Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM
e-ISSN: 2788-7421 p-ISSN: 2958-0544

Detection of anomalies and Data Drift in a time-series dismissal prediction system

Nataliya Boyko[1*], Roman Kovalchuk[1]

[1]Department of Artificial Intelligence Systems, Lviv Polytechnic National University,
12 S. Banrdera Str., Lviv, 79013, Ukraine

*Corresponding Author: Nataliya Boyko

DOI: https://doi.org/10.30880/ijcsm.2024.05.03.012
Received February 2024; Accepted April 2024; Available online July2024

ABSTRACT: The purpose of the study is to develop a system that automatically processes data based on existing and newly entered data, especially with the aim of ensuring high data quality by detecting and eliminating anomalies. The quantile filtering method, Chebyshev's inequality, Kolmogorov-Smirnov two-sample test, and others should be noted among the methods used. In the course of the research, the theoretical aspects of the methods, various principles of detecting anomalies for different types of data were considered and analysed. Different principles and approaches applied to anomaly detection in different contexts were explored. The results of the analysis and the selection of optimal methods for detecting anomalies in various types of data are important for the effective functioning of the automatic data processing system. This will make it possible to achieve accuracy and reliability in the detection of anomalies and ensure high quality of data used in the machine learning system.

Keywords: Data Quality pipeline, multimodal data, logical data, numerical data, machine learning algorithm

1. INTRODUCTION

Data quality is a critical issue in machine learning systems, especially those dealing with Time-Series data that is constantly being updated. Ensuring high-quality and consistent data inputs through automated pipelines has become essential for the smooth functioning and reliability of such systems. This study focuses on developing an automated Data Quality Pipeline for Time-Series data using a machine learning model for employee retention as an example case.

The complexity of such tasks in the real world consists of the constant updating of data, the need for their structuring, processing, processing and further application in an intelligent data analysis system. A pipeline advantages for mechanically process updated inputs is obvious. The more pristine the information is, the smoother it becomes for the machine learning algorithm to process, leading to increased predictability of the outcome. The aforementioned pipeline for ensuring Data Quality encompasses the following stages:

- Downloading information retrieved from the data archive.
- Data pre-processing and filtering.
- Finding irregularities present in the data.
- Identifying variance between new and past anomalies.
- Development of an interactive report utilizing updated data.
- Verifying Data Drift in the new dataset.
- Importing pre-processed data into the database.
- Recording of parameters, metrics, and the outcomes of pipeline execution.

It should be noted that the above case is conditional. Still, issues like this should be considered when studying the distribution of salaries as practice, i.e., when finding abnormalities observed in the salary column. When checking the Data Drift column, it is important to take into account the nuances. If salaries, salary distribution, and average value for "Quality Management" have decreased by k% in the last four months while maintaining an identical distribution, historical data indicates a shift. However, no Data Drift should be recorded apart from the conditional mean, as the distribution of salaries remains similar.

In the study of A. Iturria et al. [1], a new algorithmic sequence is developed to improve the process of predicting the occurrence of errors in time series by means of online tools. The framework that has been proposed includes the normalization of streaming data, as well as the online assessment and identification of anomalies using prediction errors. It proposes the use of neural networks in a series of online recurrent extreme learning machines is proposed. To solve the problem of predicting the occurrence of errors in time series using neural networks EORELM-AD, which

have a high level of adaptation (according to the results of the study) to perform the target functions of on-line prediction of defects in time series data.

The study of P. Kumari and M. Saini [2] examines the system for detecting anomalies and defects in the time series of audio-visual data arising from the functioning of surveillance and monitoring equipment, which form an urgent problem of detecting and fixing these deviations, which may not be noticed by the operator and not receive an appropriate response, creating consequential threats. It is proposed to use neural network tools for deep learning of multimodal data with their current correlation and defect detection. To solve the problem of detecting anomalies and defects in the time series of audio-visual data, it is proposed to use neural network tools that undergo deep learning with simultaneous correlation of audio and video streams of timing data used in surveillance and monitoring systems.

The study of C. Hegde [3] notes the low efficiency of modern tools for analysing and detecting errors in time series of data obtained in industrial sectors, where the time parameter is key. The use of analysis tools using AI on the model of functioning, focused on streaming data, is proposed. To solve the problem of detecting errors and anomalies in the time series of industrial information flow data, the effectiveness of using AI based on the streaming data-oriented functioning model is suggested and proved. Therefore, it is necessary to consider the analysis of the operationalization of the selected solutions, where amidst the main conditions are:

- autonomy;
- adaptability of visualization;
- flexibility of modification of the logic of the anomaly detection method;
- withstand distribution shifts.

X.X. Yin et al. [4] developed an effective tool for detecting anomalies in time series and predicting them using machine learning methods. The study suggests Time Series Based Data Explorer, which is a tool for advanced time series analysis. This tool provides the ability to visualize and explore time series, identify their features, and perform regression, convolution, and other operations for further analysis. In addition, the proposed method uses machine learning algorithms, such as recurrent neural networks, to model and forecast time series, as well as to detect anomalies in streaming data.

In their work, D. Sulem et al. [5] proposed a new method that explains the anomalies found in time series by generating counterfactual explanations. Counterfactual explanations are alternative scenarios that show how the data must change to remove an anomalous observation. The study proposes a new approach to counterfactual explanations using generative models and machine learning methods. The authors experiment with different model architectures and show that their methods are capable of generating diverse and reasonable counterfactual explanations.

The objective of this study is to develop a pipeline that automates data filtering, anomaly detection, comparison against historical data, and reporting of data drift within an existing machine learning system. The pipeline will carry out these tasks each time new data arrives so that high-quality inputs are fed into the model, preventing deterioration in its predictive capability. This study analyzes suitable approaches and develops an optimal solution for the modeling context. The automated pipeline aims to improve model performance by ensuring pristine, consistent data.

2. THEORETICAL OVERVIEW

2.1 METHODS FOR ANOMALY DETECTION AND APPLICATION TO SPECIFIC FEATURES

Classically, among the statistical methods for anomaly detection, there are different conditions for using particular methods, which narrow its application to some extent. Some methods rely solely on the empirical distribution's normality. On the other hand, some methods can work regardless of the normality of the empirical distribution. For instance, Chebyshev's Inequality method does not require normality of the empirical distribution. The method uses the *k*-sigma rule, better known as "Three sigmas for normal distribution". Therefore, that the data's characteristics when selecting an anomaly detection method for each feature should be considered. Data characteristics could be the variation, the median, the mean, and the type of distribution, normal or non-normal. The MonthOnSalary column is incremental, corresponding to the number of months a person is on the same salary. After conducting statistical tests, it was concluded that it does not correspond to a normal distribution, and therefore Chebyshev's boundary method will be used (Fig. 1).

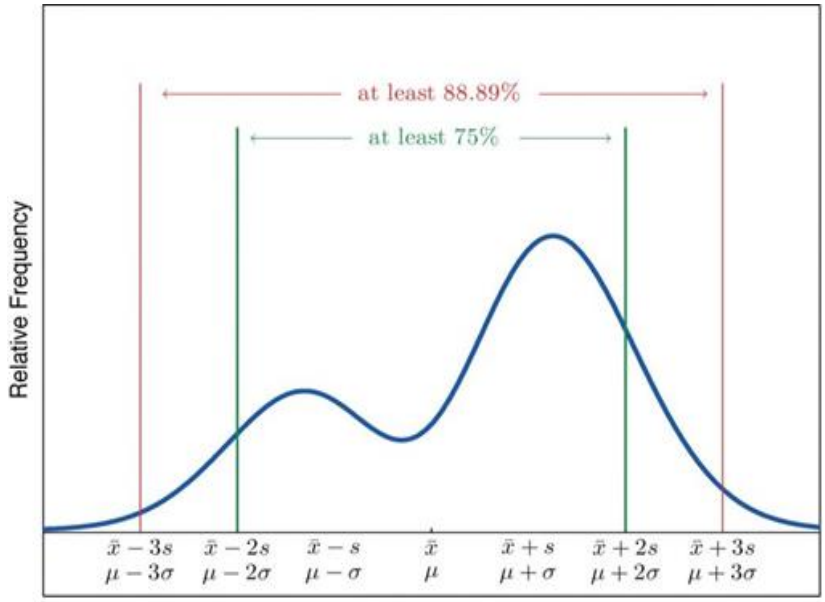

FIGURE 1. - Percentage of samples within Chebyshev's inequality boundaries (depending on *k*-value)

It could be clear that the dependence of the minimum number of samples that fall into a non-normal distribution, respectively, on the selected value of *k*, in this case on *k*=1, 2, 3 (Fig. 2).

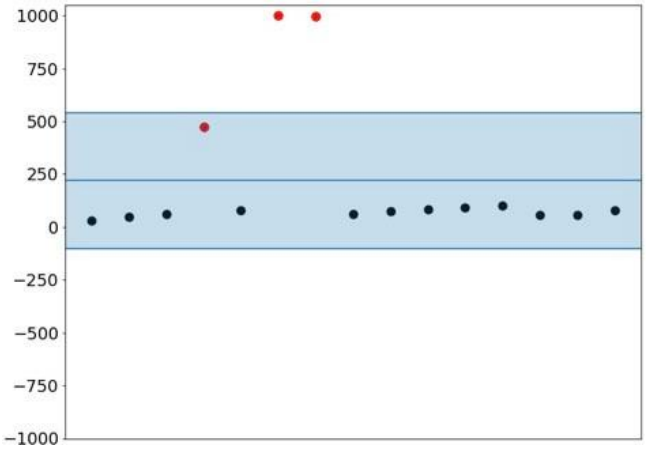

FIGURE 2. - Anomaly detection, using Mean+STD

The application performance management (APM) column is numerical, responsible for the percentage of how much profit the company receives relative to the employee's salary. Because the median and mean are radically different, there is high variation, with values deviating from the median by up to 170 thousand percent, which is an apparent anomaly for a normal distribution [6]. The influence of radical values on the average is shown, and this method is proposed as an alternative for finding anomalies (Fig. 3).

