THE USA JOURNALS
THE AMERICAN JOURNAL OF ENGINEERING AND TECHNOLOGY (ISSN – 2689-0984)
VOLUME 06 ISSUE12

PUBLISHED DATE: - 19-12-2024
DOI: - https://doi.org/10.37547/tajet/Volume06Issue12-12 **PAGE NO.: - 119-140**

RESEARCH ARTICLE **Open Access**

LEVERAGING MACHINE LEARNING FOR RESOURCE OPTIMIZATION IN USA DATA CENTERS: A FOCUS ON INCOMPLETE DATA AND BUSINESS DEVELOPMENT

Md Sumsuzoha
Master of Science in Business Analytics, Trine University

MD Sohel Rana
Executive Ph.D. in Business Analyst, University of Cumberlands

Md Shahidul Islam
MBA in business analytics, International American University

Md Khalilor Rahman
MBA, Business analytics, Gannon University, Erie, PA, USA

Mitu Karmakar
School of Business, International American University, Los Angeles, California, USA

Md Sazzad Hossain
MBA, business analytics, Gannon University, Erie, PA, USA

Reza E Rabbi Shawon
MBA Business Analytics, Gannon University, Erie, PA

Corresponding Author: Md Sumsuzoha

Abstract
Data centers form the cornerstone of modern digital infrastructure, enabling operations from e-commerce and streaming to cloud computing and artificial intelligence. The United States, at the forefront of technology, houses some of the world's most extensive and technologically advanced data centers as a part of its economic and technological framework. This study aimed to explore how different ML techniques can be used for optimizing resource utilization by data centers in the US, focusing on strategies to handle incomplete data and implications for business development. The dataset was retrieved from the GitHub repository, which provided a rich dataset of resource usage metrics and operational data from U.S. data centers. It contained complex and fine-grained information that was necessary to optimize data center performance and deal with incomplete data challenges. A detailed description of the dataset and its key attributes was provided. It was designed for analyzing resource usage patterns in data centers, putting much emphasis on energy efficiency, workload distribution, and operational reliability. This integration of time-series data with sensor readings and performance logs provided a comprehensive overview of resource consumption and environmental conditions in data center operations. This dataset was curated for the engagement of machine learning models in the study and optimization of resource consumption along with the challenges of missing data. Analysis of resource utilization in US data centers was accomplished using the application of various models for machine learning, most notably, Logistic Regression, Random Forest, and Support Vector Machines; Retrospectively, the Random Forest and SVM models seem to be robust and reliable, placing the Random Forest slightly above, given their performance is nearly perfect for training and testing. The application of machine learning techniques holds huge potential for the reformation of resource management in US data centers. These models analyze a pattern in historical data to predict future resource demands, thus allowing optimized resource allotment and minimizing operational costs.

Keywords Resource Optimization; Energy Efficiency; Data Centers; Incomplete Data; Machine Learning; Digital Infrastructure; Business Development, USA.

INTRODUCTION

Background and Motivation

As per Rahman et al. (2023), data centers form the linchpin of modern digital infrastructure, enabling everything from e-commerce and streaming to cloud computing and artificial intelligence. The United States, at the forefront of technology, houses some of the world's most extensive and technologically advanced data centers as a part of its economic and technological framework. Recent studies have indicated that U.S. data centers make up about 2% of the nation's total electric use, and this trend is probably continuing to increase with the growth in the number of computational resources the modern world requires. Shawon et al.(2023b), asserted that the scale and power involved in such facilities outline the importance of efficiently using resources to achieve sustainability. To that effect, efficient resource utilization not only cuts operation costs but also meets global efforts to ensure carbon emissions are reduced. An area where inefficiency reins in include cooling systems, hardware utilization, and power distribution. It is on this basis that addressing such inefficiencies becomes quite important in maintaining the operational viability of data centers to support their expansion in the digital economy (Alam et al., 2023; Abdulazeez et al. 2024).

Problem Statement

Notwithstanding, most data centers still suffer from ineffective use of their resources due to factors such as imbalance in workload, inefficient cooling methods, and underutilization of hardware. Incomplete data is one of the most critical barriers to optimization (Al Mukaddim et al., 2023a; Feng et al., 2024). Missing sensor readings, missing hardware logs, and incomplete environmental data pose a problem to the traditional resource management system, leading to less-than-optimal decisions. The inability to handle incomplete data weakens the reliability and robustness of resource optimization. The specific value of machine learning lies in providing transformative opportunities regarding the aforementioned challenges exacerbated by incomplete data that plagues traditional optimization approaches so often (Buiya et al. 2023; Ezeigweneme et al., 2024). This study aims to explore how different ML techniques can be used for optimizing resource utilization by data centers in the US, focusing on strategies to handle incomplete data and implications for business development. Machine Learning potentially improves efficiency, reduces cost, and unleashes paths toward sustainable growth by applying predictive analytics, anomaly detection, and reinforcement learning. Examples of incomplete data are being treated with imputation techniques or robust model design to meet the promise of ML in these most critical environments.

