Основными такими методами являются винсоризация (winsorizing) и множественное вменение (multiple imputation).

Винсоризация – это процедура, которая уменьшает влияние выбросов на среднее значение и дисперсию [12]. Винсоризация включает определение пороговых значений. Выборочные наблюдения, значения которых лежат за пределами определенных заранее установленных значений, преобразуются, чтобы приблизить их к значению отсечения. Исходные данные, лежащие за пределами пороговых значений, преобразуются, чтобы они больше не считались выбросами. Однако следует отметить, что измененные значения являются искусственными и иногда могут быть неприемлемыми. Процесс винсоризации заключается в следующем: любое значение переменной выше или ниже квантиля k на каждой стороне распределения переменных заменяется значением самого k-го квантиля. Например, если $k = 5$, то все наблюдения выше 95-го квантиля изменяются на значение этого 95-го квантиля, а значения ниже 5-го квантиля аналогично изменяются на соответствующее значение. По сравнению с простым удалением выбросов коррекцию с помощью винсоризации можно считать менее экстремальным вариантом, поскольку она «перекодирует» выбросы, а не удаляет их полностью. Часто берут $k = 5$ и, следовательно, оно используется как значение по умолчанию. В результате винсоризации мы получаем модифицированную выборку, в которой выбросы изменены нормальными значениями. Дальнейшие расчеты проводятся для модифицированного образца.

Для реализации метода винсоризации в Jupyter Notebook будем использовать функцию winsorize из библиотеки scipy, аргументами которой являются набор данных и limits – кортеж процентных значений, которые нужно вырезать на каждой стороне массива.

На рис. 6 представлены результаты применения метода для акций компании «Сбербанк».

Рассмотрим теперь *метод множественного вменения*. Метод вменения часто используется при обработке пропущенных значений; но он также может быть применен и при работе с выбросами. При этом выбросы удаляются (становятся как бы пропущенными значениями), а затем заменяются оценками, основанными на оставшихся данных.

Методы вменения можно разделить на две подгруппы: однократное вменение и множественное вменение. Большинство методов однократного вменения все выбросы или пропущенные значения заменяются средним значением, не учитывая неопределенность вмененных значений. Это означает, что методов однократного вменения распознают вмененные значения как фактические, не принимая во внимание стандартную ошибку, что приводит к смещению результатов. Поэтому лучшей альтернативой и более надежным методом вменения является множественное вменение. Множественное вменение – это метод вменения, основанный на статистике, когда для каждого пропущенного значения создается не одно, а множество вменений. Это означает многократное заполнение пропущенных значений, создание

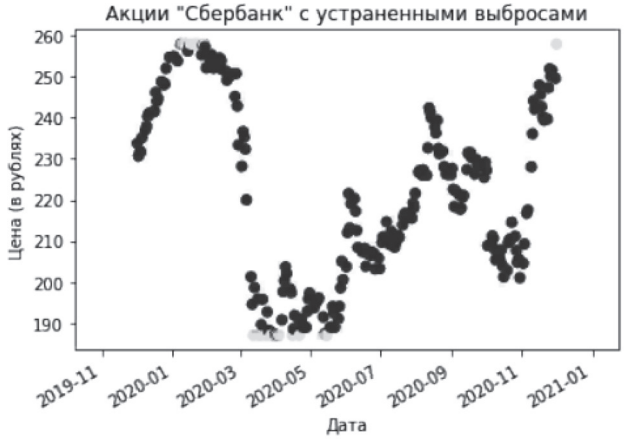

Рис. 6. График исходных данных с устраненными выбросами методом винсоризации
Fig. 6. Graph of the initial data with eliminated outliers by the winsorization method

