Research Article

Data Warehouse, Detection and Transfer of Anomalies in Retail Data

Onur CIRKIN

[1] Harran University, Gtech, Orcid ID: https://orcid.org/0000-0002-6061-2338, crkn.onur@gmail.com.tr.

Reference: Cirkin, O. Data Warehouse, Detection and Transfer of Anomalies in Retail Data. The European Journal of Research and Development,3(2), 46-53.

Abstract

In this article, we offer some suggestions for anomaly detection on the data received from the source to the Data warehouse. As a result, it is aimed to prevent the entry of dirty and noisy data into the data warehouse. We think that knowing that there is clean and healthy data in the data warehouse will be resistant to anomalies in the processed data used for data science. In order to reach our goal, studies were carried out on the data in the retail sector. We aimed to determine our theoretical thoughts from some topics such as user erroneous login data in the retail and energy industry, abnormal sales over employees during the campaign period, product stock abnormality, and incorrect pricing. When we examined many studies, we saw that they made anomaly detection after estimation. Before taking the data from the source to the data warehouse, we thought that anomaly detection would be more efficient and healthier. Analysis and results were evaluated on the data obtained in the wiseboard retail project of Gtech company.

Keywords: *DataWarehouse, Anomaly Detection, LSTM, One-class SVM, Isolation Forest*

1. Introduction

Data Warehouse is used to archive and analyze data from sales, customer segmentation, products, salaries, or other daily operations by creating an independent database from an organization's live database. It is crucial for the Data Warehouse to operate efficiently. Our main tasks include providing decision-makers with fast and up-to-date access to reports. However, performance issues can arise with the Data Warehouse. The ETL process ensures the loading and updating of data into the Data Warehouse. After going through the ETL (Extract, Transform, Load) process, the data is transferred to the Data Warehouse. Once the data transfer is complete, the most suitable methods for processing the data in the optimal format should be considered, as they affect performance criteria. Working with complex queries, triggering procedures, and

ensuring the timeliness of business intelligence can pose various challenges in big data. We aim to provide a recommendation system for methods such as parallel processing and partitioning to address these challenges. Proper partitioning of data is crucial because incorrect partitioning can lead to lower performance. We can achieve parallelism in queries by distributing data to machine nodes. Distributing data across machines to work in a distributed manner is an operation aimed at improving performance. However, providing parallelism without considering the machine's capacity will increase the cost instead of improving query performance. The purpose of our research thesis is to identify anomalies based on the source data integrated into the Data Warehouse. Creating alert systems based on these findings will enable us to establish the Data Warehouse without abnormal situations. Many studies indicate that anomaly detection is often based on predictive data. In this aspect, anomaly detection can be performed before data processing. We continue our research and investigations for the retail sector.

2. Materials and Methods

We applied our proposed approaches to the data obtained from fashion retail and real data. The results obtained have demonstrated that these methods work well on both types of data [1]. In retail, especially in fashion retail, supply chain optimization is crucial for cost control, improving customer satisfaction, managing inventory, and ultimately increasing profits [15].

We focused on the important issues in data warehousing supply. The detection of anomalies in valuable data and ensuring the health of the data used for prediction are essential for performance. To find answers to these issues, we conducted work using two algorithms. In this section, we will discuss the proposed algorithms for anomaly detection, which include Long Short-Term Memory (LSTM), One-Class Support Vector Machines, and Isolation Forest algorithms.
The distribution of the data by date is as follows in figure 1:

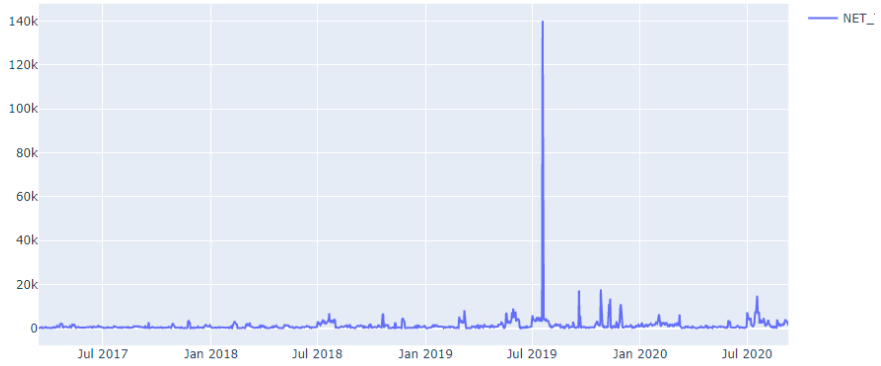

Figure-1- The distribution of the data by date

2.1.LSTM (Long Short-Term Memory)