Research Questions

RQ1: How can machine learning techniques be applied to optimize resource usage in USA data centers?

This research question aimed at understanding how machine learning algorithms can help in improving efficiency in data center operations. Analyzing the trend of past data on energy consumption, server utilization, and several other factors, machine learning models can identify patterns and thereby predict the future need for resources. This information can thus be used to optimize resource allocation, reduce energy consumption, and minimize operational costs.

RQ2: How can incomplete data be addressed in machine learning models for data center resource optimization?

This research question aims to explore imputation, data augmentation, or robust learning algorithms that can help deal with missing data for the accuracy and reliability of machine learning models. Addressing incomplete data challenges will improve the performance of machine learning models and enhance data center resource optimization. Data centers often generate large volumes of data; however, due to several reasons such as sensor failures, network outages, or human error, this data can be incomplete.

RQ3: What are the business development implications of optimizing resource usage in data centers?

The question deals with the higher-order business value linked to data center resource optimization. The improvement in profitability and sustainability for a data center will result from reductions in energy consumption and operational costs. Besides, optimized resource usage will enable data centers to scale their operations more efficiently and respond to changing business needs. Moreover, data-driven insights gained from machine learning models can help identify new business opportunities and improve service delivery.

Significance of the Study

The significance of this research lies in its capability to resolve the pressing challenges confronted by U.S. data centers, which are instrumental to the digital economy yet burdened by inefficiencies in resource utilization. This work will advance not only the technical capability of data centers but also their alignment with goals of sustainability and cost reduction through machine learning for optimization. Addressing incomplete data makes this optimization robust and reliable in real-world conditions, considering that data centers must function under such conditions. From the perspective of business development, it has far-reaching implications, as this may mean reduced operational costs, improved reliability in service delivery, and increased competitiveness in the marketplace. This has contributed to wider goals of environmental stewardship and economic development by assuring that data centers continue to serve the growing digital infrastructure safely and efficiently.

LITERATURE REVIEW

Overview of Data Center Operations and Resource Usage in the USA

Islam et al. (2023), posited that data centers have emerged as the foundational building block of the digital economy, supporting a diverse array of applications and services, ranging from cloud computing and e-commerce to AI and critical infrastructure systems. Among all types of facilities in the United States, data centers rank as some of the most power-intensive, consuming about 2% of the total electricity consumption in the nation. The general operations will include server management, cooling systems, storage optimization, and networking infrastructure (Hasanuzzaman et al.; Gong et al., 2024). These centers rely on a delicate balance between computation requirements, power consumption, and physical space, all at adequate uptime and availability.

According to Karmakar et al. (2023), the most critical concerns for resource management facing US data centers are related to energy efficiency. The cooling systems themselves, required to maintain hardware at ideal operating temperatures, consume enormous amounts of energy. Inefficiently utilized servers that are often left running for redundancy further exacerbate the problem. As per Gebreyesus et al. (2023), other challenges include the integration of renewable energy sources and real-time workload distribution across servers. All of these challenges require highly sophisticated approaches, where optimization of resources should guarantee reliability and scalability of operation. ML can outline some very promising ways to solve these complexities through intelligent decision-making in resource management automation (Khan et al., 2023).

Machine Learning for Resource Optimization

Predictive Modeling

Nasiruddin et al. (2023), articulated that Predictive modeling is among the cornerstones of machine learning applications in data centers that involve the forecasting of requirements well in advance from historical and real-time data to enable proactive resource allocation. Examples of some of the applied techniques include regression analysis, neural networks, and time series modeling. For instance, ML-driven predictive models may forecast peak workload periods and take actions that involve the pre-emptive allocating of extra resources to avoid latencies and system overload. These models shine in situations such as seasonal e-commerce traffic or periodic data processing spikes (He et al., 2024; Kumari et al. 2024).

Clustering

Clustering algorithms, like k-means or hierarchical clustering, play a significant role in workload analysis concerning workload patterns and resource optimization. The group data points of similar configuration into one group (Shawon et al. 2024; Li et al., 2024). For example, by applying different clustering methodologies, workloads can be segmented based on the requirements associated with workload processing and then dispatched onto servers that are capable of matching processing needs. This ensures that resources are neither over-provisioned nor underutilized, hence contributing to energy savings and cost reductions (Sumon et al. 2024).

Decision Trees and Ensemble Methods

As per Shil et al. (2024), decision trees and ensemble methods, like random forests and gradient boosting machines, are of particular importance to resource allocation decisions. The algorithms offer interpretable models that forecast optimal settings related to server power distribution, cooling, and workloads. Decision-making is a big positive attribute in generally dynamic environments where operational parameters often change. For example, the decision tree model could suggest operating a cooling power reduction in a zone for low workloads while sustaining efficiency without losing reliability (Krothapalli et al., 2023; Ma et al., 2024).