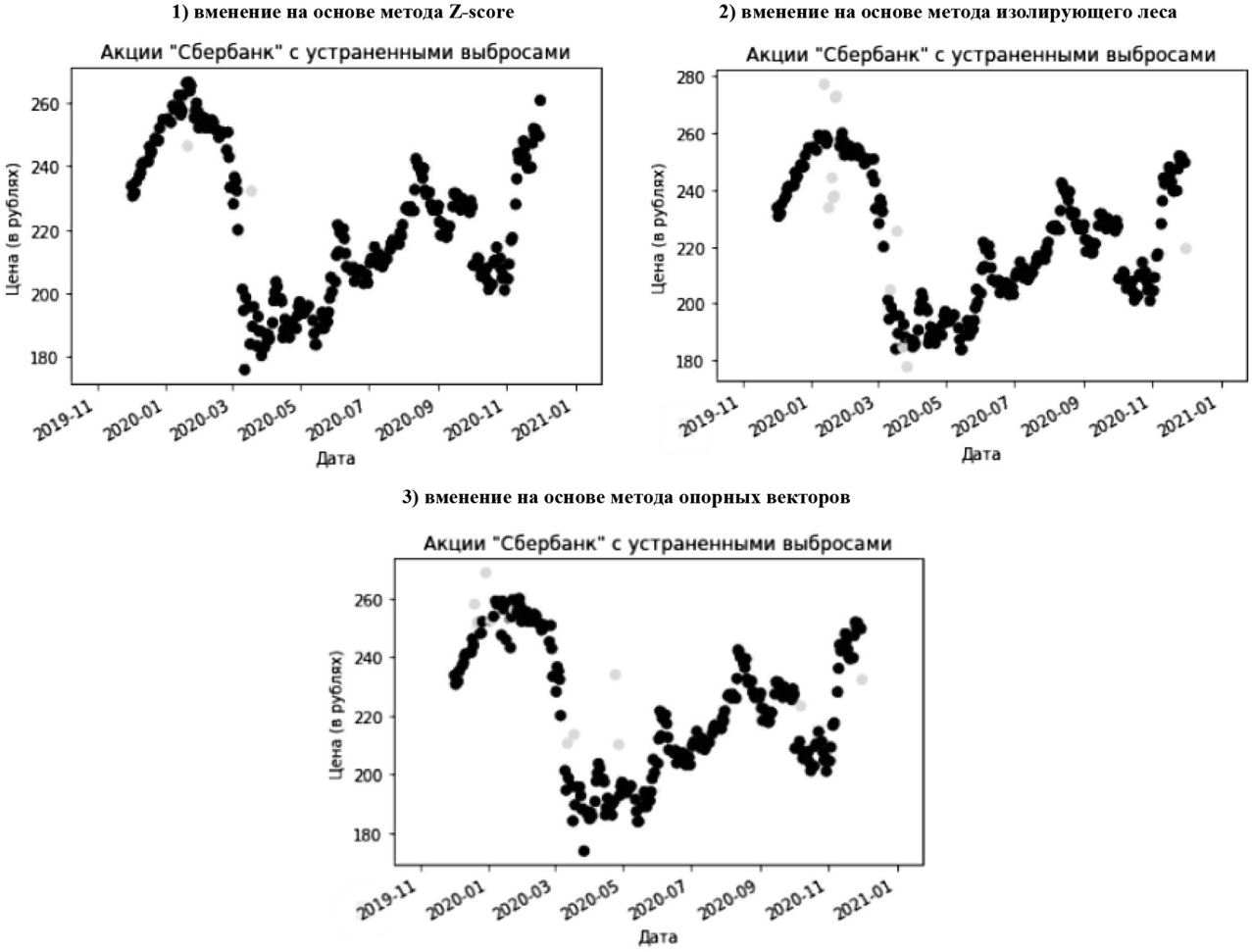

Рис. 7. Графики исходных данных с устраненными выбросами методом вменения на основе трех методов обнаружения выбросов

Fig. 7. Graphs of initial data with the eliminated outliers by imputation method based on three outlier detection methods

нескольких «полных» наборов данных [15]. Для множественного вменения был разработан ряд алгоритмов. Одним из хорошо известных алгоритмов является множественное вменение по цепному уравнению (Multiple Imputation by Chained Equations (MICE)) [15].

На практике множественное вменение в Python реализуется не так просто, однако в библиотеке sklearn есть функция IterativeImputer, которая может быть использована для множественного вменения. Главный ее аргумент, который мы будем использовать это max_iter – максимальное количество раундов вменения, которые необходимо выполнить перед возвратом вменений. Обычно выполняется 10 вменений и по умолчанию функция использует max_iter=10, поэтому будем использовать это значение. И раз мы применяем множественное, а не однократное вменение, то аргумент sample_posterior надо изменить на True, который по умолчанию принимает значение False.

