International Journal of Women in Technical Education and Employment
ISSN: 2811-1567. Volume 5 – Issue 1. July 2024
https://fpiwitedjournal.federalpolyilaro.edu.ng

Predictive Analytics for Resource Optimization in Data Warehousing and Data Mining Using Random Forest

Akinbola S. M.* & **Buoye P. A**
Department of Computer Science, The Federal Polytechnic, Ilaro.
*akinbola.serifat@federalpolyilaro.edu.ng

ARTICLE HISTORY

Received: March 19, 2024
Revised: March 22, 2024
Accepted: April 4,2024

Abstract

The present study investigates the utilization of predictive analytics methods, particularly random forest, for the purpose of optimizing resource allocation in data warehousing and data mining settings. Organizations are depending more and more on data warehouses and data mining in the big data era to glean valuable insights from enormous datasets. Optimizing memory allocation, processing power, and storage, on the other hand, is essential to ensuring the efficacy and efficiency of these analytical procedures. In order to enable proactive resource allocation and optimization techniques, this study explores the ability of random forest models to forecast resource use trends based on historical data. As part of the research technique, historical data on resource utilization metrics in data warehousing and data mining contexts are gathered and analyzed. These measurements consist of timestamps and the matching amount of data warehouse storage used. The random forest model showed that it could identify trends in past data, which made it possible to predict future storage needs. The projections of the model function as a decision support system, giving stakeholders practical information for maximizing resource use.

Keywords: data mining, data warehousing, random forest, resource optimization, timestamp

Citation
Akinbola S. M. and Buoye, P. A. (2024). Predictive Analytics for Resource Optimization in Data Warehousing and Data Mining Using Random Forest. *International Journal of Women in Technical Education and Employment*, 5(1), 33-39

Introduction

Organizations in this ever-changing field of information technology are faced with an unparalleled amount of data coming from a variety of sources. Data warehousing and data mining are two related disciplines that are evolving as a result of the strategic necessity of navigating this sea of information. The symbiotic relationship between data warehousing and data mining emerges as a beacon, leading the transformation of raw data into useful knowledge, as organizations strive to extract meaningful insights rather than just collect data.

The onset of the digital era has brought to an exponential increase in data generation. Organizations face both opportunities and challenges due to the vast amount and diversity of information available, which ranges from sensor data and market trends to customer transactions and social media interactions. In this sea of data, the capacity to strategically use information has emerged as a key difference. Once merely an afterthought, data has evolved into a strategic asset that, with the right management, can open up a world of opportunities. The idea of data warehousing is fundamental to efficient data management. A data warehouse is a centralized location that gathers, arranges, and keeps data from many sources (Lauren, 2024). It is the cornerstone of analytical processing (atlan.com, 2022). In addition to making, it easier to retrieve historical data, this organized storage gives organizations a centralized platform from which to

extract insights that can guide decision-making. The science of data mining enhances the framework of data warehousing. Data mining is the process of identifying patterns, correlations, and trends within large databases by looking behind the surface. Its goal is to convert unprocessed data into meaningful insights that may be used to support predictive analysis and lay the groundwork for well-informed decision-making.

The real transformational power appears when data warehousing and data mining work together harmoniously. While data warehousing creates the organized framework for data retrieval and storage, data mining takes the lead, exploring the depths of the data to find undiscovered information treasures. Organizations are able to predict future scenarios and comprehend historical trends because to this dynamic interaction.
For businesses navigating the complexities of a competitive landscape, the marriage of data warehousing and data mining is not merely a technological alliance; it is a strategic imperative. Informed decision-making, customer insights, and operational efficiency become achievable goals as organizations harness the power of their data.

A data warehouse is a specialized repository designed to store and manages large volumes of structured and sometimes unstructured data from various sources within an organization. Unlike traditional databases, data warehouses are optimized for analytical processing, providing a platform for extracting valuable insights through complex queries, reporting, and data mining.

The primary purpose of a data warehouse is to offer a centralized and unified view of an organization's data, providing decision-makers with a comprehensive understanding of business performance, trends, and historical patterns. By integrating data from multiple sources, a data warehouse enables users to make informed decisions, uncover hidden patterns, and derive actionable intelligence.

Data Mining is the process of uncovering meaningful patterns, correlations, and knowledge from large volumes of data. It involves the application of various algorithms and techniques to analyze datasets, identify trends, and extract actionable insights. The goal is not merely to collect information but to transform raw data into knowledge that can drive informed decision-making and strategic planning.