Reinforcement Learning

Wang et al. (2024), stated that Reinforcement learning has recently become one of the state-of-the-art machine learning approaches for data center optimization. The nature of the RL models, which learn through trial and error to maximize a reward signal, makes them suitable for managing dynamic resources. For instance, using an RL algorithm, cooling systems can autonomously read real-time environmental conditions and optimize energy use independently of human adjustments (Shawon, 2023c). Notable applications include Google's collaboration with DeepMind, in which reinforcement learning reduced cooling energy consumption by 40% within their data centers.

Handling Incomplete Data in Machine Learning

According to Zhang et al. (2024), missing data in a data center operation is a pervasive challenge that might result from hardware failures, malfunctioned sensors, or network failures. Most ML models require large-scale, high-quality datasets for good training and deployment, and hence the treatment of missing data becomes crucial.

Imputation Techniques

Imputation is the process of estimating the values of missing data for a complete dataset. The simple forms mean or median imputation, only take the average of available data for the missing ones. These methods, though simple, tend to oversimplify many complex patterns in data. Richer imputation methods such as KNN or matrix factorization substantially improve accuracy by considering the relationships between variables (Rahman et al., 2023). For instance, KNN uses the correlation between server workloads and cooling demands for missing reading imputation when a dataset of power consumption for a data center is provided.

Interpolation Techniques

Interpolation uses mathematical functions to estimate missing data points within a given range. Linear interpolation, spline interpolation, and polynomial interpolation are common, depending on the complexity of the dataset. For example, in sensor networks that monitor temperature across a data center, missing values will be reconstructed by using spline interpolation to keep the monitoring continuous.

Data Augmentation

Debnath et al. (2023), stated that Data augmentation is a type of technique involved in adding to the size of available datasets with artificially created new records. It would include techniques of bootstrapping or even the generation of synthetic data using variants of GANs for generating realistically distributive samples. This should particularly be the case when missing data occurs in complex patterns, or where not much was initially available from training. Instead, GANs could fill historical lacunas by simulating probable patterns in workload distributions. Robust ML Architectures Equally important is the art of designing ML models that are insensitive to incomplete data. To handle the uncertainty and variability presented by missing data, ensemble models, dropout layers in neural networks, and

Bayesian frameworks are particularly adept. For instance, ensemble models, such as random forests, aggregate predictions from a set of decision trees and therefore reduce dependence on any single tree where data may be missing (Buiya et al. 2024).

Business Development Implications of Resource Optimization

The optimization of resource utilization in data centers has far-reaching ramifications for business development, particularly within the U.S. context, where data centers play a crucial role in both technological and economic growth as described below:

Operational Cost Reduction. As stated earlier, resource optimization can positively affect operational costs, being some of the highest financial burdens for data centers. Consequently, it leads to great savings for the data center with minimized energy consumption through effective distribution of workload and dynamic adjustment of cooling. For example, predictive models estimating a peak period and pre-dedicating resources avoid overprovisioning, hence a reduction in utility bills. In return, these savings might be reinvested to facilitate an increase in infrastructure by applying more advanced technologies, again encouraging a cycle of growth and innovation.

Improved Efficiency and scalability. Optimized resource use ensures data centers operate at peak in handling increased workloads without increased proportional consumption of resources. This aspect of scalability is important in a business perspective where demands for digital services are growing each day. Efficient operations minimize downtime and improve the reliability of services, thus improving customer satisfaction and retention (Islam et al., 2024).

Data-Driven Decision-Making. Adopting ML-driven resource optimization encourages a fact-based decision-making culture. With predictive analytics and real-time monitoring, operators gain profound insight into the performance metrics they need to make informed strategic planning. This places the decision-makers in a good position to develop an investment plan in energy-efficient hardware or renewable energy sources, based on forecasts from data-driven projections about cost savings in the long term (Hasanuzzaman et al., 2023).

Data Collection and Preprocessing

Data Sources

The dataset was retrieved from the GitHub repository, which provided a rich dataset of resource usage metrics and operational data from U.S. data centers. It contained complex and fine-grained information that was necessary to optimize data center performance and deal with incomplete data challenges. A detailed description of the dataset and its key attributes was provided. It was designed for analyzing resource usage patterns in data centers, putting much emphasis on energy efficiency, workload distribution, and operational reliability (Pro-AI-Robikul, 2023). This integration of time-series data with sensor readings and performance logs provided a comprehensive overview of resource consumption and environmental conditions in data center operations. This dataset was curated for the engagement of machine learning models in the study and optimization of resource consumption along with the challenges of missing data. The analyst will have more actionable insights into data center operation for efficient and sustainable performance in the digital economy by making this dataset accessible.

