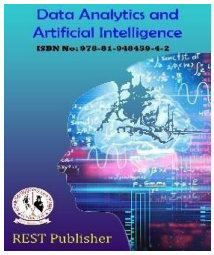
Data Analytics and Artificial Intelligence

Vol: 4(1), 2024

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/

DOI: https://doi.org/10.46632/daai/4/1/8



Machine Learning-Based Assessment of IoT System Performance: A Comparative Study of Linear and Random Forest Regression

Raju Chavan

Principal Architect, Quality Engineering, T-Mobile

Corresponding Author Email: nbrl585210@gmail.com

Abstract: *The rapid growth of Internet of Things (IoT) devices has led to vast data streams that require advanced analytical methods for meaningful interpretation. This study uses machine learning regression models - linear regression (LR) and random forest regression (RFR) - to predict IoT performance scores. This highlights the need for reliable predictive tools that support better decision-making by examining past performance data and predicting future outcomes. This approach involves thorough pre-processing, feature selection, and model building, in which LR provides interpretation and RFR captures complex nonlinear patterns. The dataset includes various performance metrics collected from IoT devices, which are divided into training and testing subsets to ensure robust model validation. The results demonstrate that both models achieve strong predictive accuracy on the training data, with data points clustering closely in the best prediction line. Due to its ability to handle feature interactions and nonlinear relationships inherent in IoT data, the random forest regression model exhibits superior performance. Testing data estimation reveals good generalization capabilities, with predictions maintaining reasonable accuracy on unseen data, although there are expected small deviations from the correct predictions. The comparative analysis highlights the tradeoffs between model complexity and interpretability. While linear regression provides simplicity and transparency in understanding variable relationships, random forest regression provides enhanced predictive power essential for complex IoT scenarios. The study confirms that machine learning regression techniques can effectively support IoT analytics, enabling proactive monitoring, anomaly detection, and performance optimization. These findings have significant implications for IoT system management, resource allocation, and predictive maintenance strategies in smart environments.*

Keywords: Internet of Things, Linear Regression, Random Forest Regression, Performance Prediction, Machine Learning, IoT Analytics

1. INTRODUCTION

The Internet of Things (IoT) has emerged as a major technological revolution of the 21st century, reshaping the way physical devices connect to digital platforms and how people interact with their surroundings. In essence, IoT refers to a vast network of smart devices, sensors, and systems that operate autonomously, collecting, sharing, and processing data over an internet connection, without the need for continuous human input. This technological revolution extends the traditional internet which primarily connected computers and mobile devices to the physical world, enabling everything from home appliances and vehicles to industrial machinery and urban infrastructure to become "smart" and interconnected [1]. The evolution from the "Internet of Computers" to the "Internet of People" through social media and collaborative platforms has now progressed to the "Internet of the Internet," where physical objects are participants in a digital ecosystem. This advancement reflects more than technological advancement; it represents a fundamental shift in how we perceive and interact with our surroundings [2]. The vision of IoT is to move from "anywhere, anytime" to "anything, anytime," connecting billions of devices to their environment, creating an environment that continuously senses, communicates, and acts upon them [3]. However, the practical realization of this vision has only become possible in recent years due to the convergence of several key technological advances [4]. The proliferation of affordable wireless connectivity, the miniaturization of sensors and computing devices, the emergence of cloud computing infrastructure, and the development of sophisticated data analytics capabilities