It is the computational process of using techniques from the fields of artificial intelligence, machine learning, statistics, and database systems to find patterns in massive data sets. Extracting information from a data set and transforming it into a comprehensible structure for later use is the main objective of the data mining process.

In the dynamic realm of information technology, the interplay between data warehousing and data mining has garnered significant attention, representing a pivotal convergence point for extracting actionable insights from the burgeoning sea of data. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information, as defined by Richa et al.(2017). The paper explained the concept of data mining and data warehouse with examples. Gayatri et al. (2018) concluded, in their paper, Analysis on Data Warehousing and Data Mining, that data warehouse is an inner store of a theme leaning, included, time-variant and non-volatile compilation of data from different sources.

Mohmmed (2022) provided evidence of the value of data warehousing and data mining by illustrating how the latter technique might assist decision-makers in reaching better conclusions.

Shuangshuang et al. (2018) concluded that Data warehouse technology provides an effective solution for the development of decision support system (DSS) while studying the application of data warehouse and data mining in fracturing engineering system.

Data warehousing and data mining jointly contribute to the realm of business intelligence. Turban, et.al (2014) underscore their role in enhancing decision support systems, emphasizing the iterative process of extracting, transforming, and mining data to drive strategic decision-making.

Pankaj and Madha (2022) after their paper review to study the Influence of Data Mining and Data Warehouse on Strategic Planning explain the impression, benefits and drawbacks of data mining and warehousing. They then concluded that the major useful feature of data warehouses is that they allow data from many sources to be integrated in one place. Han, et.al (2000) discuss the implementation of data warehousing in healthcare, demonstrating its efficacy in improving decision-making processes and patient outcomes.

The random forest can reduce the variance of regression predictors through bagging while leaving the bias mostly unchanged (Lihua et. al, 2023). He said, predicting with random forest can reduce the variance of regression predictors through bagging while leaving the bias mostly unchanged. Lingiun et .al (2018) said that random forest is a valuable tool for institutional research predictive analytics tasks that is very easy to apply, flexible, and computationally inexpensive.

Practical Implications of Leveraging Predictive Analytics For Resource Optimization

Leveraging predictive analytics for resource optimization holds significant practical implications across various organizational aspects. Through precise resource requirement forecasting grounded in past data and predictive models, enterprises can circumvent over-allocating resources, resulting in financial savings on hardware, software licensing, and cloud services. By enabling proactive resource management using predictive analytics, businesses may resolve resource limitations and bottlenecks before they have an impact on system performance. Organizations may guarantee that key workloads receive sufficient resources, which will increase overall performance and system responsiveness, by optimizing resource allocation.

By allocating resources dynamically based on predictive insights, businesses may adapt in real-time to changing workload patterns and business requirements. Because of its scalability and flexibility, it can quickly adjust to changing needs and maintain ideal performance levels even during times of high demand. Predictive analytics also enhances system reliability and stability by proactively identifying potential capacity constraints and mitigating risks of downtime and performance degradation. By optimizing resource allocation, organizations can minimize the likelihood of service disruptions and ensure consistent service delivery. Efficient workload management is facilitated through predictive analytics, allowing organizations to prioritize and allocate resources based on workload criticality and urgency. This approach optimizes resource utilization and maximizes productivity across the organization.

Methodology

In the dynamic landscape of data-driven enterprises, the efficient management of resources within data warehousing and data mining environments is paramount for sustained success.

With ever-growing information, complex analytical procedures, and a constant need for real-time insights, predictive analytics plays an increasingly important role in resource optimization for unmatched efficiency in businesses. Efficient resource optimization is crucial for modern organizations to manage the massive amounts of data they handle. This includes energy usage, network bandwidth, computing power, and storage, posing a challenge to businesses that must carefully balance cost and performance. Utilizing historical data, statistical algorithms, and machine learning approaches to forecast future resource needs, predictive analytics emerges as a strategic solution.
By adopting a proactive stance, organizations can preemptively allocate resources, optimize performance, and enhance overall operational efficiency. Beyond resource allocation, predictive analytics should be able to contribute to informed decision-making. By forecasting future demands and performance trends, organizations can make strategic decisions to enhance overall system efficiency and meet evolving business objectives.