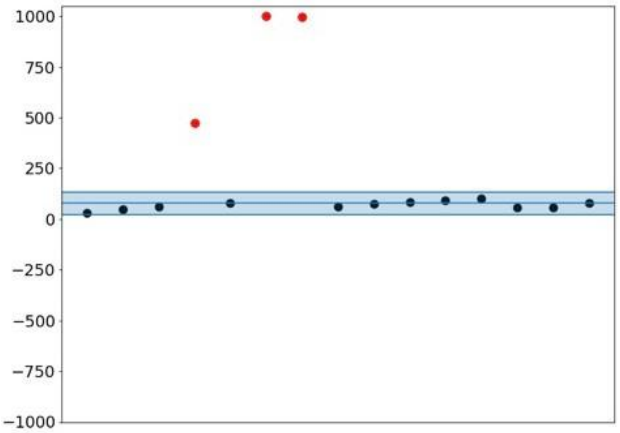

FIGURE 3. - Anomaly detection, using Median+IQR

In conclusion, Median+IQR is a good method for finding anomalies when methods using the mean or standard deviation cannot be applied due to their susceptibility to contamination by anomalies. However, the method is sensitive to the characteristics of the distribution, therefore, it may be more appropriate to use it in conjunction with other methods for the certainty of the result. The WageGross feature is numerical, responsible for the employee's salary, and the distribution is multimodal. Multimodality may be due to the overlap of many normal distributions. For example, the superimposed salaries of Junior Java Software Engineer, Intermediate Java Software Engineer, and Senior Java Software Engineer are different, and together they form a multimodal distribution. To conduct accurate search for anomalies in this column, such a cumulative distribution should be grouped by ManagementLevel, and by the already mentioned JobFamilyGroup, which would provide with more "normal" distributions that are easier to work with (Fig. 4).

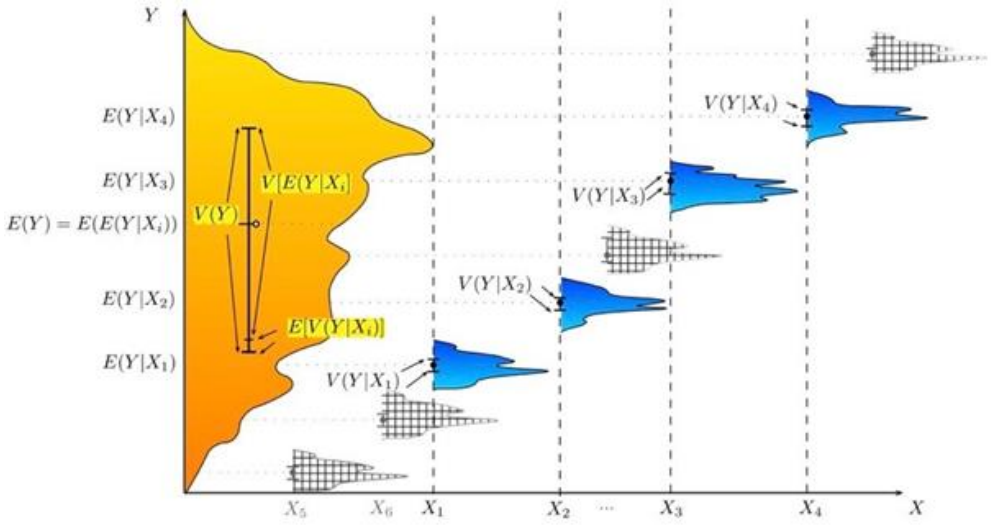

FIGURE 4. - Example of dividing multimodal distribution by the ANOVA method

Source: A. Gelman [7].

For the already mentioned APM column, it is possible to improve the result of finding anomalies by using the machine learning method, Isolation Forest. This method uses the contamination parameter, which corresponds to the set's anomaly contamination percentage [8]. It is necessary to determine the percentage of pollution from Median+IQR and pass it as a parameter to Isolation Forest. Using these two methods in a pair, avoid selecting the contamination parameter since each pipeline run will regenerate it based on the existing distribution. Since the data in this problem is not static, it is a Time-Series problem; hence it is essential to support the automatic determination of all possible parameters and not to set them manually. Automation will ensure the independence of the final solution, as there will be less need to adjust the pipeline manually.

2.2 METHODS FOR DETECTING DATA DRIFT

Regarding the problem of determining Data Drift from the point of view of designing a machine learning system, it is clear that there are two general approaches to solving it. The first, more general, is the use of ready-made packages, for example, Evidently AI, Tensorflow Data Validation, which offer a more general solution for verifying Data Drift between two datasets. The second is using self-made methods of validation based on self-selected statistical tests. As mentioned, package tools use a common approach for all features. For instance, the WageGross column has a multimodal distribution for which the general approach will not work. Evidently AI uses the Kolmogorov-Smirnov test for numeric columns, the Pearson-Chi-Square test for categorical data, and the Z-Test for binomial data. Although the test for the Evidently AI interface was written, its signature must strictly accept two datasets, the same as for all other columns, and return a P-Value, which limits from using the tests (Fig. 5).

FIGURE 5. - Example of Data Drift report generated by Evidently AI

Source: Data Drift report [9].

Evidently AI has a very modest set for customizing the Data Drift report, although it uses Plotly for visualization. For example, if it is needed to change the batch distribution graph to the own, with a different name of the x or y axis, add details when hovering, sort, or group the result, Evidently AI will not suit because it has no tools for customization. Based on the mentioned information, absence of customizing the mentioned parameters, it was concluded that the use of Evidently AI is not appropriate in this case. In order to determine Data Drift, it is necessary to test the null hypothesis suggesting that two datasets are equivalent. The first is a dataset with historical records before the inspection date, the other with new data after the inspection date, and which also includes historical data. Two-Sample Test by The Kolmogorov-Smirnov, the Z-Test if the normal distribution and the Mann-Whitney U Rank Test could be used to check numerical columns. When choosing tests, one should, first of all, consider their limitations [10-14].

In order to eliminate the shortcomings of the above tests, they could be combined and used as a more general solution, for instance: with small numbers of samples ($n<100$) – Mann-Whitney U Rank Test; with the number of samples ($n>100$) – Kolmogorov-Smirnov Two-Sample Test; with normal distribution – Z-Test. Speaking of a test that only works with the normal distribution, one should also consider a test that tests the normality of the dataset. The test combines the assessment of the different kurtosis coefficients and asymmetry of the distribution [15, 16]:

- there is no minimum number of samples, but for smaller sets ($n<30$), the chance of the second type of error significantly increases;
- accuracy of the determination increases with the number of samples;
- combines the estimation of the discrepancy of the coefficients of kurtosis and asymmetry of two sets;
- test checks whether the empirical distribution is theoretically normal;
- test ignores the skew of the distribution.

3. MATERIALS AND METHODS

The focus of the investigation is a system designed to predict the likelihood of a particular employee's termination within a specified time in the company. Many possible independent variables characterize the system; as an illustration,

the model utilizes approximately 200 features, including some generated ones. Since this problem should be considered a Time-Series problem, when creating interactive reports, checking data for validity, presence and searching for outliers and anomalies, attention should be paid to trends and seasonality. For example, it is logical to assume that there is a trend towards increasing salaries in a particular group of occupational categories, from now on called Job Family Group. So, it could be assumed that the JobFamilyGroup with the value "Architecture" will grow by an average of n% over the last four months, and the JobFamilyGroup "Quality Management" will fall by an average of k% over the same four months.

To get a competent anomaly assessment that meets the nuances of the needs and specifics of the data, need to choose the correct method for each type of data. Given that no single method would satisfy all the needs and metrics and, at the same time, take into account the specifics of the data, need to approach each situation competently. For example, searching for anomalies by excluding only the upper and lower boundary values is simply impractical for multimodal data. The multimodal data from a mathematical statistics perspective is an overlay of several other distributions and dependencies. So, finding anomalies in each component distribution, the assessment would be more accurate. From the machine learning side, the model rarely uses a feature alone, but instead tries to combine it with others. Anomalies within this distribution distort the model's view of the variable and confuse it instead of focusing on its statistical characteristics. Given the subject area, the WageGross column does not have the same law for all employees but depends most on its ManagementLevel and JobFamilyGroup. Hence, searching for anomalies in multimodal data depending on subcategories will be more interpretative. For the model and for the researchers, it can help eliminate anomalous records and to correct them for the data owners. Following these examples, the ANOVA+Percentile filtering method was used to find anomalies in multimodal data, using the WageGross column as an example.

The following methods were used to find anomalies: ANOVA, Percentile Filtering, Isolation Forest, Median+IQR, Chebyshev's Inequality, Rule-Based. Such a set of methods and their combinations is necessary for a competent assessment of anomalies, considering the different types and features of the data. For example, ANOVA + Percentile Filtering searches for anomalies in multimodal data into subcategories. The search by simple sifting is quite adequate, sufficient and interpretative. For example, easily explain to the data owner that this employee's salary is anomalously high or low, given his JobFamilyGroup and ManagementLevel. A more obvious example is Rule-Based filtering, when the given data is Boolean, searching for anomalies by excluding impossible situations based on business logic provides a clear and correct result. For Data Drift detection, a combination of statistical tests for different types of distributions was used, along with methods for separating multimodal distributions. The following statistical tests that were used: Kolmogorov-Smirnov Two-Sample Test, Mann-Whitney U Rank Test, Z-Test, Normal-Test. The preliminary steps to detect Data Drift were:

- Join the SQL database with a query;
- read the parameters from the YAML file, which are necessary for the pre-processing itself and the execution of methods or statistical tests;
- filter records from the Employee History and Vacancy History tables so that their date column is within the time limits from the parameters (starting date is 2020 due to COVID-19 pandemic);
- download an Excel file with filters that indicate the needed data;
- detect anomalies in WageGross, MonthOnSalary, APM columns and VacancyHistory table;
- remove or correct, if possible, previously detected anomalies.

4. RESULTS

Considering the non-normal distribution in which the median value differs from the mean, next, take the APM column as an example. The APM column specifies the profit percent, and company losses from employees' salaries. This distribution "out of the box" is not normal. The smallest value of the column is about minus 200 million percent, and the largest is 150%. In that case, see where the distribution's density is more significant and forms a "more normal" distribution (Fig. 6).