Copyright@ REST Publisher

have collectively created an ecosystem in which IoT can thrive [4]. Today, several billion devices are connected to the Internet, and this number is expected to grow exponentially in the coming years, with trillions of connected objects expected to be connected [5]. The architectural framework of IoT typically consists of several interconnected layers, each performing unique but complementary functions. At a basic level, the sensing layer consists of physical devices with various sensors temperature, motion, location, humidity, and more that collect data from their environment [6]. The middleware layer provides processing capabilities and service management functions, enabling interoperability between heterogeneous devices and systems [7]. Practical applications of IoT span nearly every sector of modern society, demonstrating the versatile and transformative potential of this technology. In healthcare, IoT enables remote patient monitoring, personalized medicine, and improved hospital operations through connected medical devices and health sensors [9]. Agriculture is benefiting from precision farming techniques that improve irrigation, fertilization, and pest control based on continuous environmental monitoring. In the consumer space, smart homes equipped with connected appliances, security systems, and environmental controls offer unprecedented convenience and efficiency [12]. Despite its enormous promise, IoT faces significant challenges that must be addressed to realize its full potential. Security and privacy concerns are paramount as the proliferation of connected devices creates a multitude of potential vulnerabilities and attack vectors [13]. The sheer volume of IoT applications is generating enormous amounts of data, posing challenges in terms of storage, processing, and analysis. Interoperability issues arise due to the lack of global standards, as different manufacturers and platforms often use incompatible protocols and architectures [14]. Energy efficiency concerns are particularly important for battery-powered devices that need to operate autonomously for long periods of time [15]. Questions about governance, regulatory frameworks, and ethical considerations surrounding data ownership and use require careful attention [16]. These organizations work to develop common protocols, security frameworks, and architectural guidelines that ensure the reliable, secure operation of IoT systems on a global scale. The integration of these technologies will enable more intelligent, autonomous, and efficient systems that can adapt in real time to changing conditions and user needs. For developing countries and emerging economies, IoT offers unique opportunities to overcome traditional infrastructure limitations and address important challenges in areas such as healthcare access, agricultural productivity, and sustainable development [18].

2. MATERIALS AND METHODS

Sensor Accuracy: Sensor accuracy is a critical performance metric in IoT systems, measuring how closely sensor readings match the true values of the monitored parameters. High accuracy ensures reliable data collection, which is the basis for making informed decisions in applications ranging from health monitoring to industrial process control.

Network Latency: Network latency refers to the time delay in data transmission from IoT devices to processing centers or end users. Low latency is essential for real-time applications that require immediate action, such as autonomous vehicles, emergency response systems, and industrial automation. High latency can result in delayed responses, reduced system performance, and potential security risks in time-critical IoT applications.

Device Uptime: Device Uptime measures the percentage of time that IoT devices are operational and accessible. High uptime is critical for maintaining continuous monitoring and service delivery in mission-critical applications such as healthcare, security systems, and infrastructure monitoring. Frequent device failures or downtime can lead to data gaps, compromised system reliability, and increased maintenance costs.

Data Rate: Data rate refers to the amount of information transmitted by IoT devices per unit of time. Optimal data rates ensure sufficient information flow for accurate analysis while avoiding network congestion. Applications requiring high-resolution monitoring, such as video surveillance or environmental sensing, demand higher data rates, while simple condition monitoring may require minimal bandwidth.

Power Consumption: Power consumption is a key consideration for battery-powered IoT devices, especially in remote or inaccessible locations. Efficient power consumption extends the life of the device, reduces maintenance frequency, and lowers operational costs. Energy-efficient designs that incorporate sleep modes, optimized transmission schedules, and energy harvesting techniques are essential for sustainable IoT applications.

Performance Score: Performance Score refers to a comprehensive metric that integrates multiple factors, including accuracy, latency, runtime, data rate, and power efficiency. This comprehensive measurement helps system administrators assess IoT system performance, identify optimization opportunities, and ensure that applications meet specific service level agreements and operational requirements.

Linear regression is one of the most basic and well-understood machine learning algorithms used in IoT analytics. In IoT applications, LR proves to be particularly valuable for scenarios that require transparency and interpretability. For example, in smart building energy management systems, LR can model the relationship between environmental factors and energy consumption, allowing facility

managers to understand which factors affect energy use the most. The simplicity of LR makes it ideal for edge computing scenarios where computational resources are limited, enabling real-time predictions directly on IoT gateways or edge devices without the need for a cloud connection. However, LR operates under several assumptions, including linearity of relationships, independence of errors, and homogeneity. These constraints limit its performance when dealing with complex, nonlinear patterns frequently encountered in IoT data streams. Despite these limitations, LR serves as an excellent baseline model, and the explanatory power of this model is valuable for many important applications.