In the era of big data, scalability is a critical challenge. Predictive analytics provides a scalable solution by offering insights into future resource

needs, allowing organizations to plan for expansions and efficiently manage increasing data volumes. To illustrate anticipating resource requirements in a data warehousing and data mining context, we consider a scenario where we use a dataset containing historical information about data storage usage in a data warehouse. We'll employ a simple machine learning algorithm, specifically random forest, to predict future storage requirements based on historical trends.

The random forest model is used to predict future storage requirements based on the timestamp. The dataset contains timestamps and corresponding storage usage in a data warehouse.

Random forest was used because is a versatile ensemble learning algorithm that can handle complex relationships and interactions between features. It's robust to outliers and noise in the data and can handle both numerical and categorical features. Random Forest often provides high predictive accuracy and is less prone to over-fitting compared to individual decision trees.

Data collection

The data used for the analysis was collected from Kaggle. Kaggle is an online repository for dataset of different types. The dataset consist of files of various sizes and the time of arrival on the computer memory.

Table 1: Size of file and time of storage

id timestamp	size (mb)
0 01/01/2000 00:00:00	105
1 01/01/2000 00:00:01	79
2 01/01/2000 00:00:02	130
3 01/01/2000 00:00:03	168
4 01/01/2000 00:00:04	190
5 01/01/2000 00:00:05	199
6 01/01/2000 00:00:06	169
7 01/01/2000 00:00:07	178
8 01/01/2000 00:00:08	186
9 01/01/2000 00:00:09	105
10 01/01/2000 00:00:10	99
11 01/01/2000 00:00:11	130
12 01/01/2000 00:00:12	165
13 01/01/2000 00:00:13	196
14 01/01/2000 00:00:14	199
15 01/01/2000 00:00:15	169
16 01/01/2000 00:00:16	178

17 01/01/2000 00:00:17	180
18 01/01/2000 00:00:18	133
19 01/01/2000 00:00:19	165
20 01/01/2000 00:00:20	178
21 01/01/2000 00:00:21	137

Result

A Pearson correlation coefficient of 0.2748 suggests a weak positive linear relationship between the timestamp (converted to seconds) and the size (in MB) variables in the provided dataset.

Calculate correlation coefficient

correlation = df['timestamp_seconds'].corr(df['size (mb)'])

print("Pearson Correlation Coefficient:", correlation)

Pearson Correlation Coefficient: 0.2748324489733002

In the correlation coefficient's range of -1 to 1, 0.2748 is more closely aligned with 0 than with 1. This suggests a weak positive linear relationship, which means there is a modest tendency for the size (in MB) to increase along with the timestamp's increase (showing subsequent time points). The strength of this relationship isn't that strong, though.

The provided data must first be preprocessed, divided into features (independent variables) and targets (dependent variables), and then the Random Forest Regression model must be trained. Let's continue writing Python code to manipulate data using the pandas library and create the Random Forest Regression model using the scikit-learn package.

The plot below displays the actual storage usage (black dots) and the predicted storage usage (blue dots). As can be seen in fig. 1 below, the regression indicated that the space used by the files (data) will decrease as the time advances (future time). From top left to bottom right, the regression line sloped downward. Using historical data, the random forest model determines the relationship between the timestamp and storage utilization. Future needs for storage in the data warehouse can be predicted using predicted values. These forecasts can be used by planners to proactively assign storage capacity, guaranteeing peak data warehouse performance even as data quantities increase.

The average squared difference between the predicted and actual values is roughly 1682.47 square megabytes, according to a mean squared error (MSE) of 1682.47.