Data Pre-processing

A code snippet in Python was incorporated into the preprocessing pipeline of machine learning: Firstly, the code started by standardizing column names to ensure consistency across multiple datasets. Secondly, it concatenated these datasets vertically, eliminating any duplicate rows. To handle missing values, the code filled in missing entries with the median value of the respective column. Thirdly, the dataset was reordered based on timestamps for chronological analysis. Finally, the resulting merged dataset was saved as a CSV file for any further use. The cleaning process ensured that all data quality and consistency have been pre-processed and increased, in terms of completeness for it to be suitable for model training and evaluation. Handling missing values is an important data pre-processing step. Imputation techniques included the replacement of missing values with estimated ones. Mean, median, or mode were used to replace the values in simple imputation methods. Other versions, such as more sophisticated ones like K-Nearest Neighbors, impute based on the values of similar data points. Another interpolation technique was deployed involving the estimation of missing values from the trend of the surrounding data points. Depending on the complexity of the trend, either simple or spline interpolation is applied.

Exploratory Data Analysis (EDA)

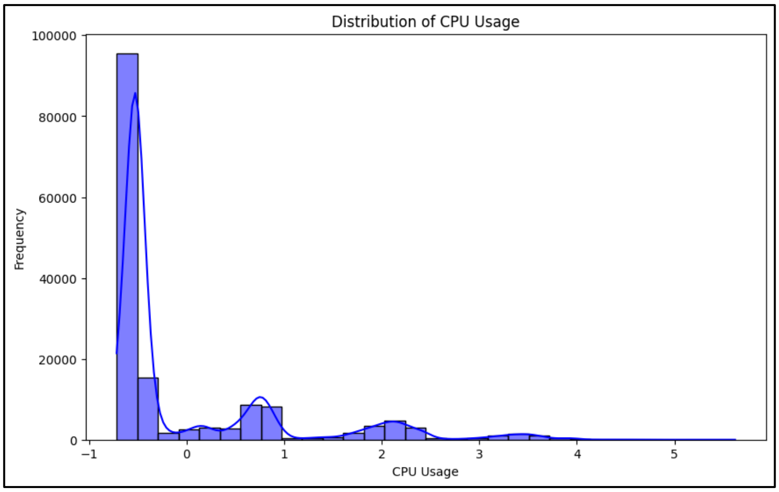


Figure 1: Displays the Distribution of CPU Usage

This histogram displays the distribution of CPU usage with a superimposed density curve. There is a highly right-skewed, or positively skewed, distribution with several distinct peaks. The most prominent peak near the 0 mark reflects very high frequency, about 95,000, of instances with low CPU usage. Several other, much smaller peaks exist at intervals, notably around the values for CPU usage of 1, 2, and 3, but their frequencies are much lower, in the range of 5,000 to 10,000. This pattern suggests that the system is normally running at low CPU utilization, with periods of moderate to high usage creating secondary peaks. The long right tail extending beyond CPU usage of 5 indicates very high CPU utilization, though these occur with very low frequency.

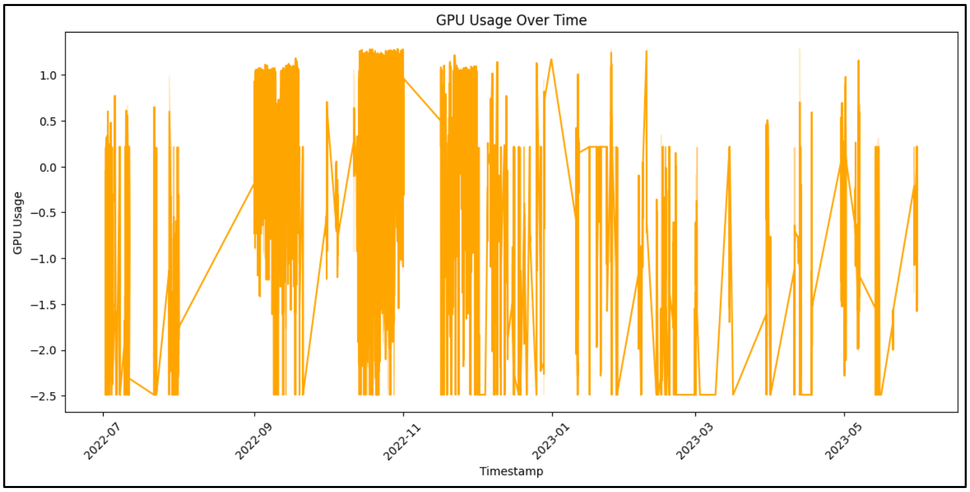


Figure 2: Portrays GPU Usage Over Time

This time series chart above shows a pattern of GPU usage during almost a year, from mid-2022 to mid-2023, in a highly turbulent manner. The values range from -2.5 to 1.0 with very sudden fluctuation between these two bounds. Notice the high activities in September and November of 2022, that it is denser and has fluctuations at a very fast pace. Starting in January 2023, this pattern becomes much more random and less dense with clearly separated spikes. Such a negative value for this metric is somewhat unusual for the case of GPU usage; they might reflect either a measurement scale that is normalized, including possible anomalies in monitoring. Thus, these would reflect batch job-type processing workloads in nature and probably mean that periodic and not quite steady, strong computations should run on the GPUs but also could run in discontinuity.