Random Forest Regression refers to a sophisticated ensemble learning approach that overcomes many of the limitations of linear models. This ensemble method introduces randomness in the selection of training samples (bootstrap aggregation or packing) and in the selection of features considered in each segment, reducing over fitting and improving generalization. For IoT applications, RFR offers several compelling advantages. The algorithm naturally handles nonlinear relationships, making it well-suited for complex IoT scenarios such as predictive maintenance in industrial settings, where equipment failure patterns may depend on complex interactions between multiple sensor measurements. RFR exhibits robustness to outliers and noisy data, common characteristics of real-world IoT sensor streams where communication errors, sensor failures, or environmental interference can degrade data quality. In smart agriculture applications, RFR excels at predicting crop yields by integrating diverse data sources, including soil moisture sensors, weather stations, satellite imagery, and historical agricultural records. The algorithm automatically captures the complex interactions between these variables, without the need for explicit feature engineering. Similarly, in healthcare IoT applications, RFR can predict patient health outcomes by analyzing data from multiple wearable sensors and medical devices, identifying subtle patterns that linear models might miss.

3. ANALYSIS AND DISCUSSION

TABLE 1. Internet of Things Descriptive Statistics

	Sensor Accuracy	Network Latency	Device Uptime	Data Rate	Power Use	Performance Score
count	20.0000	20.0000	20.0000	20.0000	20.0000	20.0000
mean	88.8500	43.7500	95.8000	12.2250	4.9150	81.3000
std	5.1019	12.4810	2.8396	2.1413	0.7506	10.1260
min	80.0000	27.0000	90.0000	9.0000	4.0000	64.0000
25%	84.7500	32.5000	94.0000	10.7250	4.2000	73.7500
50%	89.5000	44.0000	96.5000	12.1000	4.8000	82.5000
75%	93.2500	52.7500	98.0000	14.0500	5.4500	89.5000
max	96.0000	65.0000	99.0000	15.8000	6.2000	96.0000

This dataset provides insights into the performance of an Internet of Things (IoT) system by analyzing six key parameters across 20 observations. The results show that sensor accuracy is consistently high, with an average of 88.85% and a narrow spread, indicating reliable data capture. The minimum value of 80% and a maximum of 96% represent acceptable variability, and the median value closely matches the mean, confirming consistency in sensor performance. Device uptime also shows strong consistency, with an average of 95.8%, indicating that devices are active most of the time across the network with minimal downtime. Network latency shows high variability. The data rate averages 12.23 Mbps, with moderate variability, indicating stable data transfer capacity.