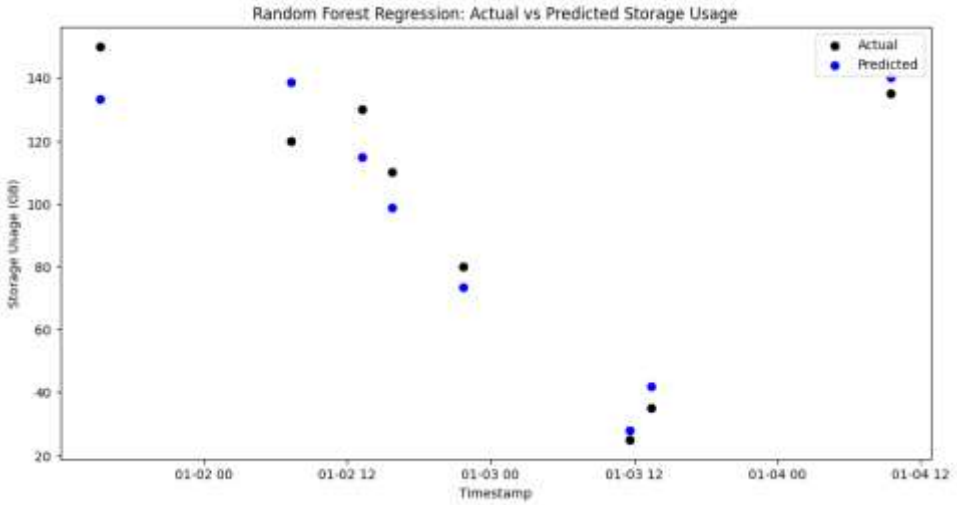

Figure 1: Random Forest Regression showing actual vs predicted storage

Conclusion

Analyzing the time-space dynamics of server storage in a data warehouse with Random Forest Regression is a powerful way to extract insights and maximize resource use. Organizations may efficiently simulate the complex interactions between time and storage capacity inside their server infrastructure by utilizing this potent ensemble learning technique. Because of its exceptional ability to manage the intricate, nonlinear relationships present in time-series data, Random Forest Regression is especially useful for predicting storage needs over time. A comprehensive knowledge of the factors impacting storage consumption, such as temporal trends, seasonal patterns, and potential anomalies, is made possible by its capacity to capture interactions among several predictor variables.

The random forest model demonstrated its capability to learn patterns in historical data, enabling the anticipation of future storage requirements. This proactive stance empowers organizations to allocate resources efficiently, aligning with the anticipated demands of data warehousing and data mining operations.

The model's predictions serve as a decision support system, providing stakeholders with actionable insights for optimizing resource utilization. Decision-makers can leverage these forecasts to make informed choices on storage infrastructure, ensuring scalability and performance as data volumes evolve.

Anticipating resource requirements fosters operational efficiency. By staying ahead of potential bottlenecks and ensuring adequate resource provision, organizations can maintain seamless data processing, uphold system performance, and mitigate the risk of resource-related disruptions.

Reference

Atlan.com (2022). Cloud Data Warehouses: Cornerstone of the Modern Data Stack. Retrieved on 20th of January 2024, from www.atlan.com.

Deshpande, S. P., & Thakare, V. M. (2010). DATA MINING SYSTEM AND APPLICATIONS: A REVIEW. International Journal of Distributed and Parallel Systems (IJDPS), 1(1), 32-44.

Han, J., Pei, J., & Yin, Y. (2000). "Mining frequent patterns without candidate generation." ACM SIGMOD Record.

Kim, Y., & Lee, J. (2002). "Data mining for scientific and engineering applications." Springer.

Kimball, R., & Ross, M. (2002). "The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling." Wiley.

Lauren. B & Stan. H (2024). What is data warehouse: Definition and how it works? Retrieved 20th of January, 2024, from www.ninjaone.com.

Lingiun. H, Richard A. L, Juanjuan .F, Joshua. B & Jenne. S (2018). Random Forest as a Predictive Analytics Alternative to Regression in Institutional Research. *Practical Assessment, Research & Evaluation, 23(1). Available online: http://pareonline.net/getvn.asp?v=23&n=1*

Lihua .C, Prabhashi. W. G & John. R (2023). Debias random forest regression predictors. Journal of Statistical Research 56(2):115-131

Mohmmed, A. (2022). Data Mining and Warehousing. International Journal of Research Publication and Reviews, 3(3), 1396-140.

Pankaj, S., & Madhav, S. S. (2022). Influence of Data Mining and Data Warehouse on Strategic Planning: A Review Paper. International Journal of Innovative Research in Computer Science & Technology (IJIRCST), 10(1).

Richa, P., Lalit, M., Sanjeev, B., & Janmejay, P. (2017). Data Mining and Data Warehouse: sInternational Journal on Emerging Technologies (Special Issue NCETST-2017), 8(1), 155-157.

Shuangshuang, R., Fei, S., Hao, X., Meng, L., & Jun, W. (2018). The application of data warehouse and data mining in fracturing engineering system. 2nd International Symposium on Resource Exploration and Environmental Science. IOP Conf. Series: Earth and Environmental Science, 170, 022080.

Turban, E., Sharda, R., & Delen, D. (2014). "Business Intelligence: A Managerial Perspective on Analytics." Pearson.