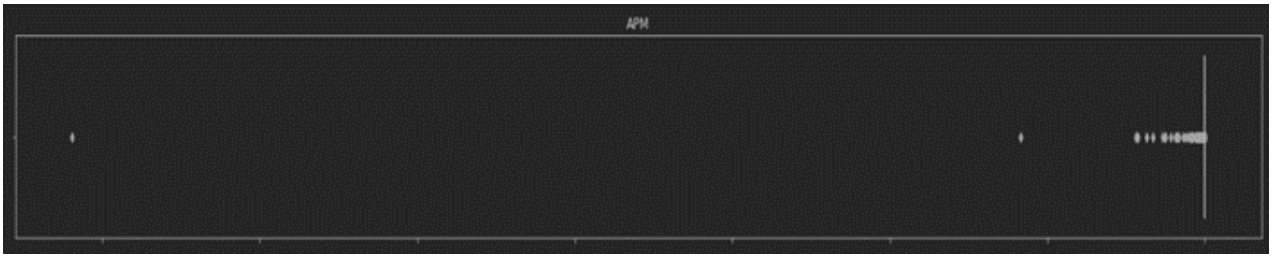

FIGURE 6. - Boxplot of the APM column without excluding extreme values (grey point determines anomalies)

Even though these values are quite rare against the background of the total number of samples, they are absolutely tremendous. Therefore, any method that uses the average value to one degree or another is an entirely inappropriate exercise. For instance, imagine a situation where a new batch of data was received after a certain month, with additional anomalies in the APM column. Although the number of such anomalies is small, their value is still tremendous against the background of the distribution; for example, out-of-the-box samples with an APM value of -200 million percent or more distorted the average value. The value of the distribution within the boxplot ranges from -50 to 150 (Fig. 7).

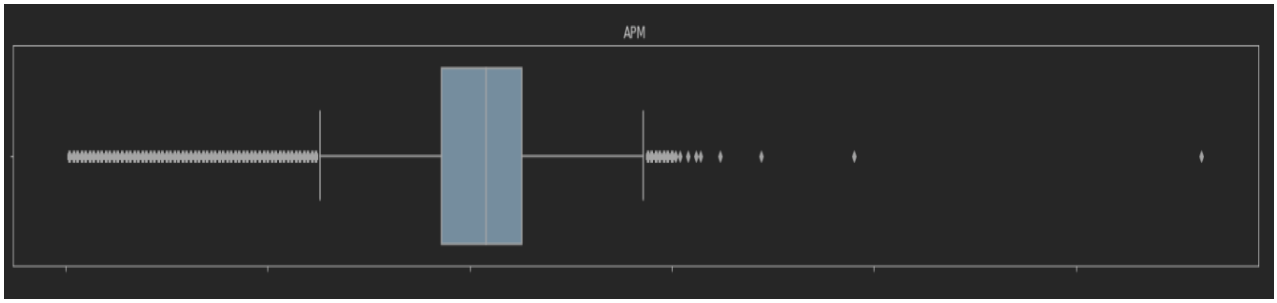

FIGURE 7. - Boxplot of the APM column scaled to boxplot's whiskers (area within determines the distribution)

The mean will change significantly depending on the minority of anomalous data, which is undesirable. It is therefore desirable to use a more stable metric that characterises the distribution and is representative in the given situation. The Median+IQR+iForest method aims to determine the percentage of contamination of the data frame according to the APM column and pass the contamination percentage value to the Isolation Forest method, which accepts this value as a hyperparameter. Thus, eliminate the need to define the contamination percent parameter for the Isolation Forest model, and justify the statistical determination of anomalies, Median+IQR, by the machine learning method, Isolation Forest. Thus, the confidence in choosing exactly anomalous values increases. Fig. 8 illustrates what will be considered an anomaly according to the upper and lower limits, with the parameter k=1.5.

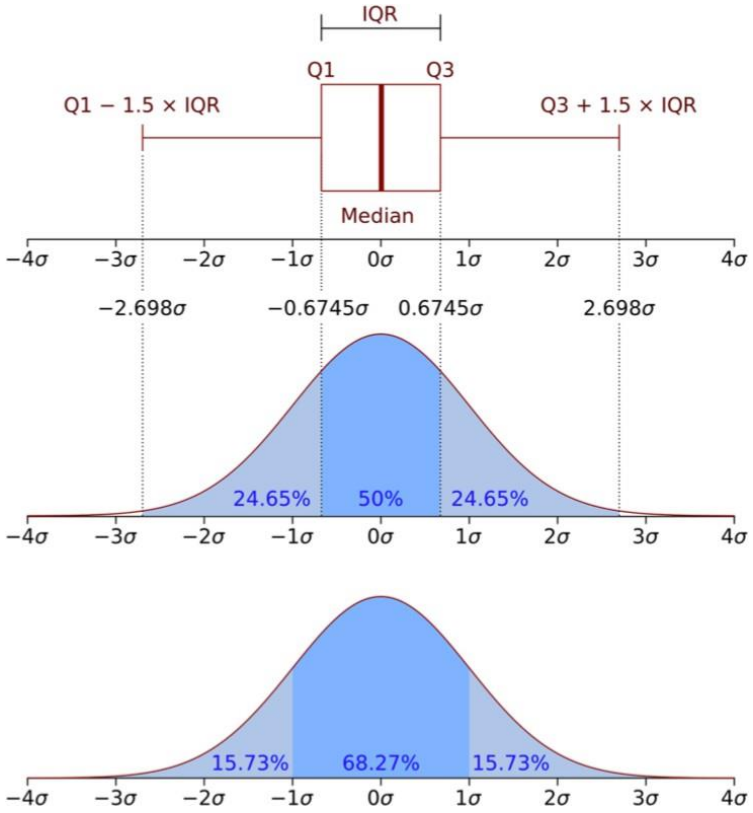

FIGURE 8. - Building boxplot by MEDIAN+IQR method, with set value *k*=1.5 (q3+1.5*IQR)

The parameter k=1.5 is considered standard, but the value "3" was increased. In the work "Why "1.5" in IQR method of outlier detection?" (2019), it was determined that the value k=1.5 was considered normal for a normal

distribution, which was connected to the relationship of the density distribution with the number of standard deviations (STD). To confirm the statement, a QQ-Plot for the normal and the APM column distribution should be outputted (Fig. 9).

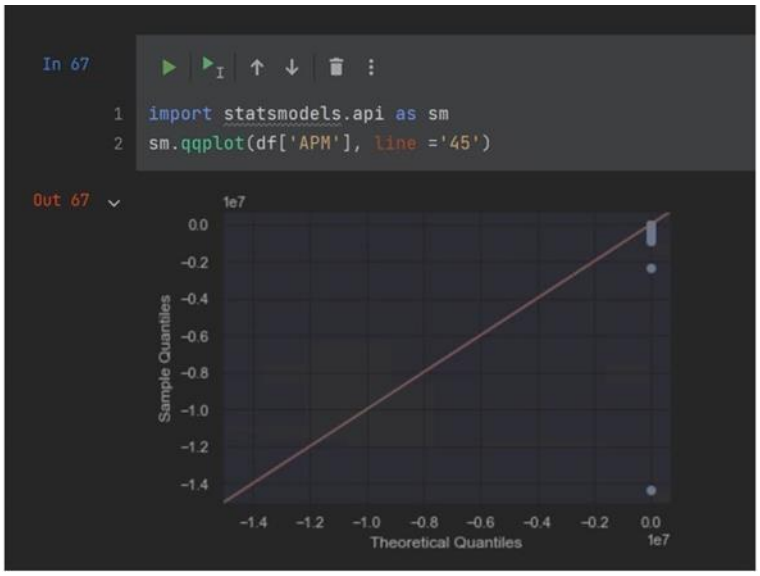

FIGURE 9. - Checking empirical distribution with out-of-box APM column values for normal distribution

Fig. 9 shows that the out-of-box data does not correspond to a normal distribution. Also, it was noticed that the already mentioned obvious anomalies with enormous values. Therefore, the empirical distribution out-of-box does not fall under the definition of the normal theoretical distribution. Considering a single test to check the inconsistency of the empirical distribution with the theoretical one is enough; any number of tests would not grant consistency between the two, but only of the hypothesis probability [17]. Considering the QQ-Plot, it is clear that cleaned from anomalies data does not correspond to a normal distribution either (Fig. 10).

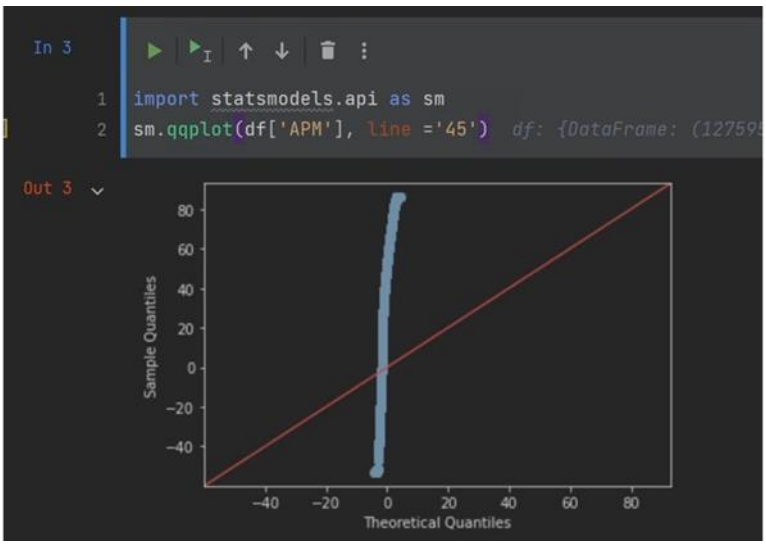

FIGURE 10. - Checking empirical distribution with cleaned APM column values for normal distribution

The question arises whether correct the application of the method is by just changing the parameter *k*. The answer is affirmative, based on the connection of the Chebyshev's limit for big numbers and the connection with the central limit theorem. The Chebyshev's law of large numbers talks about the relationship between the density of the distribution and the deviation, the sigma parameter. This theorem works not only for normal distributions but also for non-normal ones, however the relationship of density-deviations from the mean decreases [11]. So, if for a normal distribution, in order to determine the main bunch of the distribution, it is needed the 3-sigma rule; for non-normal distribution – 6-sigma. For specifics, given a table of confidence and density of the selected distribution to the chosen sigma parameter for non-normal distribution (Fig. 11).

k	Min. % within *k* standard deviations of mean	Max. % beyond *k* standard deviations from mean
1	0%	100%
$\sqrt{2}$	50%	50%
1.5	55.56%	44.44%
2	75%	25%
3	88.8889%	11.1111%
4	93.75%	6.25%
5	96%	4%
6	97.2222%	2.7778%
7	97.9592%	2.0408%
8	98.4375%	1.5625%
9	98.7654%	1.2346%
10	99%	1%

FIGURE 11. - Relation of selected samples to *k* standard deviations

Recalling the central limit theorem, combined the idea of increasing the parameter *k* with MEDIAN+IQR to increase the confidence for non-normal distribution with the Chebyshev's limit method and the parameter *k* in it. It was indicated how the algorithm works with the selected value of *k*=3 (Fig. 12).