Copyright@ REST Publisher

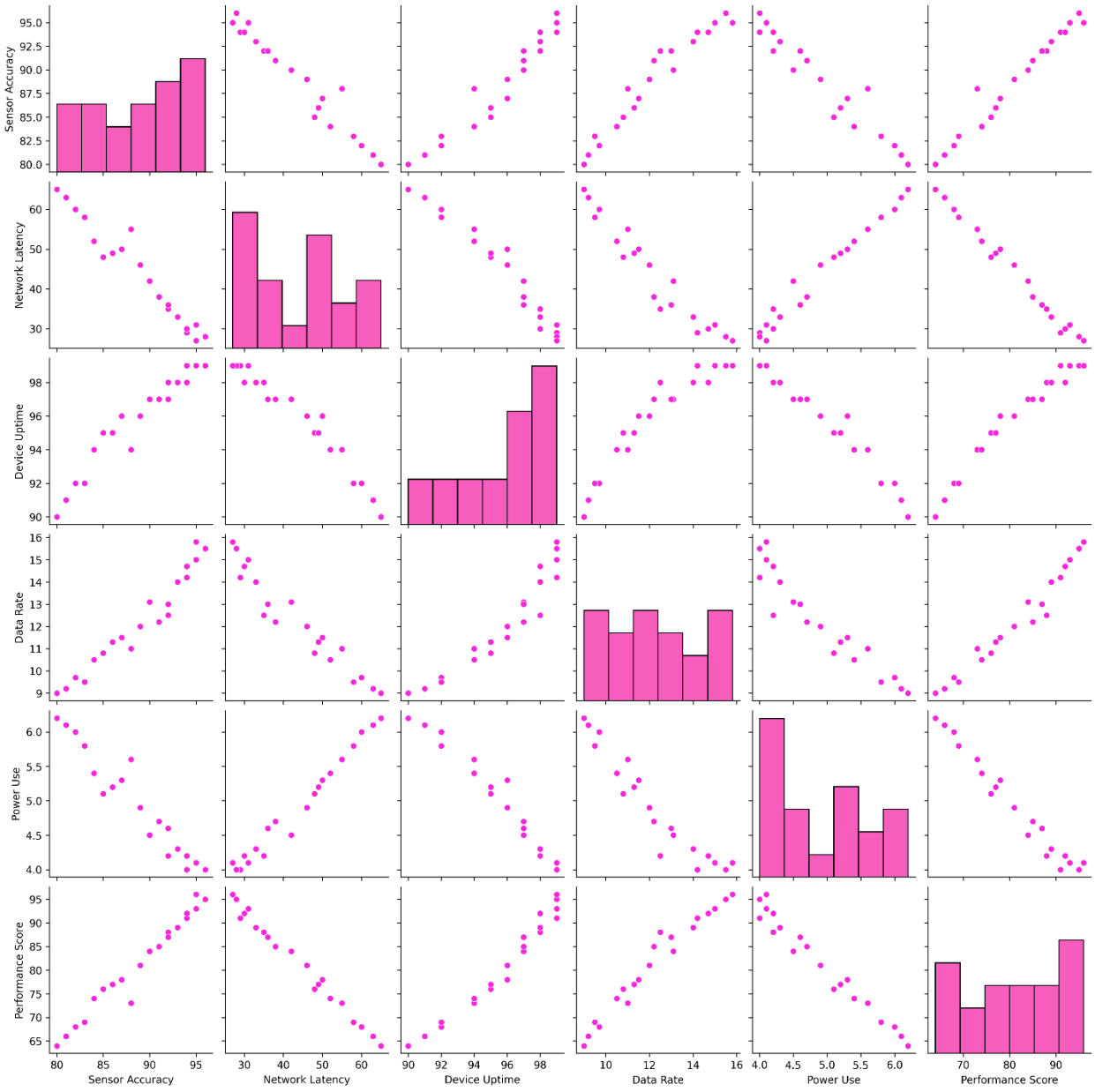


FIGURE 1. Internet of Things Effect of Process Parameters

The scatterplot matrix illustrates the relationships between key IoT performance variables, revealing several clear patterns. Sensor accuracy, device uptime, data rate, and performance score all show strong positive relationships with each other. As sensor accuracy increases, device uptime and data rate also increase, ultimately contributing to overall performance. This indicates that well-calibrated sensors and consistent device performance directly improve system performance. Network latency shows a consistent negative relationship with most variables. Higher latency corresponds to lower sensor accuracy, reduced uptime, and weaker performance scores. This pattern highlights latency as a limiting factor in IoT environments, where delays can disrupt data flow and system responsiveness. Power consumption shows moderate variation and a generally negative relationship with performance variables, indicating that more efficient devices tend to perform better.