LSTM (Long Short-Term Memory) is an artificial recurrent neural network (RNN) architecture widely used in deep learning. It consists of three control gates:

The Forget Gate
The Input Gate
The Output Gate

Recurrent neural networks have a loop structure. Each gate has a sigmoid neural network and dot product formulas. The sigmoid layer ensures that the input values, which represent a sequence of input vectors x = {x1, x2, ..., xt, ...}, are scaled between [0, 1] to control the flow of information. When used as an RNN for time series data, LSTM reads a sequence of input vectors x = {x1, x2, ..., xt, ...}, where xt ∈ ℝm represents an m-dimensional input vector for the variables at that specific time step. It decides which past data should be forgotten using the following formula:

$$ft = \sigma(Wf .[ht-1, xt] + bf), (1)$$

Here, Wf and bf are the weight matrix and bias of the forget gate. Then, LSTM operates on the cell state. It determines which information should be retained. The input is then processed through a tanh function to bring the values within the range of (-1, 1) and the resulting values are element-wise multiplied. The calculated It value, which represents the input gate value, and the C̃t value simultaneously generated by the tanh layer, are used to compute the Ct value that will be updated, as follows:

$$it = \sigma(Wi .[ht-1, xt] + bi), (2)$$

$$C̃t = \tanh(Wc .[ht-1, xt] + bc), (3)$$

and

$$Ct = ft * Ct-1 + it * C̃t, (4)$$

(Wi, bi) and (Wc, bc) are the weight matrices and biases controlling the input gate and the memory state, respectively. After passing through the tanh function, the output gate is calculated as follows:

$$ot = \sigma(Wo .[ht-1, xt] + bo), (5)$$

$$ht = ot * \tanh(Ct). (6)$$

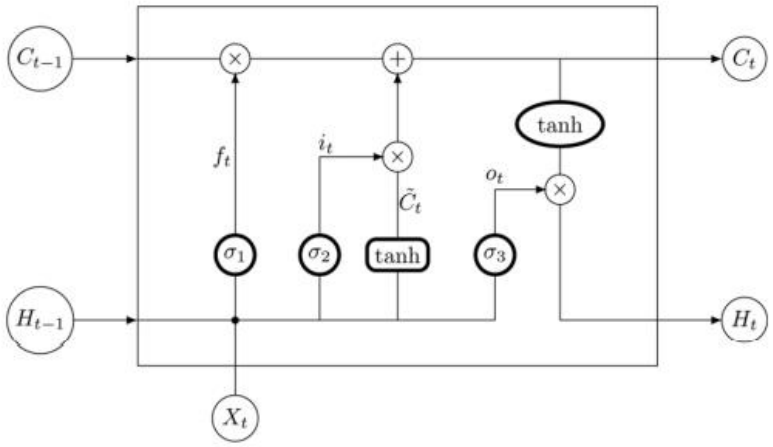

Figure-2- The Model

Different variations of LSTM have been used to solve various problems [16]. However, we will test our original algorithm for this specific case.

2.2. One-class Destek Vektör Makinesi

One-Class Support Vector Machine (OCSVM) is a machine learning algorithm. It aims to capture the support of the distribution.

It draws a linear line to separate the data points placed within a plane. The plane contains two different classes, and the decision on which class future data points will belong to is made at this stage. The biggest advantage of OCSVM compared to other algorithms is its minimal probability of misclassification.

The data is transformed into a feature space using a kernel function employed by the algorithm. The plane that separates the two classes is defined to be as far away from the center as possible. This plane is analyzed using the formula provided below.

$$\min\left(\frac{1}{2}\|w\|^2+\frac{1}{vl}\sum_{i=1}^{l}\xi_i-\rho\right)$$

$$\left(w.\Phi(x_i)\right)\geq\rho-\xi_i \quad i=1,2,...,l \quad \xi_i\geq 0$$

Here, w and ϱ are the parameters of the hyperplane, Φ is the kernel function, v is the allowed rate of misclassified values (outliers), l is the number of objects in the training set, and ξ is the error parameter.

The outputs of the data set are calculated as follows.

$$f(x)=\operatorname{sgn}\left(w.\Phi(x)-\rho\right)$$

2.3. Isolation Forest

It is based on decision tree algorithms. It is a powerful algorithm for detecting outliers in data. It produces two results in anomaly detection. Firstly, it generates a categorical label indicating whether the observation is anomalous or not. Secondly, it produces a score or confidence value. In this study, we observed the result generated by the confidence value, specifically the "-1" values.