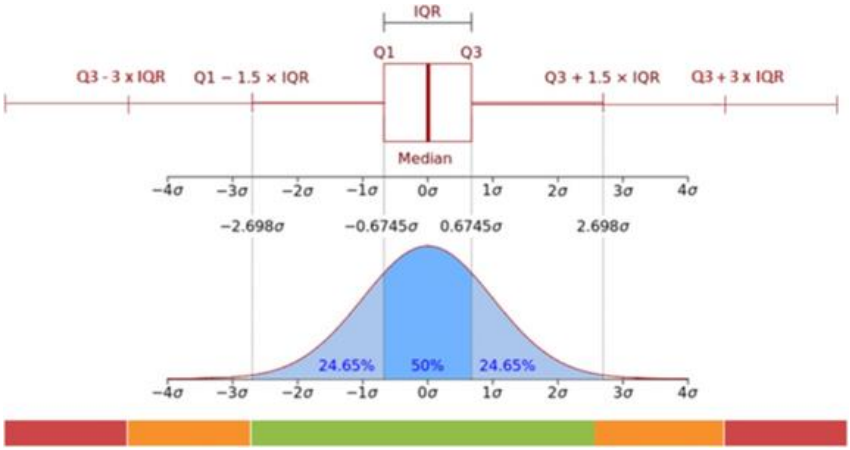

FIGURE 12. - Building boxplot by MEDIAN+IQR method, with set value *k*=3 (q3+3*IQR)

In conclusion, the contamination percentage was calculated according to the MEDIAN+IQR method and used as the value of a corresponding parameter in the Isolation Forest method. A flowchart of the algorithm for finding anomalies in the APM column was presented on Fig. 13.

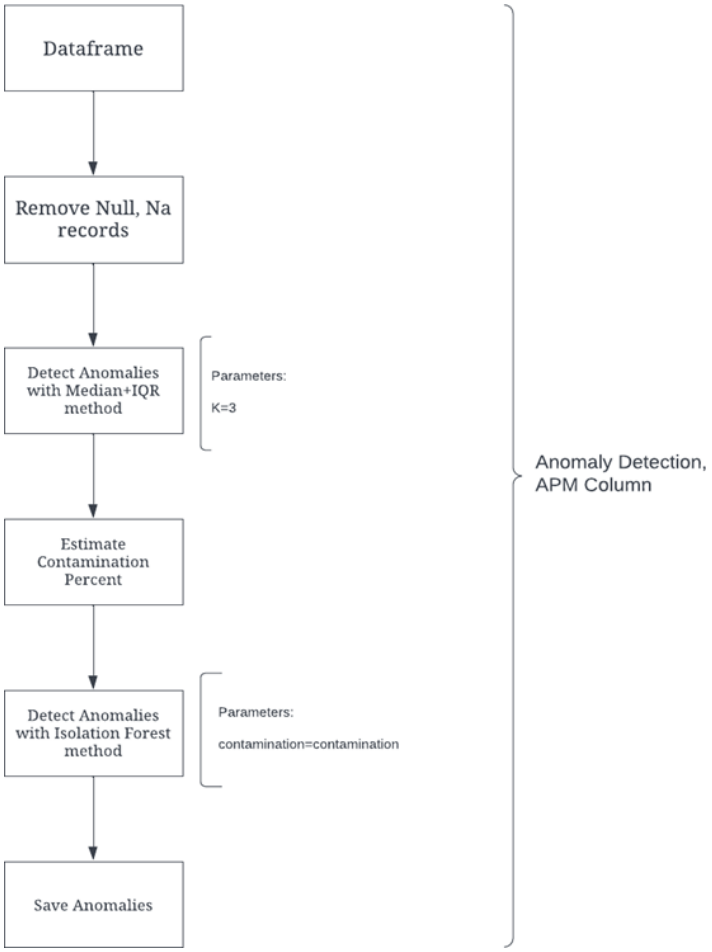

FIGURE 13. - Anomaly detection flowchart for APM column using MEDIAN+IQR+IFOREST method

Consider finding anomalies for a non-normal distribution where the mean is representative; therefore, logically, it allows for using methods that use the mean. Next, analyse the search for anomalies in such a distribution using the MonthOnSalary column as an example. The MonthOnSalary column is responsible for the number of months the employee spent on the same salary. When the salary changes, the value of the column becomes one. The column corresponds to the incremental data type; hence the hypothesis about the data abnormality can arise. Checking it for normality, then visually, it can be seen that the distribution does not look like a normal one at all. Since the distribution is not normal, there are no radical outliers, the median and average values are close to each other and fall into the distribution [18]. The MonthOnSalary column is responsible for the number of months the employee spent on the same salary. Between the visual comparison, next, check this distribution with the help of QQ-Plot (Fig. 14).

FIGURE 14. - Checking empirical distribution with MonthOnSalary column values for normal distribution

The application of the Chebyshev's boundary method on normal and non-normal distributions was compared. Firstly, compare the statistical number of distribution samples, which considered abnormal for normal and non-normal distributions, with the identical value of k=3 (Fig. 15).

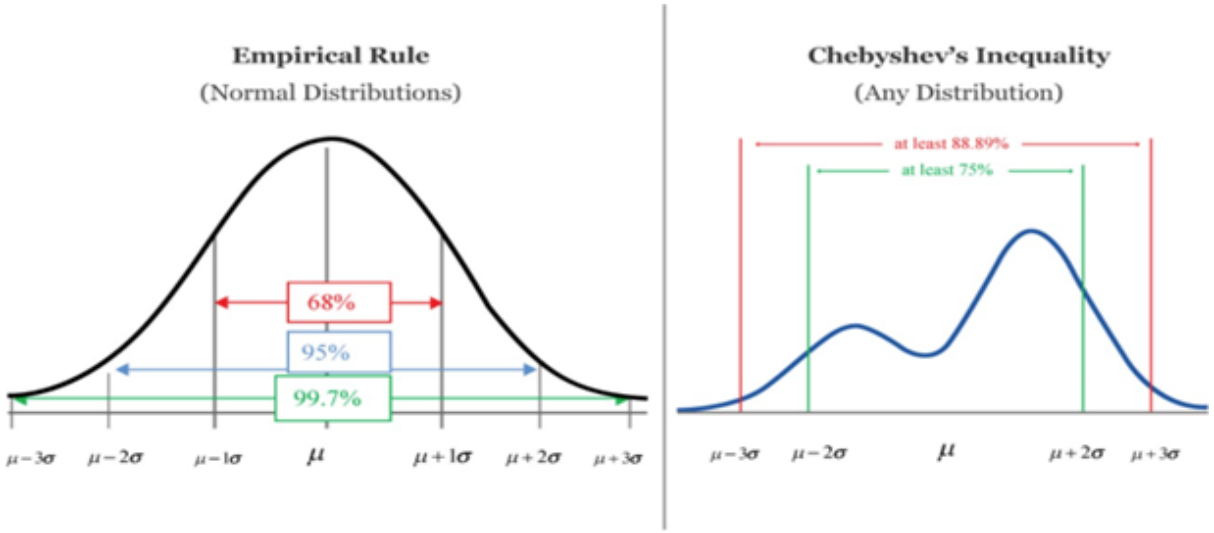

FIGURE 15. - Anomaly detection on normal and non-normal distributions using the Chebyshev's boundary method

The percentage of such samples is less for non-normal one; therefore, it is needed to increase the k parameter to one corresponding to the number of anomalies of other methods used and the number of anomalies in the normal distribution.

Consider the anomaly detection process for a logical data type using the OnSite column as an example. The column determines whether the employee's office corresponds to the "OnSite" office. It accepts a value of zero or one. As mentioned in the introduction, the presence of anomalies in the data type is checked by the so-called "Rule-Based" algorithm [19]. Rule-Based algorithm checks for response parameters in the data, but without using statistical or machine learning methods. To do this, it is needed to clear records from NaN and Null values and then check the data relevance (Table 1).

Table 1. - Description of the OnSite column parameter values

Column	Value	Description	Condition
OnSite	1	The office assigned to the employee corresponds to the "OnSite" office	The variable "Country" of the record is equal to the value "Ukraine"
	0	The office assigned to the employee	The variable "Country" of the record is not equal

	does not corresponds to the “OnSite” office	to the value “Ukraine”

From the Table 1, it can be seen that anomalies will be all columns that do not meet the correct rule. With the help of the NumPy Python package, selected anomalies using a logical operation (Fig. 16).

```
outliers_df = df[
    np.logical_or(
        np.logical_and(
            df['Country'] == 'UKRAINE',
            df[col] == 1
        ),
        np.logical_and(
            df['Country'] != 'UKRAINE',
            df[col] == 0
        )
    )
]
```

FIGURE 16. - Anomaly detection using a logical operation for logical data type (using On Site column as an example)

A flowchart for finding anomalies for a logical data type was also presented using the example of the “On Site” column (Fig. 17)