Copyright@ REST Publisher



FIGURE 2. Internet of Things Effect of Process Parameters

The correlation heat map shows strong and consistent relationships between IoT system variables. Sensor accuracy, device uptime, data rate, and performance score are all highly positively correlated, with coefficients above 0.95. This indicates that improvements in sensor accuracy, stable device operation, and higher data transfer rates directly contribute to overall system performance. These variables work together to improve reliability and responsiveness across the IoT network. Network latency and power consumption show strong negative correlations with performance-related variables. High latency is associated with lower sensor accuracy, reduced uptime, and poor performance scores, highlighting its disruptive effect on system performance. Similarly, increased power consumption corresponds to poor performance, indicating that efficient devices are operating more efficiently. The nearly perfect negative correlations between latency and performance (–0.99) and power consumption and uptime (–0.98) reinforce these trends.

TABLE 2. Linear Regression Models Performance Score Train and Test performance metrics

Linear Regression	Train	Test
R2	0.9990	0.9706
EVS	0.9990	0.9708
MSE	0.1180	2.1224
RMSE	0.3435	1.4569
MAE	0.2988	1.0387
Max Error	0.6605	3.2921
MSLE	0.0000	0.0003
Med AE	0.2580	0.6892

Copyright@ REST Publisher

The linear regression results indicate a highly accurate predictive model, with strong performance on both the training and test datasets. The R^2 and EVS values for training (0.9990) show that it explains almost all the variation in the sample data, while the test values (around 0.97) confirm strong generalization with little performance degradation. This demonstrates that the model effectively captures the underlying patterns without significant over fitting. The error metrics also support this conclusion. The MSE and RMSE increase from training to testing, but only marginally, indicating that the prediction errors are very small. Similarly, the MAE increases moderately on the test set (1.0387), indicating that the average prediction error is still within an acceptable range.

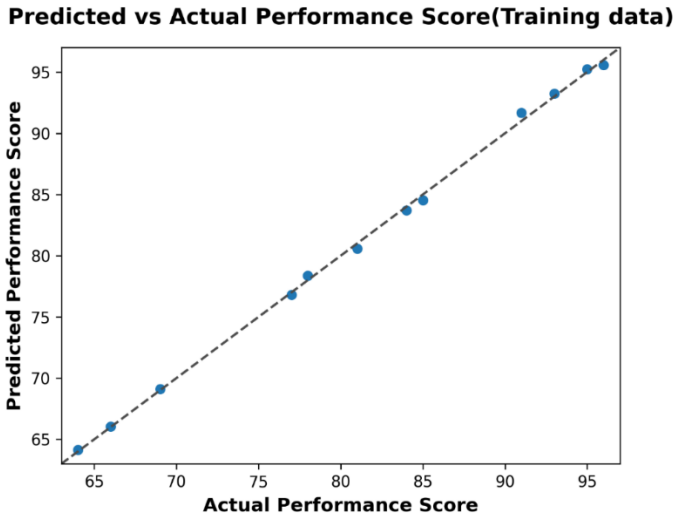


FIGURE 3. Random Forest Regression Performance Score Training

The graph demonstrates a strong relationship between the actual and predicted performance scores in the training data. Most of the data points are located very close to the diagonal reference line, indicating that the model's predictions closely reflect the actual values. This indicates very low error and shows that the model has successfully learned the underlying data patterns. The tight clustering of points on the dashed line, without large deviations or outliers, highlights the high accuracy of the model, ensuring reliable performance across all score levels.



FIGURE 4. Random Forest Regression Performance Score Testing

This graph shows the relationship between actual and predicted performance scores for the test data, providing insights into how well the model generalizes beyond the training set. Most of the points are reasonably close to the diagonal reference line, indicating that the model maintains good predictive accuracy even when applied to unseen data. This indicates that the model is not over fitted and can reliably estimate performance scores in real-world situations. Although the alignment is slightly less tight compared to the training data, the deviations are small and consistent. Some points fall moderately above or below the reference line, reflecting the small prediction errors expected in any test set.