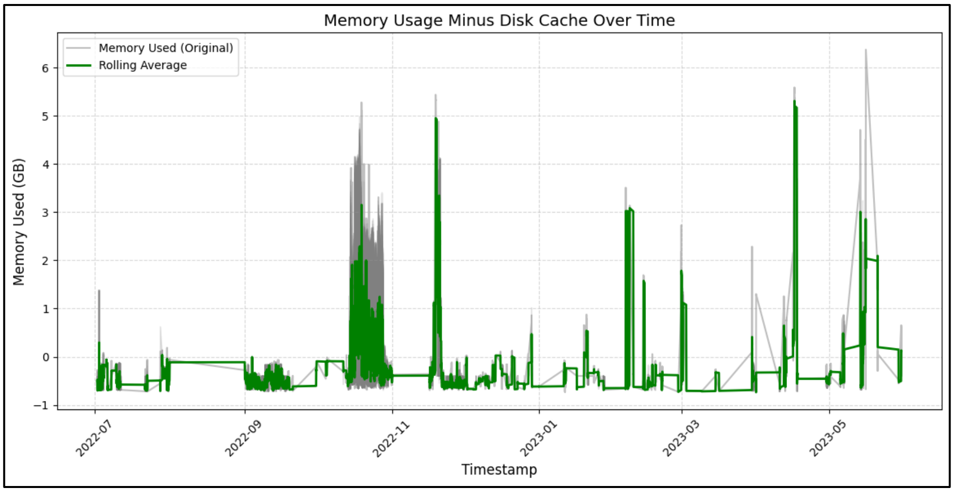


Figure 3: Exhibits Memory Usage Minus Disk Cache Over Time

The above time series graph shows memory usage, minus disk cache, over about a year from mid-2022 to mid-2023; it depicts raw measurements in gray and a rolling average in green. Overall, memory usage is around 0 GB, with a baseline between -1 GB and 0 GB, but there are several large spikes throughout the period. The most prominent spikes are in November 2022, reaching about 5 GB, and in

May of 2023, whereby the usage spikes over 6 GB. The moving average is represented by the green line, which smooths the high-frequency fluctuations and reveals the underlying patterns in the memory usage spikes, usually short-lived before returning to baseline. The negative values of the baseline indicate that essentially, the disk cache exceeds that amount of real memory usage in normal operation, which in general is a characteristic of well-optimized systems where memory not used is used for caching.

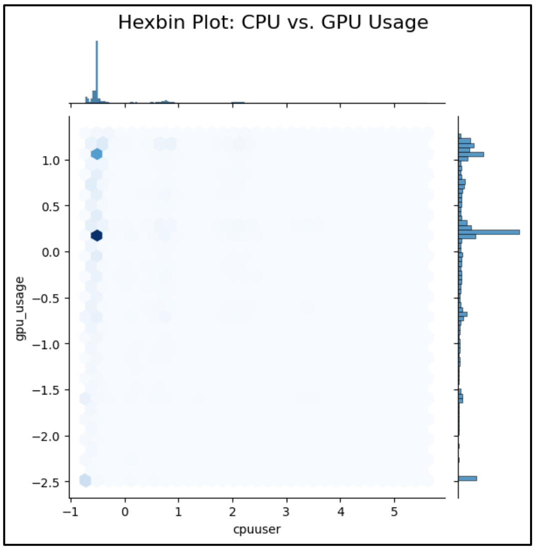


Figure 4: Visualizes Hexbin Plot: CPU vs. GPU Usage

This hexbin plot visualizes the relationship between CPU and GPU usage despite color intensity showing the density. We notice several interesting patterns in that plot: there is for example a strong concentration at CPU usage of around 0 and GPU usage within 0.2, shown by the quite dark blue hexagons falling in this region. The vertical distribution of GPU usage spans from -2.5 to 1.0, while CPU usage extends from -1 to about 5. In particular, there are distinct horizontal bands of activity at GPU usage levels of approximately -2.5, 0.2, and 1.0, indicating that the GPU tends to operate at these specific levels regardless of CPU utilization. The marginal histograms on the edges give some indication that the CPU usage has a right-skewed distribution with many peaks, but the usage of the GPU seems more discrete, even multi-modal, with quite clear states of preference. This is a pattern indicative of CPU and GPU usage being relatively independent, with significant GPU operations happening in relatively distinct states rather than continuously over time.