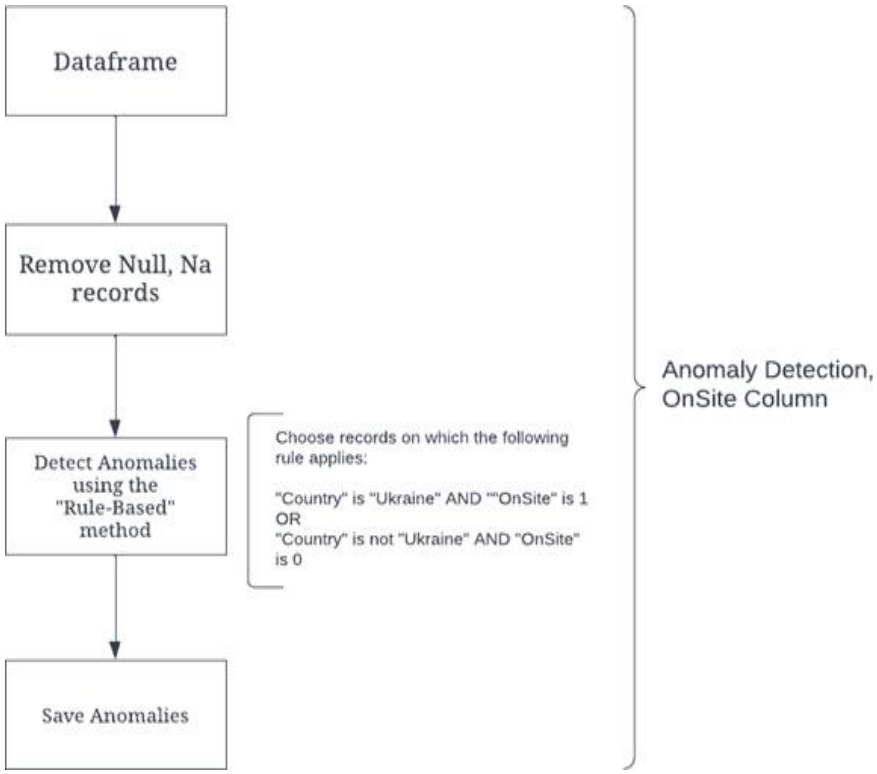

FIGURE 17. - Anomaly detection for a logical data flowchart using the “Rule-Based” method (using “On Site” column as an example)

Consider the search for anomalies for a complex logical data type using the Vacancy History table as an example. The Vacancy History table is a separate table in the database that the model can later use. All three columns can acquire two states – 1 and 0 (Table 2).

Table 2. - Description of Vacancy History table parameter values

Column	Value	Description
Opened	1	Vacancy is open
	0	Vacancy has not been opened yet
Closed	1	Vacancy is closed
	0	Vacancy is not closed yet
Cancelled	1	Vacancy is cancelled
	0	Vacancy is not cancelled

Based on business logic, a vacancy cannot be in a state where all three columns are equal to one. To search for anomalies under such a condition, used the logical operator from the NumPy package. Fig. 18 below shows the flowchart for using this operator.

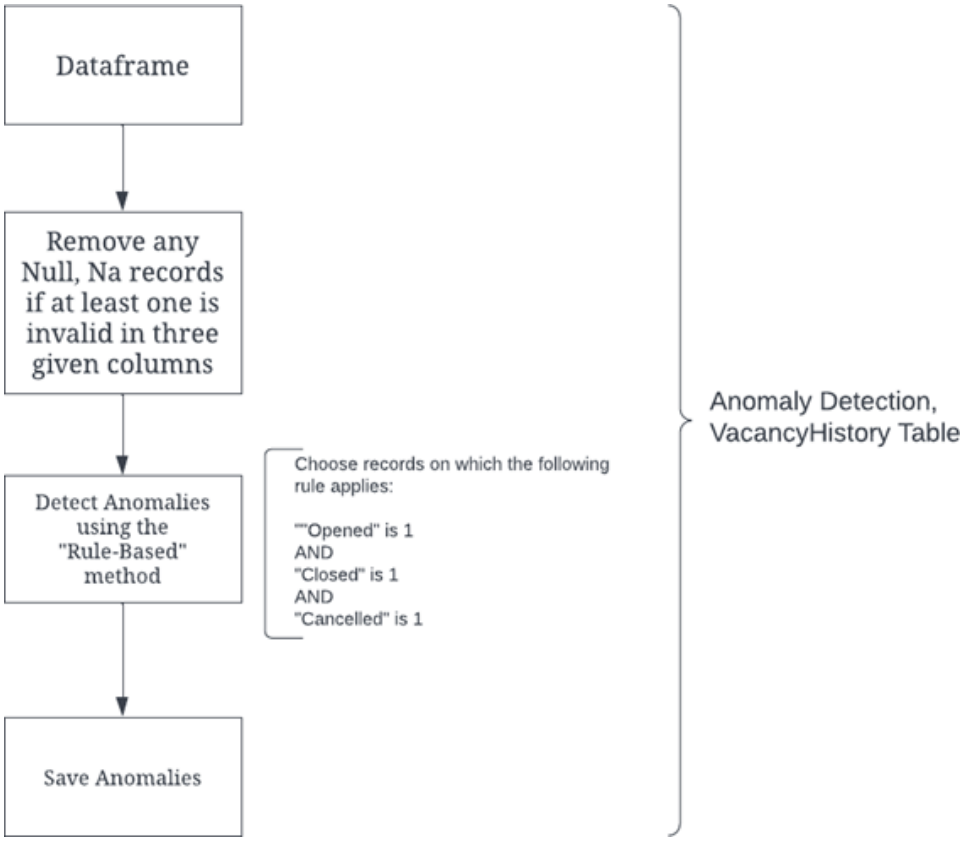

FIGURE 18. - Anomaly detection for a complex logical data flowchart using the "Rule-Based" method (using "Vacancy History" column as an example)

Data Drift is testing the hypothesis of whether a data distribution of new data corresponds to a past data distribution that did not include this data (Fig. 19).

FIGURE 19. - Illustration of Time-Series line

Fig. 19 shows that the historical data is before 2021-11-01, and the new (production) data is after. The situation is classic for Time-Series systems that work with data that is constantly updated and supplemented over time. The task is to check whether the new data's properties correspond to the historical ones' properties. To do this, it is needed to check for the presence of a common theoretical distribution, the historical distribution, and the new (production) data distribution. The Data Drift will be detected if the equation "P(X)Historical Data! = P(X)New (Production) Data" is true. There are several ways to test the equality of two distributions, but considered statistical methods for a multimodal distribution using the WageGross column as an example. As a reminder, the WageGross column has a multimodal distribution, which should be divided into groups, firstly by the category ManagementLevel and then by JobFamilyGroup.

To successfully apply statistical tests, it is necessary to meet their performance requirements. For instance, the Z-Test requires a normal distribution, the Kolmogorov-Smirnov Two-Sample Test requires the absence of anomalies, and the Pearson Normality Test, like all others, requires a minimum number of samples for comparison. The initial step involves removing any anomalies from the data frame, such as Null, NaN, and 0 values, as identified in the previous section. Next, the data frame should be divided into groups based on ManagementLevel, and then further divided by Job Family Group within each ManagementLevel group. Finally, it should be verified whether the historical sample size is sufficient for the Pearson Normal Test to be conducted on each group. The dataset's sample size was compared to historical data to identify any discrepancies. If the number of samples falls below the threshold, the Data Drift test result for the group is recorded as having insufficient data (see Fig. 20).

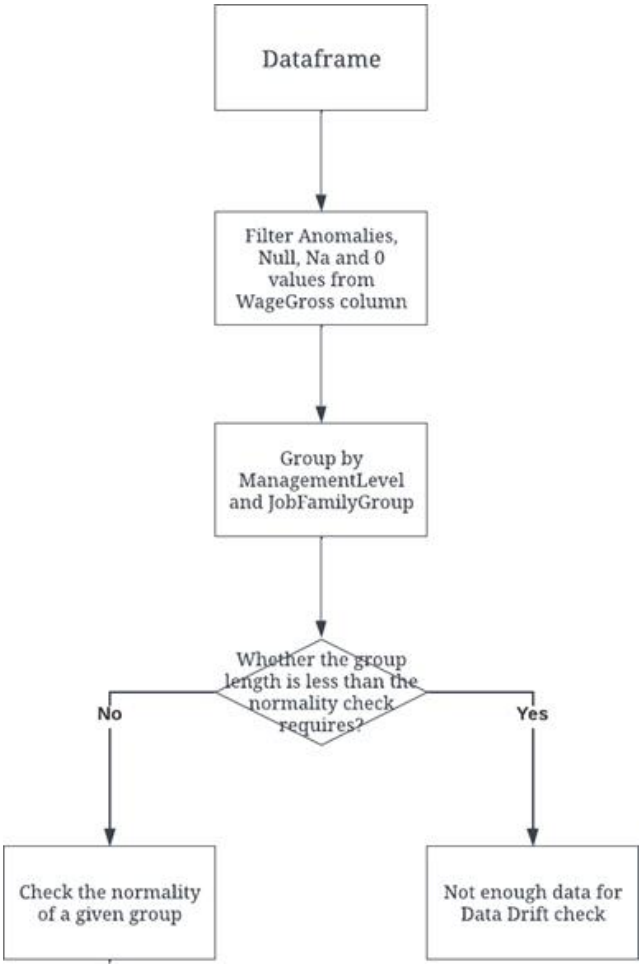

FIGURE 20. - Flowchart of first three steps of Data Drift detection for multimodal data on an example of the WageGross column

If the number of samples for a group is greater than the minimum number of samples for Pearson's normality test, performed a group normality test; see documentation of the method in the SciPy package. The test will return a P-Value from which determined whether the distribution of the group is normal. If the value is less than the threshold, rejected the hypothesis that the distribution of this group is normal. Otherwise, if the value is greater than the threshold, then cannot be rejected the hypothesis about the normality of the distribution of this group (Fig. 21).