TABLE 2. Random Forest Regression Models Hydrophobicity Train and Test performance metrics

Random Forest Regression	Train	Test
R2	0.9960	0.9677
EVS	0.9960	0.9738
MSE	0.4615	2.3276
RMSE	0.6793	1.5257
MAE	0.5563	1.2494
Max Error	1.3850	2.5200
MSLE	0.0001	0.0004
Med AE	0.4375	1.2000

The performance results of the random forest regression model demonstrate its strong predictive ability for both training and test data. The high R^2 values for the train (0.9960) and test (0.9677) sets indicate that the model accounts for most of the variance in the performance scores, reflecting a good fit and reliable generalization. Similarly, the high EVS values confirm that the model effectively identifies key patterns in the data. Error metrics such as MSE, RMSE, and MAE are low in both datasets, indicating consistent prediction accuracy. Although the test errors are slightly higher, they remain acceptable, suggesting minimal overfitting and consistent model behavior. Even the largest errors are relatively small, and the very low MSLE values show negligible logarithmic discrepancies.

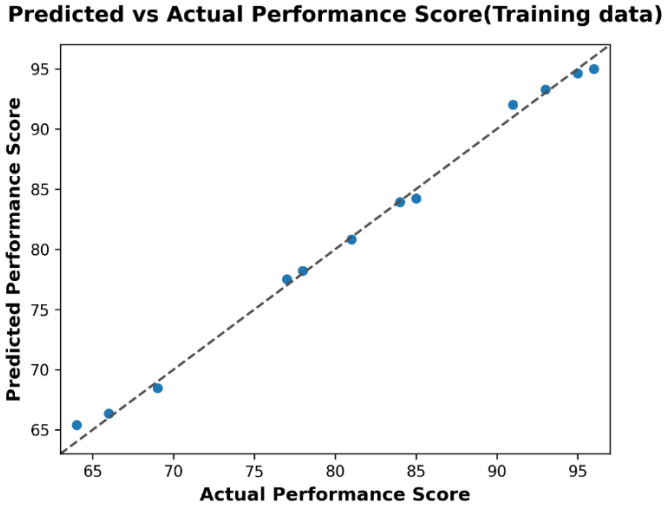


FIGURE 5. Random Forest Regression Hydrophobicity Training

This scatterplot visualizes the relationship between actual and predicted performance scores on the training data, with the dashed diagonal line representing correct predictions, where the predicted values match the actual values exactly. This plot shows excellent model performance because all data points cluster very closely on the diagonal reference line. This tight alignment indicates that the model's predictions are highly accurate over the entire range of performance scores from approximately 65 to 95. The standard deviation shows that the model performs well in both the low and high score ranges, without any systematic bias. The minimum deviation from the best line indicates a strong model fit with low prediction error. Points slightly above or below the diagonal indicate small overpredictions or underpredictions, respectively, but these deviations appear to be very small.