METHODOLOGY

Feature Engineering and Selection

Feature engineering in this regard should focus on the derivation of meaningful temporal and behavioral patterns from the data on system resource usage. Key engineered features could include rolling statistics, standard deviation, min, and max-over various time windows, such as 1-minute, 5-minute, and 1-hour for CPU, GPU, and memory usage. Other features could be created by calculating the rate of change, peak-to-average ratios, and utilization patterns during specific periods of peak hours versus off-hours. Most valuable in this respect can be the cross-metric feature-engineered variables, such as the ratio between CPU and GPU utilization, memory usage normalized by the CPU activity, or composite metrics that express the overall system load. Time-based features should encode time components, including an hour of the day, day of the week, and month, by using respective sine and cosine transformations to capture the periodic patterns of resource utilization.

Feature selection was driven by both statistical significance and domain relevance. The most important feature selection criteria involved the importance score of features from tree-based models such as Random Forests and XG-Boost, correlation analysis to identify and remove redundant features, and stability selection by iterating multiple times through model training. Features that show high VIFs were kept with great care so that multicollinearity cannot take place. Feature selection gave priority to features that demonstrated consistency with their predictive power across time and system states. Besides, the features should be checked for their computational efficiency and real-time availability since this is seemingly a monitoring system to which rapid calculations would be advantageous. Such a combination of filter methods-statistical tests, wrapper methods-recursive feature elimination, and embedded methods-LASSO regularization provided a robust approach toward identifying the most relevant features while preserving model interpretability.

Model Selection and Justification

Analysis of resource utilization in US data centers was accomplished using the application of various models for machine learning, most notably, Logistic Regression, Random Forest, and Support Vector Machines; each best fits different characteristics of either or both of the aspects relating to the data under discussion or even for the objective of performing prediction: **Logistic Regression** represents the statistical model used while attempting to achieve binary classifications such as overloading versus not overloading a certain server within predetermined conditions. It is simple and interpretable; hence, it forms a very good starting point for selecting the main predictors of operational anomalies. **Random Forest** is an ensemble learning method based on decision trees, especially efficient while dealing with complex nonlinear relationships between data. By aggregating multiple trees' predictions, robustness and accuracy are gained for tasks such as the prediction of energy consumption patterns or the classification of server workloads. They are particularly effective in classification and regression problems in high-dimensional spaces. **Support Vector Machines **classify data by drawing a hyperplane that separates data points, making them useful in anomaly detection in resource usage or classification of the efficiency of cooling systems under changing conditions. Each of these models brings distinct strengths to the table, enabling a comprehensive approach to data center optimization. A choice of these models is justified according to the characteristics of the dataset and possible goals of the analysis.

Logistic regression fits binary outcomes and, therefore, stands as a baseline model in many respects, such as interpretability and low computational needs. Random Forest handles missing data in these data well, and generally provides superior performances on noisy and complicated data; hence, it also offers excellent representations concerning incomplete and multifaceted datasets. This will also shed light on the most influential variables causing resource utilization based on the feature importance of this metric. Also, since events like failure are so rare and critical in a data center, SVMs can handle high-dimensional and imbalanced datasets. All these factors together bring about interpretability, robustness, and precision, hence providing great help in predicting and reaching effective decisions over the optimal utilization of resources in a data center.

Train and Test Framework

A rigorous training and testing framework is pivotal to guarantee that the machine learning algorithms generalize well to unseen data and perform effectively in real-world scenarios. The commonly used splitting of data into the training-testing set is in an 80-20% fashion, where model training is done on 80% of the data, and the rest of it, 20% is used for testing purposes. This division ensures that the model is evaluated on a subset of data it has not seen during training, providing an unbiased assessment of its predictive accuracy. Care is taken to ensure the data split maintains the distribution of key variables, particularly when dealing with imbalanced classes, such as anomalies in server workloads or rare energy efficiency issues. Stratified sampling is used in classification to maintain class proportions in both training and testing sets.

To enhance further the reliability of model evaluation, k-fold cross-validation is applied. In this technique, a training set is divided into k subsets or "folds." Training on k-1 folds and validation on the remaining one is performed by the model while rotating through all folds until each subset has served as the validation set. Typical choices for k are 5 or 10; the choice balances computational efficiency with the thoroughness of evaluation. This helps to avoid overfitting on a particular training subset and provides a better, more realistic estimate of the performance of the model. If there is a time series or a sequence dependency, methods such as time-based cross-validation or walk-forward validation should be applied to preserve order when splitting the data.

Hyperparameter Tuning

Model hyperparameter optimization is a crucial process to avoid underfitting or overfitting to get the peak performance. Grid search is an exhaustive approach where a model needs to consider every different combination with pre-specified values of hyperparameters and then select the best setting. For instance, a Random Forest model would make use of grid search to tune the number of trees via n_estimators, maximum tree depth via max_depth, and minimum number of samples required to split a node via min_samples_split. While computationally expensive, the warranty with grid search is that the best combination of parameters within the predefined space will be found.