Обнаруженные раннее в исходных данных выбросы нужно удалить, сделав их пропущенными значениями. Так как мы применяли три метода обнаружения выбросов, то и вменение мы будем проводить для каждого набора выбросов соответствующих компаний.

На рис. 7 показаны результаты множественного вменения с обнаруженными различными методами выбросами для акций компании «Сбербанк».

Методы винсоризации и множественного вменения были применены и для данных акций компаний «Аэрофлот» и «Газпром».

Результаты сравнения методов коррекции выбросов по среднеквадратическому отклонению (СКО) для акций рассматриваемых компаний представлены в таблице

Таблица 5 (Table 5)

Значения СКО для методов коррекции выбросов

Standard deviation values for outliers' correction methods

	Аэрофлот	Газпром	Сбербанк
Винсоризация	16,234	27,371	22,456
Z-score и вменение	14,903	25,9	22,883
Isolation Forest и вменение	14,809	25,493	22,369
SVM и вменение	14,772	25,364	22,072

Как можно увидеть лучший результат дает совместное применение метода опорных векторов (SVM), который обнаруживает выбросы, и метода множественного вменения, который их устраняет.

Заключение

В работе рассмотрены методы обнаружения и коррекции выбросов, показаны их возможности и реализация на реальных данных.

Для использованных в работе исходных данных лучший результат показала реализация алгоритма множественного вменения по обнаруженным методом опорных векторов выбросам.

Практическая реализация методов обнаружения и устранения выбросов, предложенная в данной работе, может являться инструментом вычисления более точных показателей в любой сфере, например, для улучшения прогнозирования цены акции.

В рамках дальнейшей работы можно рассмотреть оптимизацию параметров, используемых в методах обнаружения и коррекции выбросов, для исследования их влияния на результаты работы моделей.

Литература

1. Ардан С. Д. Прогнозирование банкротства методами машинного обучения // Информационное общество. 2021. № 1. С. 56–67.

2. Девянин И.С. Предварительная обработка данных для машинного обучения // Фундаментальные и прикладные исследования в физике, химии, математике и информатике. 2021. С. 117–121.

3. Копырин А.С., Видищева Е.В. Оценка влияния аномалий на результаты анализа массивов экономических данных // Modern Economy Success. 2021. № 2. С. 235–240.

4. Чернова В.В. Применение методов машинного обучения для выявления аномалий с банковскими картами // МНСК-2021. 2021. С. 107–107.

5. Шибзуков З.М. О принципе минимизации эмпирического риска на основе усредняющих агрегирующих функций // Доклады Академии наук. 2017. Т. 476. № 5. С. 495–499.

6. Aggarwal C.C. Outlier analysis. Springer Science & Business Media. 2013. 455 с.

7. Aguinis H., Gottfredson R.K., Joo H. Best-practice recommendations for defining, identifying, and handling outliers // Organizational Research Methods. 2013. Т. 16. № 2. С. 270–301.

8. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey // ACM Computing Surveys. 2009. Т. 41. № 3. С. 15–58.

9. Cousineau D., Chartier S. Outliers detection and treatment: A review // International Journal of Psychological Research. 2010. Т. 23. № 1. С. 59–68.

10. Cunningham P., Cord M., Delan, S. J. Supervised Learning // Machine Learning Techniques for Multimedia. Springer Berlin Heidelberg. 2008. С. 21–49.

11. Fei T.L., Kai M.T., Zhi-Hua Zhou. Isolation Forest // 2008 Eighth IEEE International Conference on Data Mining. 2008.

12. Frey B.B. The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation. 2018.

13. Gorelik V., Zolotova T. Method of Parametric Correction in Data Transformation and Approximation Problems // Lecture Notes in Computer Science (LNCS). Springer. Т. 12422. С. 122–133.

14. Hodge V., Austin J. A survey of outlier detection methodologies // Artificial Intelligence Review. 2004. Т. 22. № 2. С. 85–126.

15. Khan S.I., Hoque A.S. M.L. SICE: an improved missing data imputation technique. J Big Data. 2020. № 7. Article number 37.

16. Omar S., Ngadi A., Jebur H. Machine Learning Techniques for Anomaly Detection: An Overview // International Journal of Computer Applications. 2013. Т. 79. № 2. С. 33–41.