Copyright@ REST Publisher

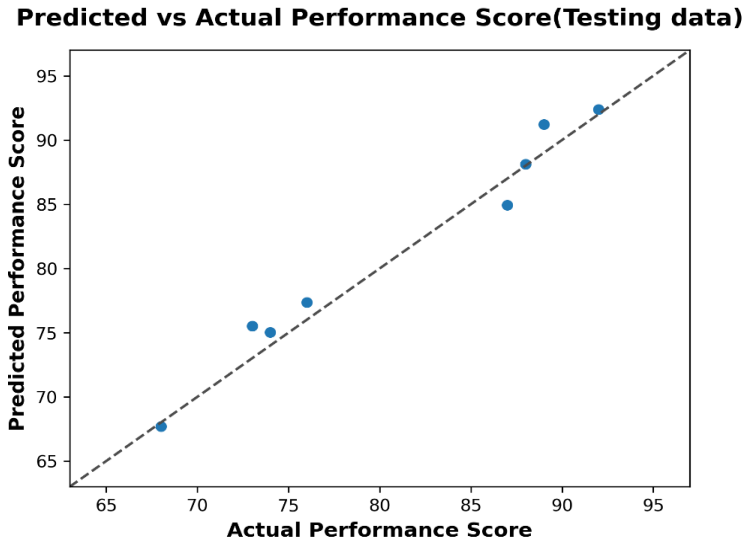


FIGURE 6. Random Forest Regression Hydrophobicity Training

This scatterplot shows the predictive performance of the model on the test data, comparing the predicted and actual performance scores. The dashed diagonal line represents the ideal situation where the predictions match the actual values perfectly. The plot demonstrates good but not perfect model generalization. The data points generally follow the diagonal reference line, which indicates reasonable predictive accuracy over a score range of approximately 68 to 93. However, there is a significantly higher scatter compared to typical training data patterns, which is expected and healthy—it indicates that the model is not simply memorizing the training examples, but has learned common patterns. Some observations include small deviations from the ideal line, especially in the mid-range scores where the points show small overpredictions and underpredictions. The gap between the points and the reference line represents the prediction error.

4. CONCLUSION

This paper successfully demonstrates the effectiveness of applying linear regression and random forest regression techniques to Internet of Things performance prediction challenges. A comprehensive evaluation of the two algorithms reveals their unique advantages and applicability in different IoT environments, providing valuable insights for practitioners and researchers in the field. Experimental results confirm that both linear regression and random forest regression can effectively model IoT performance patterns, achieving high accuracy on training data with close alignment to the real values. The linear regression model offers computational efficiency and interpretability, making it suitable for scenarios requiring explicit decision-making processes and real-time predictions with limited computational resources. Its straightforward implementation and low overhead make it particularly attractive for resource-constrained IoT edge devices. Random forest regression demonstrates superior prediction capabilities by leveraging ensemble learning to capture complex, nonlinear relationships within IoT data. The model's robustness to outliers and ability to handle high-dimensional feature spaces make it suitable for sophisticated IoT applications involving multiple sensor inputs and complex environmental factors. Experimental validation confirms that both models maintain good generalization performance, which is essential for reliable real-world application. The small performance degradation observed from training to test data indicates healthy model behavior, indicating successful learning without overfitting. This characteristic ensures that the models can adapt to changing IoT environments where data patterns may evolve over time. It contributes to the growing body of knowledge about machine learning applications in IoT systems, demonstrating practical approaches for performance optimization and predictive analytics. Future research directions include exploring deep learning architectures for more complex IoT scenarios, implementing hybrid models that combine multiple algorithms, and building adaptive learning systems that continuously update predictions as new data becomes available. Exploring real-time streaming analytics and distributed learning architectures could further enhance IoT system intelligence.

REFERENCES

[1]. Rose, Karen, Scott Eldridge, and Lyman Chapin. "The internet of things: An overview." The internet society (ISOC) 80, no. 15 (2015): 1-53.