As a more computationally efficient alternative, random search randomly samples a subset of the hyperparameter combinations and evaluates them. Random search often outperforms grid search in finding near-optimal solutions at considerably lower computational costs. Random search was especially useful in models with lots of hyperparameters, such as Support Vector Machines or neural networks. Besides, more complex approaches like Bayesian optimization an approach that iteratively builds a probabilistic model of the objective function and selects hyperparameters to evaluate, based on expected improvement also possible. Hyperparameter tuning integrates with cross-validation to ensure the performance of selected parameters is not specific to one dataset split. Implemented together, these approaches provide that machine learning models are optimized for precision and reliability, therefore increasing their usefulness in resource optimization for data centers.

Performance Evaluation Metrics

Evaluation of machine learning models requires a set of complete metrics that describe various aspects of model effectiveness, in particular for tasks like classification and regression. For classification, accuracy is a basic metric expressing the number of properly predicted instances against the total within a dataset. However, it is not enough in certain cases, especially when one is dealing with datasets that have imbalances in class distribution. It means that precision, recall, and the harmonic mean between the two-sensitivity F1 are more understandable for this sort of analysis: Precision describes what fraction of true positive predictions comprise all positive predictions. In cases where there is a high cost related to the so-called "false alarm", the Precision would prove itself rather valuable in such scenario analysis. On the contrary, Recall/Sensitivity is computed in the identification rate concerning positive predictions classified as such; this remains fundamental when poor detection entails huge damage to important aspects with sensitive or fatal consequences that have to do with fault anomaly detection. Another important metric in binary classification is the AUC-ROC curve, which stands for the Area Under the Receiver Operating Characteristic Curve. It gives a model the ability to trace the generally positive cases against the generally negative cases at different threshold levels.

RESULTS

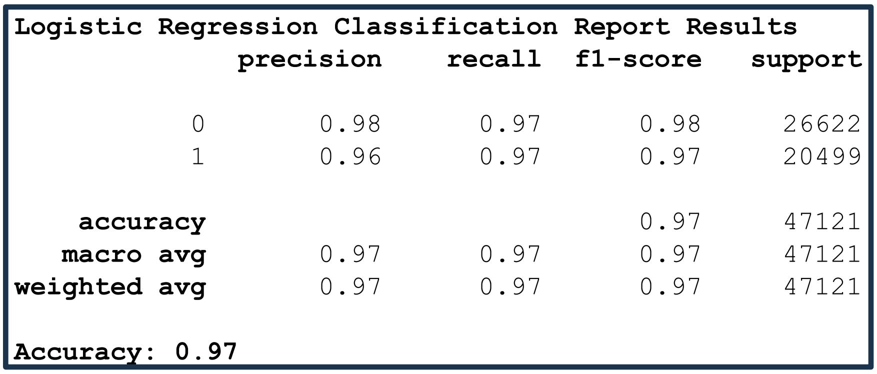
Model Performance

a) Logistic Regression

An appropriate Python code snippet was deployed during the implementation of logistic regression for classification. It instantiated a Logistic-Regression object, trained the instance on the training data, X_train, and y_train, and utilized the trained model to make predictions on the test data, X_test to produce the predicted labels, y_pred_lr. The classification report summarized the precision, recall, F1-score, and support for each class in detail. Additionally, the accuracy score is also calculated to check the overall correctness of the prediction. This report helped the analyst to understand the strengths and weaknesses of the model, informing further improvements or decisions from its predictions.

Output:

Table 1: Exhibits the Logistic Regression Classification Report Results



```
Logistic Regression Classification Report Results
              precision    recall  f1-score   support

           0       0.98      0.97      0.98     26622
           1       0.96      0.97      0.97     20499

    accuracy                           0.97     47121
   macro avg       0.97      0.97      0.97     47121
weighted avg       0.97      0.97      0.97     47121

Accuracy: 0.97
```

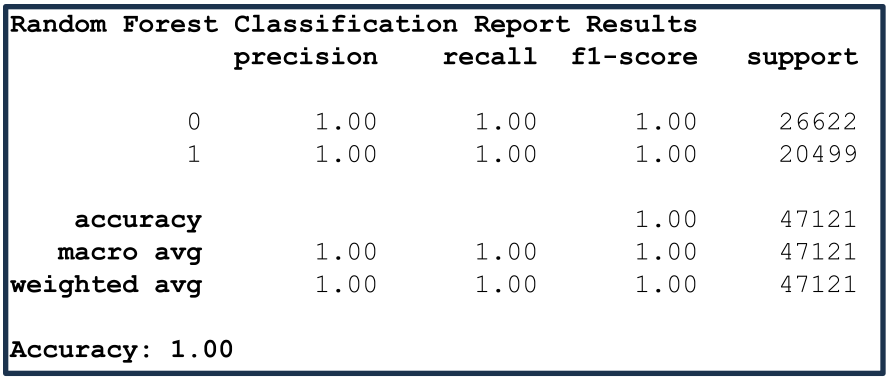
The classification report above essentially gives the details of the performance of the Logistic Regression model. This model has given a very good performance for class 0 with a high precision of 0.98, recall of 0.97, and F1-score of 0.98, reflecting well on its accuracy in identifying true positive instances. For class 1, the model has had a good performance at precision at 0.96, recall at 0.97, and an F1-score of 0.97; hence, it classified most of the true positive instances correctly for this class. The overall accuracy is 0.97, which means the model can predict class labels correctly for about 97% of the test instances. That means the logistic regression model does quite a good job of classifying data into two classes with high accuracy while the performance is balanced in both classes.