3. Results

Isolation Forest and LSTM algorithms were used for anomaly detection based on net earnings by months. The clearest example of this is the observed differences in amounts, even on consecutive days without any discounts. One of the most important factors in anomaly detection can be observed through anomaly detection through prediction and anomaly detection through classification.

The results we obtained from the Isolation Forest algorithm are as follows: a confidence interval is derived for each net amount. We observe that the values we detected as -1 indicate abnormal results.

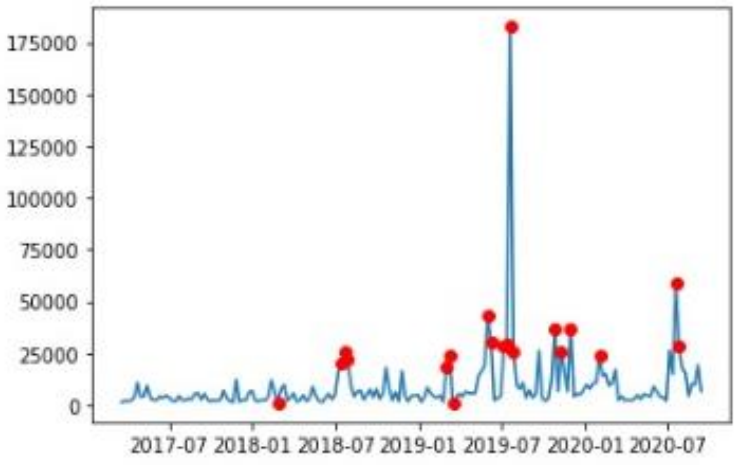

Figure-3- The graph of results

The output of the results we detected using the LSTM algorithm is as follows. Similarly, a prediction model has been developed based on the Net Amount. Abnormal values reaching the output gate through the sigmoid function have been observed as the output results for us.

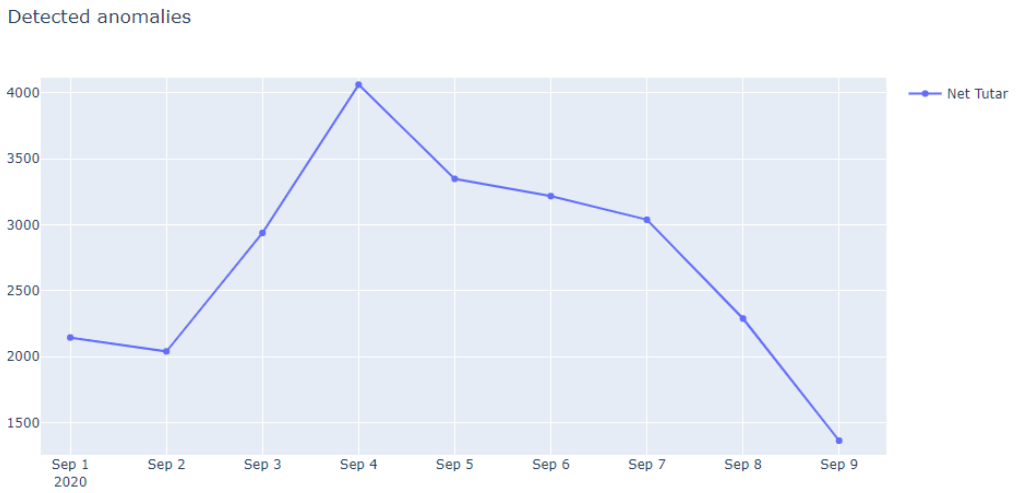

Figure-4- The graph of detected anomalies

4. Discussion and Conclusion

As a result, we were able to capture anomalies with our algorithms except for the One-Class SVM. In the initial stage, we observed that the Isolation Forest algorithm is more effective in anomaly detection.

In the anomaly detected by our Isolation Forest algorithm, we observed a discrepancy in prices despite no discounts being applied within a one-day interval. We were able to perform a historical check based on the net amount. This occurred at this stage regarding the content of the data. The ultimate goal is to detect anomalies in the source data on a column-by-column basis. Such cases should be reported and escalated to the managerial level.

In the subsequent process, the algorithms will continue to be tested on different datasets. Based on the results of the outputs, the ultimate objective is to position this study before the data is loaded into the data warehouse. This will ensure that the data warehouse is made as stable as possible, enabling the healthy sharing of prediction models and end-user reports.