[2]. Atzori, Luigi, Antonio Iera, and Giacomo Morabito. "The internet of things: A survey." Computer networks 54, no. 15 (2010): 2787-2805.
[3]. Madakam, Somayya, Ramya Ramaswamy, and Siddharth Tripathi. "Internet of Things (IoT): A literature review." Journal of computer and communications 3, no. 5 (2015): 164-173.
[4]. Rose, Karen, Scott Eldridge, and Lyman Chapin. "The internet of things: An overview." The internet society (ISOC) 80, no. 15 (2015): 1-53.
[5]. Chen, Yen-Kuang. "Challenges and opportunities of internet of things." In 17th Asia and South Pacific design automation conference, pp. 383-388. IEEE, 2012.
[6]. Fleisch, Elgar. "What is the internet of things? An economic perspective." Economics, Management, and financial markets 5, no. 2 (2010): 125-157.
[7]. Da Xu, Li, Wu He, and Shancang Li. "Internet of things in industries: A survey." IEEE Transactions on industrial informatics 10, no. 4 (2014): 2233-2243.
[8]. Tan, Lu, and Neng Wang. "Future internet: The internet of things." In 2010 3rd international conference on advanced computer theory and engineering (ICACTE), vol. 5, pp. V5-376. IEEE, 2010.
[9]. Coetzee, Louis, and Johan Eksteen. "The Internet of Things-promise for the future? An introduction." In 2011 IST-Africa Conference Proceedings, pp. 1-9. IEEE, 2011.
[10]. Villamil, Sebastian, Cesar Hernández, and Giovanny Tarazona. "An overview of internet of things." Telkomnika (Telecommunication Computing Electronics and Control) 18, no. 5 (2020): 2320-2327.
[11]. Reinfurt, Lukas, Uwe Breitenbücher, Michael Falkenthal, Frank Leymann, and Andreas Riegg. "Internet of things patterns." In Proceedings of the 21st European conference on pattern languages of programs, pp. 1-21. 2016.
[12]. Stankovic, John A. "Research directions for the internet of things." IEEE internet of things journal 1, no. 1 (2014): 3-9.
[13]. Tsai, Chun-Wei, Chin-Feng Lai, Ming-Chao Chiang, and Laurence T. Yang. "Data mining for internet of things: A survey." IEEE Communications Surveys & Tutorials 16, no. 1 (2013): 77-97.
[14]. Madakam, Somayya. "Internet of things: smart things." International journal of future computer and communication 4, no. 4 (2015): 250.
[15]. Li, Shancang, Theo Tryfonas, and Honglei Li. "The Internet of Things: a security point of view." Internet Research 26, no. 2 (2016): 337-359.
[16]. Huang, Yinghui, and Guanyu Li. "Descriptive models for Internet of Things." In 2010 International Conference on Intelligent Control and Information Processing, pp. 483-486. IEEE, 2010.
[17]. Huang, Yinghui, and Guanyu Li. "Descriptive models for Internet of Things." In 2010 International Conference on Intelligent Control and Information Processing, pp. 483-486. IEEE, 2010.
[18]. Zanella, Andrea, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, and Michele Zorzi. "Internet of things for smart cities." IEEE Internet of Things journal 1, no. 1 (2014): 22-32.
[19]. Grömping, Ulrike. "Relative importance for linear regression in R: the package relaimpo." Journal of statistical software 17 (2007): 1-27.
[20]. Kejian, Liu. "A new class of blased estimate in linear regression." Communications in Statistics-Theory and Methods 22, no. 2 (1993): 393-402.
[21]. Leatherbarrow, Robin J. "Using linear and non-linear regression to fit biochemical data." Trends in biochemical sciences 15, no. 12 (1990): 455-458.
[22]. Huang, Mengyu. "Theory and Implementation of linear regression." In 2020 International conference on computer vision, image and deep learning (CVIDL), pp. 210-217. IEEE, 2020.
[23]. Schmidt, Amand F., and Chris Finan. "Linear regression and the normality assumption." Journal of clinical epidemiology 98 (2018): 146-151.
[24]. Chen, Wei, Xiaoshen Xie, Jiale Wang, Biswajeet Pradhan, Haoyuan Hong, Dieu Tien Bui, Zhao Duan, and Jianquan Ma. "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility." Catena 151 (2017): 147-160.
[25]. Liu, Dong, Zhongrui Fan, Qiang Fu, Mo Li, Muhammad Abrar Faiz, Shoaib Ali, Tianxiao Li, Liangliang Zhang, and Muhammad Imran Khan. "Random forest regression evaluation model of regional flood disaster resilience based on the whale optimization algorithm." Journal of Cleaner Production 250 (2020): 119468.
[26]. Wang, Xuanhui, Tailian Liu, Xilai Zheng, Hui Peng, Jia Xin, and Bo Zhang. "Short-term prediction of groundwater level using improved random forest regression with a combination of random features." Applied Water Science 8, no. 5 (2018): 125.
[27]. Johansson, Ulf, Henrik Boström, Tuve Löfström, and Henrik Linusson. "Regression conformal prediction with random forests." Machine learning 97, no. 1 (2014): 155-176.
[28]. Mendez, Guillermo, and Sharon Lohr. "Estimating residual variance in random forest regression." Computational statistics & data analysis 55, no. 11 (2011): 2937-2950.