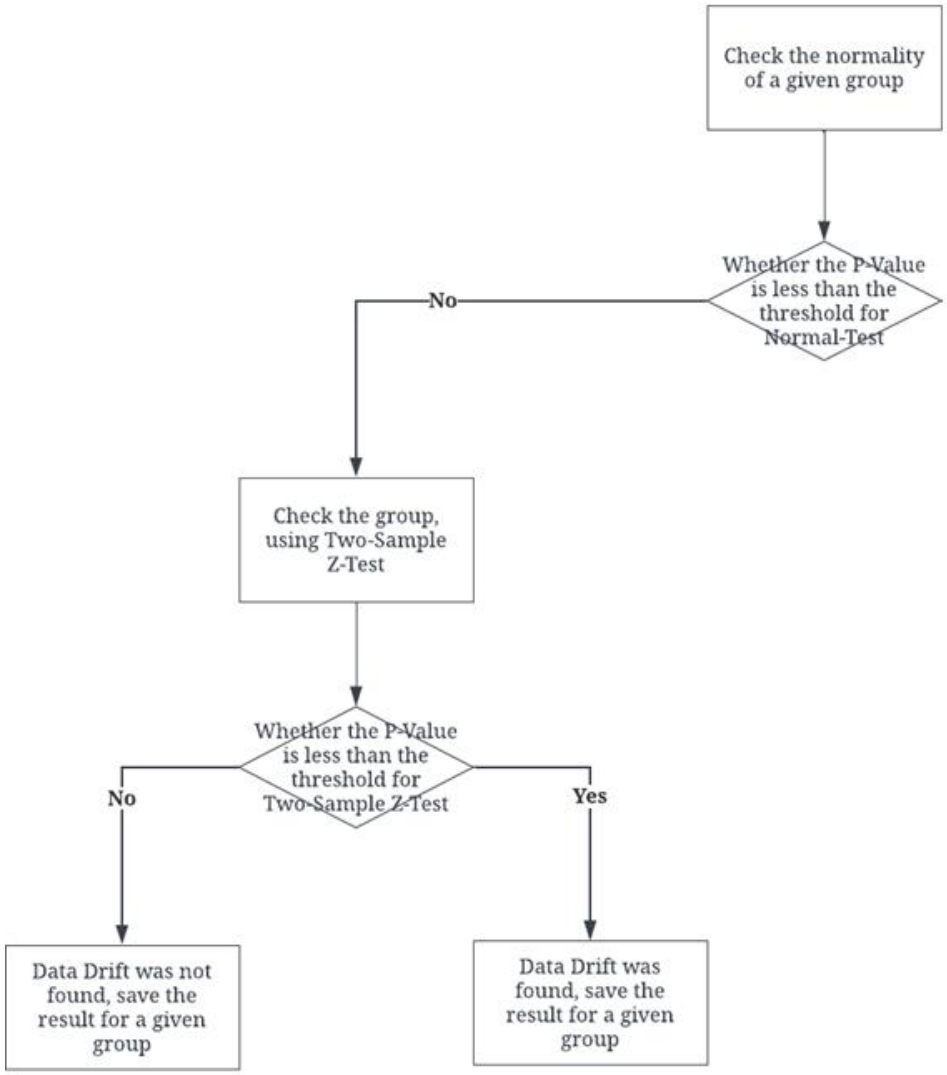

FIGURE 21. - Flowchart of next three steps of Data Drift detection for multimodal data on an example of the WageGross column (case when the group is assumed to be normal)

The distribution of the group is normal. In that case, checked the equality of the historical data and new data distributions using the Z-Test. If the P-Value found is greater than the threshold specified for the Z-Test, then recorded the absence of Data Drift in the result of this group. If the P-Value found is less than the threshold, then can be rejected the hypothesis about the equality of these distributions. Therefore, the presence of Data Drift can be recorded in the result for this group. Next case is that the group distribution is non-normal. In that matter, it is needed to check whether the number of new samples of the group is smaller than the Kolmogorov-Smirnov Two-Sample Test can handle since its threshold is higher than for the normality test mentioned earlier. If the number is insufficient, the Mann-Whitney U Rank Test should be used (Fig. 22).

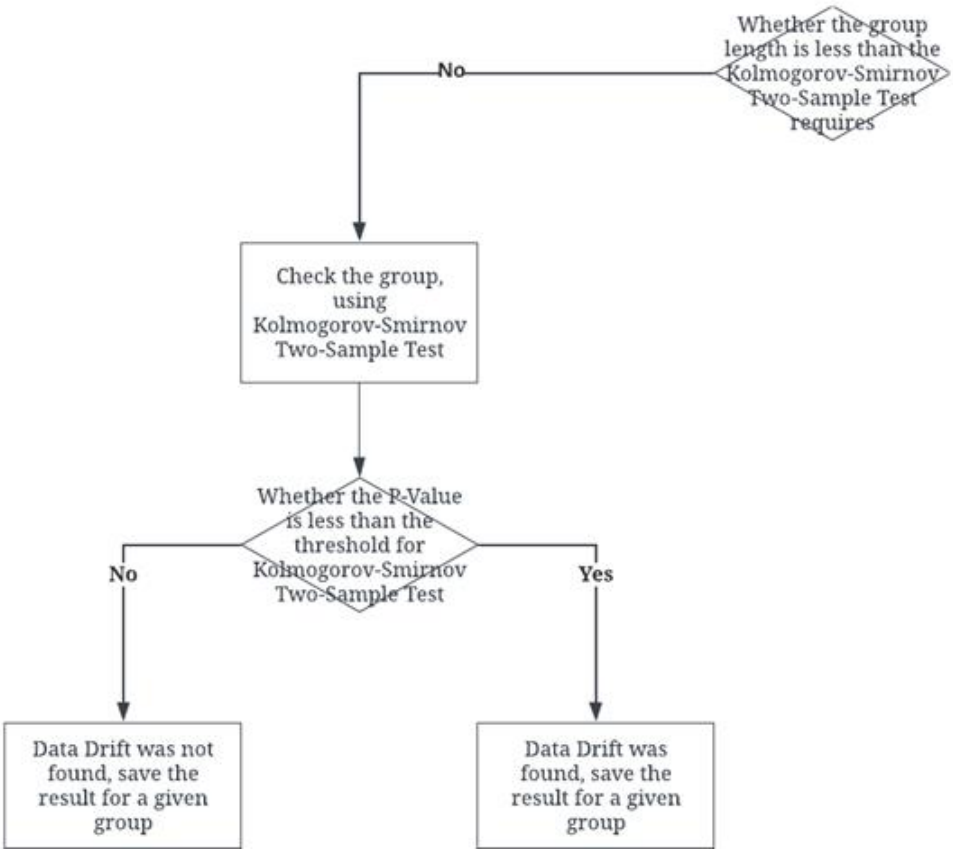

FIGURE 22. - Flowchart of Data Drift detection for multimodal data on an example of the WageGross column (case when the group is assumed to be non-normal, and the test samples is sufficient for a Kolmogorov-Smirnov Two-Sample Test)

Consider the case when there are enough samples for the Kolmogorov-Smirnov Two-Sample Test. Then, check the equality of the historical and new distributions. If the P-Value found is greater than the threshold specified for the Kolmogorov-Smirnov Two-Sample Test, then record the absence of Data Drift in the result of this group. If the P-Value found is less than the threshold, then reject the hypothesis about the equality of these distributions. Therefore, record the presence of Data Drift in the result for this group (Fig. 23).

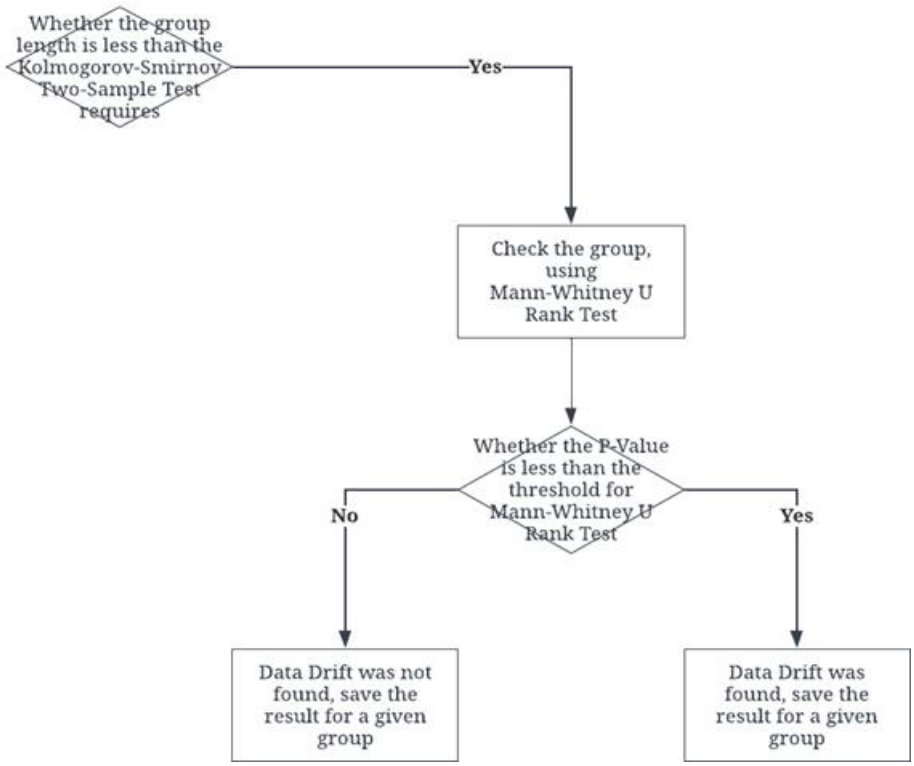

FIGURE 23. - Flowchart of Data Drift detection for multimodal data on an example of the WageGross column (case when the group is assumed to be non-normal, and the test samples is not sufficient for a Kolmogorov-Smirnov Two-Sample Test)

Consider the opposite case, when the number of historical samples in the group is less than the threshold for the Kolmogorov test, the Mann-Whitney U Rank Test was used. If the P-Value found is greater than the threshold specified for the Mann-Whitney U Rank Test, then recorded the absence of Data Drift in the result of this group. If the P-Value found is less than the threshold, then can be rejected the hypothesis about the equality of these distributions. Therefore, can be recorded the presence of Data Drift in the result for this group. Consider numerical Data Drift detection for numerical data with non-normal distribution using APM column as an example. As a first step, clean the data frame from false values and anomalies. It would be best also to create the data frame from Null and NaN values. The second step is to check whether the number of historical samples in the data frame is less than the sample threshold in the already mentioned Kolmogorov-Smirnov Two-Sample Test. The historical and new distributions were checked for equality in the third step using the Kolmogorov-Smirnov Two-Sample Test. If the P-Value found is greater than the threshold specified for the Kolmogorov-Smirnov Two-Sample Test, then recorded the absence of Data Drift in the result of this group. If the P-Value found is less than the threshold, then the hypothesis about the equality of these distributions can be rejected. Therefore, the presence of Data Drift can be recorded in the result for this group (Fig. 24).