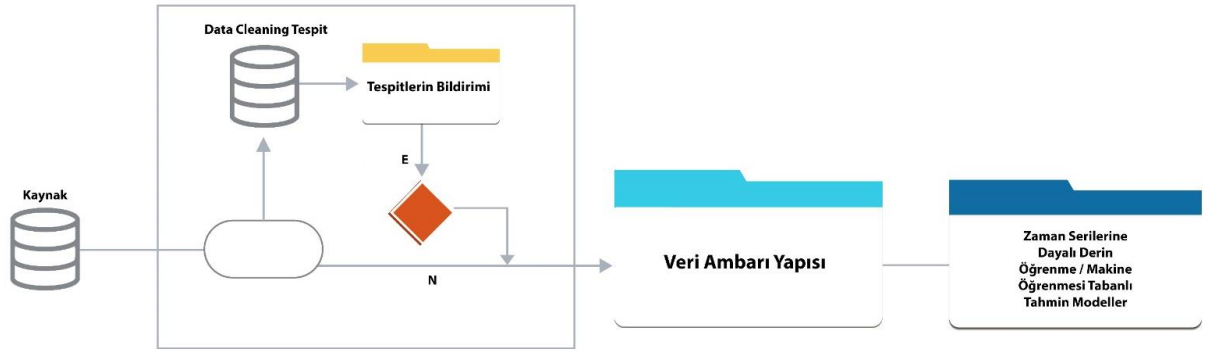

Figure-5- Model

5. Acknowledge

- Successfully transfer data to the Data Warehouse and thereby guide company target predictions with optimum accuracy.
- Prevent user and application errors.
- Contribute to increasing the company's profit margin.
- Prevent unnecessary occupation of physical space on database machines.

6. Thank You

I would like to express my gratitude to Gtech company for their inspiration and support throughout the Wiseboard Retail project.

References

[1] Nguyen, H. D., et al. "Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management." *International Journal of Information Management* 57 (2021): 102282.

[2] Jansen, Maarten, Laurens Swinkels, and Weili Zhou. "Anomalies in the China A-share market." *Pacific-Basin Finance Journal* 68 (2021): 101607.

[3] Hampton, Harrison, and Aoife Foley. "A review of current analytical methods, modelling tools and development frameworks applicable for future retail electricity market design." *Energy* (2022): 124861.

[4] Oliveira, João Pedro, and Rui Dinis Sousa. "Unsupervised Anomaly Detection of Retail Stores Using Predictive Analysis Library on SAP HANA XS Advanced." *Procedia Computer Science* 181 (2021): 882-889.

[5] Ramakrishnan, Jagdish, et al. "Anomaly detection for an e-commerce pricing system." U.S. Patent Application No. 17/721,594.

[6] Chen, Xu, et al. "GraphAD: A Graph Neural Network for Entity-Wise Multivariate Time-Series Anomaly Detection." *arXiv preprint arXiv:2205.11139* (2022).

[7] Vincent, Vercruyssen, Meert Wannes, and Davis Jesse. "Transfer learning for anomaly detection through localized and unsupervised instance selection." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 04. 2020.

[8] Putra, Hafid Yoza. "Fraud detection at self-checkout retail using data mining." *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE, 2020.

[9] Pourhabibi, Tahereh, et al. "Fraud detection: A systematic literature review of graph-based anomaly detection approaches." *Decision Support Systems* 133 (2020): 113303.

[10] Leite, Roger A., et al. "Visual analytics for event detection: Focusing on fraud." *Visual Informatics* 2.4 (2018): 198-212.

[11] Laptev, Nikolay, Saeed Amizadeh, and Ian Flint. "Generic and scalable framework for automated time-series anomaly detection." *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015.

[12] Haldar, Malay, et al. "Applying deep learning to airbnb search." *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019.

[13] Liu, Xiufeng, and Per Sieverts Nielsen. "Regression-based online anomaly detection for smart grid data." *arXiv preprint arXiv:1606.05781* (2016).

[14] Shipmon, Dominique T., et al. "Time series anomaly detection; detection of anomalous drops with limited features and sparse examples in noisy highly periodic data." *arXiv preprint arXiv:1708.03665* (2017).

[15] Thomassey, Sébastien. "Sales forecasting in apparel and fashion industry: A review." *Intelligent fashion forecasting systems: Models and applications* (2014): 9-27.

[16] Greff, Klaus, et al. "LSTM: A search space odyssey." *IEEE transactions on neural networks and learning systems* 28.10 (2016): 2222-2232.