17. Patcha A., Park J. M. An overview of anomaly detection techniques: Existing solutions and latest technological trends // Computer Networks. 2007. Т. 51. № 12. С. 3448–3470.

18. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. Wiley Interdiscip Rev Data Min Knowl Discov. 2011. № 1(1). С. 73–9.

19. Zimek A., Filzmoser P. There and back again: Outlier detection between statistical reasoning and data mining algorithms // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018. Т. 8. № 6. С. 73–79.

20. Инвестиционная компания «ФИНАМ» [Электрон. ресурс]. Режим доступа: https://www.finam.ru/.

References

1. Ardan S.D. Bankruptcy forecasting using machine learning. Informatsionnoye obshchestvo = Information society. 2021; 1: 56-67. (In Russ.)

2. Devyanin I.S. Preliminary data processing for machine learning. Fundamental'nyye i prikladnyye issledovaniya v fizike, khimii, matematike i informatike = Fundamental and applied research in physics, chemistry, mathematics and informatics. 2021: 117-121. (In Russ.)

3. Kopyrin A.S., Vidishcheva Ye.V. Evaluation of the impact of anomalies on the results of the analysis of economic data arrays. Modern Economy

Success = Modern Economy Success. 2021; 2: 235-240. (In Russ.)

4. Chernova V.V. Application of machine learning methods to detect anomalies with bank cards. MNSK-2021 = MNSK-2021. 2021: 107-107. (In Russ.)

5. Shibzukov Z.M. On the principle of empirical risk minimization based on averaging aggregating functions. Doklady Akademii nauk = Reports of the Academy of Sciences. 2017; 476; 5: 495–499. (In Russ.)

6. Aggarwal C.C. Outlier analysis. Springer Science & Business Media. 2013. 455 p.

7. Aguinis H., Gottfredson R. K., Joo H. Best-practice recommendations for defining, identifying, and handling outliers. Organizational Research Methods. 2013; 16; 2: 270–301.

8. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey. ACM Computing Surveys. 2009; 41; 3: 15–58.

9. Cousineau D., Chartier S. Outliers detection and treatment: A review. International Journal of Psychological Research. 2010; 23; 1: 59–68.

10. Cunningham P., Cord M., Delan, S. J. Supervised Learning. Machine Learning Techniques for Multimedia. Springer Berlin Heidelberg. 2008: 21–49.

11. Fei T.L., Kai M. T., Zhi-Hua Zhou. Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining. 2008.

12. Frey B.B. The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation. 2018.

13. Gorelik V., Zolotova T. Method of Parametric Correction in Data Transformation and Approximation Problems. Lecture Notes in Computer Science (LNCS). Springer; 12422: 122-133.

14. Hodge V., Austin J. A survey of outlier detection methodologies. Artificial Intelligence Review. 2004; 22; 2: 85–126.

15. Khan S.I., Hoque A.S. M.L. SICE: an improved missing data imputation technique. J Big Data. 2020: 7. Article number 37.

16. Omar S., Ngadi A., Jebur H. Machine Learning Techniques for Anomaly Detection: An Overview. International Journal of Computer Applications. 2013; 79; 2: 33–41.

17. Patcha A., Park J. M. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks. 2007; 51; 12: 3448–3470.

18. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. Wiley Interdiscip Rev Data Min Knowl Discov. 2011; 1(1): 73–9.

19. Zimek A., Filzmoser P. There and back again: Outlier detection between statistical reasoning and data mining algorithms. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018; 8; 6: 73–79.

20. Investitsionnaya kompaniya «FINAM» = Investment company «FINAM» [Internet]. Available from: https://www.finam.ru/. (In Russ.)

Сведения об авторах

Татьяна Валерьяновна Золотова
Д.ф.-м.н., профессор
Финансовый университет
при Правительстве РФ, Москва, Россия
Эл. почта: tzolotova@fa.ru

Дарья Александровна Волкова
Российский университет дружбы народов,
Москва, Россия
Эл. почта: 174557@edu.fa.ru

Information about the authors

Tatiana V. Zolotova
Dr. Sci. (Physics and Mathematics), Professor
Financial University under the Government of the Russian Federation, Moscow, Russia
E-mail: tzolotova@fa.ru

Daria A. Volkova
Peoples' Friendship University of Russia,
Moscow, Russia
E-mail: 174557@edu.fa.ru