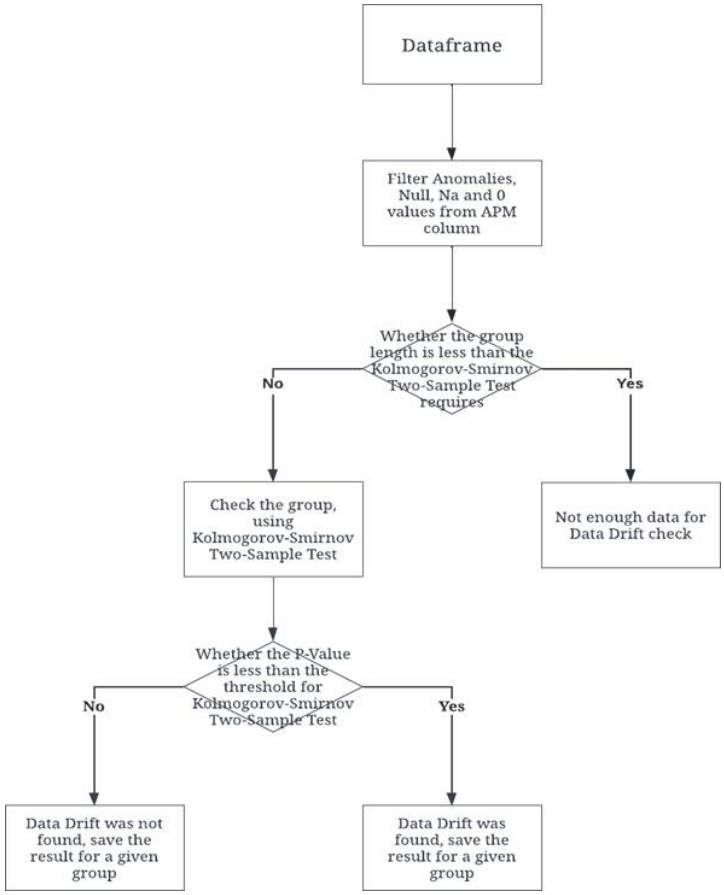

FIGURE 24. - Flowchart of Data Drift detection for numerical data with non-normal distribution on an example of the APM column

All of the results and parameters were given in Tables 3-8.

Table 3. - Parameters for finding anomalies

Method	Column	Parameter	Value
Median+IQR+iForest	APM	“iqr_k”	3
Median+IQR+iForest	APM	“n_estimators”	100
Median+IQR+iForest	APM	“max_samples”	“auto”
Chebyshev’s Inequality	MonthOnSalary	*k*	6
Chebyshev’s Inequality	MonthOnSalary	“ignore_lower_whisker”	True
ANOVA+PERCENTILE	WageGross	“quantile”	0.0015

Table 4. - Received metrics during anomaly detection

Method	Column	Metric	Datatype	Value
Median+IQR+iForest	APM	relative.detected_records	Non-normal distribution, in which the median differs from the mean	0.036
Chebyshev’s Inequality	MonthOnSalary	relative.detected_records	Non-normal distribution, in which the median is approximate to the mean	0.003
Rule-Based	OnSite	relative.detected_records	Logical	1.137e-5
Rule-Based (complex)	VacancyHistory	relative.detected_records	Complex logical (table)	0
ANOVA+PERCENTILE	WageGross	relative.detected_records	Multimodal	0.003

Table 5. - Parameters for determining Data Drift

Column	Parameter	Value
-	testing_date	2021-11-01
-	end_date	2021-12-01
-	start_date	2020-01-01
WageGross	normal_test_p_value_threshold	2.5E-01
WageGross	z_test_p_value_threshold	1E-03
WageGross	ks_test_p_value_threshold	1E-02
WageGross	mw_test_p_value_threshold	3E-02
WageGross	normal_test_p_value_threshold	1E-02
WageGross	normal_test_min_sample_size	30
WageGross	ks_test_min_sample_size	100
APM	ks_test_p_value_threshold	1E-03
APM	ks_test_min_sample_size	100

Table 6. - Received metrics for determining Data Drift

Test	Whether Data Drift is detected?	JobFamilyGroup	ManagementLevel	P-Value	Relative number of samples
KSTest	FALSE	Application Engineering	2 – Intermediate/Middle Associate	0.575987	0.141142
KSTest	FALSE	Quality Management	1 – Junior/Junior Associate	0.592674	0.054797
ZTest	FALSE	Quality Management	6 – Principal/Associate Director	0.624688	0.000654
MWTest	FALSE	Engineering and Technology – Platforms	3 – Senior/Senior Associate	0.634278	0.000588
ZTest	FALSE	Asset Management	1 – Junior/Junior Associate	0.63502	0.001973

Table 7. - Top five samples with lowest P-Value

Test	Whether Data Drift is detected?	JobFamilyGroup	ManagementLevel	P-Value	Relative number of samples
KSTest	FALSE	Business Analysis JFG	5 – Expert/Senior Manager	1.0	0.003092
KSTest	FALSE	Engineering and Technology – Data Science	3 – Senior/Senior Associate	0.999999	0.001451
KSTest	FALSE	Finance	2 – Intermediate/Middle Associate	0.999999	0.008309
KSTest	FALSE	Technical Communication JFG	2 – Intermediate/Middle Associate	0.99999	0.002352
KSTest	FALSE	Engineering and Technology – DevOps	4 – Tech Lead/Lead Associate	0.999937	0.001271

Table 8. - Received metrics for the APM column

Test	Column	Whether Data Drift is detected?	P-Value
KSTest	APM	FALSE	0.496153

As expected, the most significant number of anomalies is in the APM column. Many outliers with extreme values outside the boxplot distort the average in the APM column. It was adhered to the idea of looking for only the most obvious anomalies; since the average number of outliers is relatively small, the interval of the relative amount is 0.036-0; if multiplied by 100, this value is 3.6%-0%, where 3.6% are anomalies in the APM column. Speaking about the Data Drift, which was carried out only in the section of one month of new data, in comparison with historical ones, it is clear that the smallest P-Value belongs to the most volatile group (Table 9).

Table 9. - Data Drift detection for APM column

Test	Column	Whether Data Drift is detected?	P-Value	testing_date
KSTest	APM	FALSE	0.496153	2021-11-01
KSTest	APM	FALSE	0.11606	2021-10-01
KSTest	APM	FALSE	0.015622	2021-09-01
KSTest	APM	FALSE	0.010455	2021-08-01
KSTest	APM	FALSE	0.029831	2021-07-01
KSTest	APM	FALSE	0.042292	2021-06-01

Regarding the detection of Data Drift for the APM column, it is clear that the P-Value is lower than for the test with a cut of one month, so this column is quite volatile over time. This statement was confirmed by increasing the testing period by a couple of months. This volatility influenced the decision to lower the P-Value threshold for testing the APM column (Fig. 25).

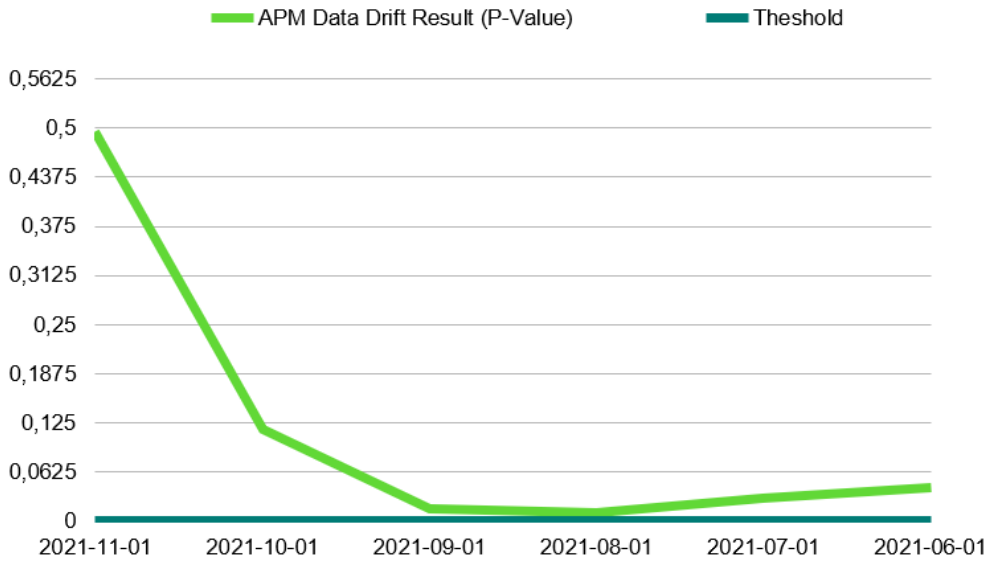

FIGURE 25. - Relation in an increasing the number of testing months to the P-Value

In this study, an automated data quality management pipeline was developed for data acquisition, pre-processing, anomaly detection, data drift analysis, and reporting. For anomaly detection, a blend of statistical and machine-learning methods was used. The Median+IQR method identified 3.6% anomalies in the non-normal APM column, while Chebyshev's Inequality found 0.3% anomalies in the incremental MonthOnSalary column. Categorical data such as OnSite were examined using rule-based techniques, and complex data in VacancyHistory was analysed using Boolean logic. The multimodal WageGross data was segmented and filtered, detecting 0.3% anomalies. Data drift analysis assessed distribution shifts using the Kolmogorov-Smirnov test, Mann-Whitney, or Z-tests, depending on the data distribution. The pipeline's modular architecture allows for continuous monitoring, feedback, optimization, and integration with retention models, enhancing data quality and providing insightful reports.

5. DISCUSSION

Research on the detection of anomalies and changes in time series data is important for the development of processing and structuring of large volumes of data. Applying methods to detect anomalies and changes in time series data has helped to improve the quality of data analysis, identify potential problems or deviations in real time and take timely measures to solve them. This study, conducted for the development of a system of automatic analysis and data processing, made it possible to understand in more detail the problems of implementing various methods for detecting anomalies, which made it possible to determine the most effective approaches to detecting anomalies, to develop optimal algorithms and models, as well as to determine the parameters necessary for their debugging. Research conducted in recent years in the field of detection of anomalies and deviations in real time has brought significant progress in the application of various methods, technologies and models of machine learning. These studies allowed not only to evaluate the effectiveness and reliability of different approaches, but also to identify their advantages and limitations. In addition, research has revealed the advantages and limitations of different machine learning models. For

example, some models may be more effective at detecting anomalies with certain data characteristics, while others may be more flexible and adaptive to changes in the data.

According to the results of the study by A. Blázquez-García et al. [20], it turned out that there was a wide range of methods for detecting anomalies in time series, which varied from classical statistical methods to more complex algorithms of machine learning and deep learning. Their advantages and limitations in use were indicated, a comparative analysis of their effectiveness was carried out, and the main challenges related to the detection of anomalies in time series were identified, such as large volumes of data, non-stationarity, noise and other factors affecting the accuracy of the results. The similarity of this study with what was described above lies in the fact that they consider different methods of detecting anomalies in time series, analyse their advantages and limitations, and also examine the effectiveness of using each of the methods. The application of various machine learning technologies and models, such as classification algorithms, clustering, neural networks, and deep learning, has made it possible to detect anomalies and deviations in real time with high accuracy. Some of the anomaly detection methods are based on comparing samples with normal behaviour, others use classification algorithms to highlight abnormal samples, and other methods use neural networks and deep learning to automatically detect complex dependencies and anomalies in data.

In the work of G. Fenza et al. [21] that devoted to the methodology of detecting anomalies in the smart power grid, presented the main steps of a methodology that takes into account changes in data for detecting anomalies in the smart grid, including building a data model, detecting changes and anomalies based on comparing current data with normal patterns or patterns. The problems of detecting anomalies in the system were also considered, considering changes in the time series of data that occur due to changes in the system structure, functional dependencies or external factors. The analysis of this work will make it possible to understand the main principles and approaches used in the methodology, which will be useful in understanding the use of this method to detect anomalies. Also, the analysis of the research will allow for evaluating the accuracy, sensitivity and specificity of the method, as well as for finding out its advantages and limitations, which is important for understanding the extent to which this approach can be applied in real situations in the smart grid. The study also revealed the advantages and limitations of different methods of detecting anomalies and deviations. Some methods may be more efficient for certain types of data or specific scenarios, while others may be less efficient or require more computing power. Understanding these advantages and limitations will help to choose the most optimal methods for specific problems of detecting anomalies and deviations.

M. Jain et al. [18] in their work about the method of detecting anomalies in networks based on the hybrid technique of detecting changes in concepts, combining K- Means algorithms clustering and the method of support vectors, presented a new method for detecting anomalies. The results of the study showed that the proposed hybrid method effectively detected anomalies in network data and made it possible to note conceptual changes in real time. Compared to other methods of anomaly detection, the proposed method demonstrates high accuracy and reliability. The main advantage of this method is the combination of two powerful approaches, such as K- Means clustering and the method of support vectors, to detect anomalies in the data. Starting with building a model based on K- Means clustering, which helps to identify normal patterns in the data, the method allows for creating a basic model that reflects the normal state of the data. Then, with the help of the support vector method, classification of new data samples is carried out, which allows detecting anomalies. The hybrid method is efficient to use because it provides better accuracy and reliability of anomaly detection due to the combination of the advantages of both methods. This method provides a more comprehensive and understandable approach to detecting anomalies in data.

In their work on the method of detecting anomalies in time series using stacked residuals of understandable components, L. Zancato et al. [22] proposed the STRIC methodology (Stacked Residuals of Interpretable Components), which is based on the decomposition of the time series into interpretable components such as trend, seasonality, cycles and noise. The results showed that this method is distinguished by its ability to interpret the results and allows for detecting complex anomalies in time series, which will make it a valuable tool for data analysis. The main advantage of this method is its interpretability, which makes it possible to understand the nature of the detected anomalies and simplifies the decision making regarding further analysis. This can be useful in the context of making decisions about further analysis and action. Understanding the nature of anomalies helps to establish the reasons for their occurrence and decide what steps should be taken for further control or correction. Interpreted components can also provide contextual information to help refine the analysis and make it more accurate and targeted to the needs of a particular situation or system. In general, the STRIC method can be applied in various fields where it is important to detect anomalies in time series, such as finance, industry, medicine and many others. Taking into account these results, approaches to detecting anomalies and deviations in real time can be more reasonably chosen for further development, depending on the specific context and characteristics of the data. In addition, these studies are the basis for further development of methods and technologies in this field, leading to improved quality of data analysis and providing more effective detection of anomalies in real time.

Thus, research on the detection of anomalies and changes in time series data is an important stage in the development of modem data analytics, which allows for ensuring high-quality and reliable data analysis, identifying potential problems and taking timely measures to solve them. The continued use of anomaly and change detection methods in time series data has the potential to improve the efficiency and accuracy of data analysis, provide predictive

problem detection, and help take the necessary actions to resolve them. Given the constant change of conditions and data, automatic anomaly detection systems become a necessary tool for effective management and monitoring of various processes and systems.

6. CONCLUSIONS

This research can be used in the data control pipeline, ensuring continuous Data Quality for the machine learning model in the Time-Series system. The data tends to change when it is essential to monitor the continuous quality of the data. Also, the research can be used in the Time-Series system, for which the automation of processes and minimal manual intervention of data scientists is vital in the future. When the performance of a machine learning model decreases due to the number of outliers, or a change in trend due to Data Drift, this solution will warn and correct anomalies. After completing this work, a data inspection pipeline that performs an automated anomaly detection and Data Drift detection phase was built. Despite the complexity of applying different methods, considering different types of data, and distributions may be the most important criterion for the success of such a pipeline is independence.

In the future, the considered study will permit the visualization of each step executed by the data validation pipeline, enabling other developers to observe the result of its work without being aware of the nuances of its implementation and saving time. It is necessary to standardize the solution using MlFlow and Docker. The following research will be the operationalization that will make it possible to use this pipeline for its intended purpose regardless of the human factor and manual interventions of data scientists to adjust, manually validate the data or run this pipeline from the development environment and instead use web visualization.

Funding

None

ACKNOWLEDGEMENT

The study was created within the framework of the project financed by the National Research Fund of Ukraine, registered No. 30/0103 from 01.05.2023, "Methods and means of researching markers of ageing and their influence on post-ageing effects for prolonging the working period", which is carried out at the Department of Artificial Intelligence Systems of the Institute of Computer Sciences and Information Technologies of the Lviv Polytechnic National University.

CONFLICTS OF INTEREST

None

REFERENCES

[1] A. Iturria, J. Labaien, S. Charramendieta, A. Lojo, J. Del Ser and F. Herrera, “A framework for adapting online prediction algorithms to outlier detection over time series,” Knowledge-Based Systems, vol. 256, 109823, 2022.

[2] P. Kumari and M. Saini, “An adaptive framework for anomaly detection in time-series audio-visual data,” IEEE Access, vol. 10, pp. 36188-36199, 2022.

[3] C. Hegde. “Anomaly detection in time series data using data-centric AI,” in *IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, 2022, pp. 1-6.

[4] X. X. Yin, Y. Miao and Y. Zhang, “Time series based data explorer and stream analysis for anomaly prediction,” Wireless Communications and Mobile Computing, 5885904, 2022. https://doi.org/10.1155/2022/5885904

[5] D. Sulem, M. Donini, M. B. Zafar, F. X. Aubet, J. Gasthaus, T. Januschowski, S. Das, K. Kenthapadi and C. Archambeau, “Diverse counterfactual explanations for anomaly detection in time series,” ArXiv, 2022. https://doi.org/10.48550/arXiv.2203.11103

[6] Why “1.5” in IQR method of outlier detection? accessed 18 May 2022, https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097

[7] A. Gelman, “Analysis of variance – Why it is more important than ever,” The Annals of Statistics, vol. 33, no. 1, pp. 1-53, 2005. https://doi.org/10.1214/009053604000001048

[8] Outlier detection with Isolation Forest, accessed 18 May 2022, https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

[9] Data Drift report, accessed 5 September 2023, https://docs.evidentlyai.com/presets/data-drift#data-drift-report

[10] M. Naaman, “On the tight constant in the multivariate Dvoretzky-Kiefer-Wolfowitz inequality,” Statistics & Probability Letters, vol. 173, 109088, 2021. https://doi.org/10.1016/j.spl.2021.109088

[11] Testing fit to distributions, accessed 18 May 2022, http://www.rguha.net/writing/notes/stats/node11.html
[12] H. B. Mann and D. R. Whitney, "On a Test of whether one of two random variables are stochastically larger than the other," The Annals of Mathematical Statistics, vol. 18, no. 1, pp. 50-60, 1947.
[13] R. C. Sprinthall, *Basic statistical analysis*. Boston: Allyn & Bacon, 2011.
[14] G. Casella and R. L. Berger, "Statistical inference," Biometrics, vol. 49, no. 1, pp. 320-321, 1993. https://doi.org/10.2307/2532634
[15] K. F. R. S. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 50, no. 302, pp. 157-175, 1900. https://doi.org/10.1080/14786440009463897
[16] R. B. D'Agostino, "An omnibus test of normality for moderate and large size samples," Biometrika, vol. 58, no. 2, pp. 341-348, 1971. https://doi.org/10.2307/2334522
[17] Hypothesis testing: Fear no more, accessed 19 May 2022, https://www.isixsigma.com/tools-templates/hypothesis-testing/hypothesis-testing-fear-no-more/
[18] M. Jain, G. Kaur and V. Saxena, "A K-Means clustering and SVM based hybrid concept drift detection technique for network anomaly detection," Expert Systems with Applications, vol. 193, 116510, 2022. https://doi.org/10.1016/j.eswa.2022.116510
[19] The empirical rule and Chebyshev's theorem, accessed 18 May 2022, https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/02%3A_Descriptive_Statistics/2.05%3A_The_Empirical_Rule_and_Chebyshev's_Theorem
[20] A. Blázquez-García, A. Conde, U. Mori, J. A. Lozano, "A review on outlier/anomaly detection in time series data," ACM Computing Surveys, vol. 54, no. 3, pp. 1-33, 2021. https://doi.org/10.1145/3444690
[21] G. Fenza, M. Gallo and V. Loia, "Drift-aware methodology for anomaly detection in smart grid," IEEE Access, vol. 7, pp. 9645-9657, 2019. https://doi.org/10.1109/ACCESS.2019.2891315
[22] L. Zancato, A. Achille, G. Paolini, A. Chiuso and S. Soatto, "STRIC: Stacked residuals of interpretable components for time series anomaly detection," accessed 18 May 2022, https://openreview.net/forum?id=VnurXbqxr0B