b) Random Forest

Python code snippet implemented a random forest classifier. It started by instantiating a Random-Forest-Classifier object with a specified number of trees, n_estimators=100, and a random state for reproducibility. The model trained on the train data X_train, y_train. After training, the model was used on the test set X_test to make predictions, y_pred_rf. The classification report presented the model's performance in detail with precision, recall, F1-score, and support for each class. Other than that, the accuracy score is also calculated as a measure of the overall correctness of the predictions. This report assisted in understanding the strengths and weaknesses of the model and thus further helped in making improvements and decisions based on its predictions.

Output:

Table 2: Depicts the Random Forest Classification Report Results



```
Random Forest Classification Report Results
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     26622
           1       1.00      1.00      1.00     20499

    accuracy                           1.00     47121
   macro avg       1.00      1.00      1.00     47121
weighted avg       1.00      1.00      1.00     47121

Accuracy: 1.00
```

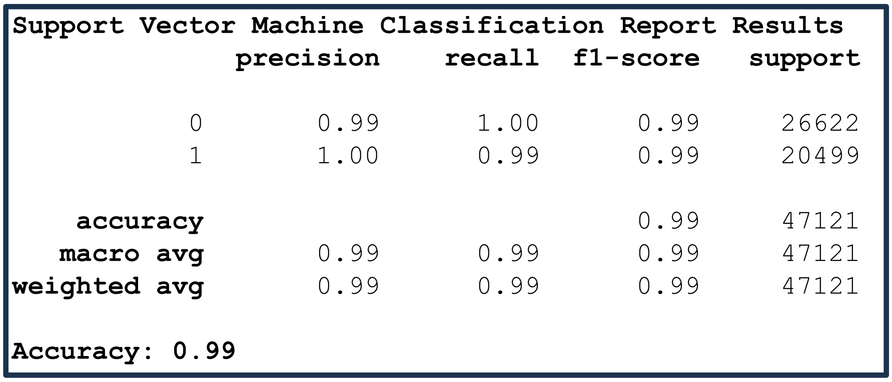
The classification report of the Random Forest model gives an outstanding performance. For class 0 and class 1, the model has a perfect precision, recall, and F1-score of 1.00. That means the model correctly identified all true positive instances and did not have any false positives or false negatives. The overall accuracy of the model is also 1.00, which further confirms the ability of the model to classify the data with 100% accuracy. Results obtained here show that Random Forest is outstanding in classifying the data into two classes of remarkably good accuracy and reliability predictions.

c) Support Vector Machines

A suitable Python code fragment facilitated the implementation of a Support Vector Machine (SVM) for classification tasks. The code begins by generating an SVC object with an 'rbf' kernel and a specified random state for reproducibility. The model was then trained on the training data (X_train, y_train). After training, the model was used to make predictions on the test data (X_test), generating the predicted labels (y_pred_svc). The classification report had the performance of the model in detail: precision, recall, F1-score, and support for each class. Besides, the accuracy score was computed as a measure that deals with the overall correctness of the predictions.

Output:

Table 3: Showcases the Support Vector Machine Classification Report



```
Support Vector Machine Classification Report Results
              precision    recall  f1-score   support

           0       0.99      1.00      0.99     26622
           1       1.00      0.99      0.99     20499

    accuracy                           0.99     47121
   macro avg       0.99      0.99      0.99     47121
weighted avg       0.99      0.99      0.99     47121

Accuracy: 0.99
```

The overall classification report of the SVM model is very impressive. In class 0, the model has performed well with high precision (0.99), recall (1.00), and F1-score (0.99); hence, it is very well at selecting true positive instances of this class. Similarly, for class 1, the precision (1.00), recall (0.99), and F1-score (0.99) are good, meaning true positive instances of this class can be correctly classified by the model. The overall accuracy of the model is 0.99, with the model correctly predicting the class labels for about 99% of the test instances. These results thus indicate that the SVM model will be effective in classifying the data into the two classes with high accuracy and balanced performances across both classes.

ROC Curve Model Comparison

The code snippet in Python compared three classification models: Logistic Regression, Random Forest, and SVM. For each model, it calculated the ROC curve and AUC. For those whose default is to predict with predict_proba such as Logistic Regression and Random Forest, it took the probabilities to plot the ROC curve. In the case of the SVM, which does not have predict_proba, the decision function was used instead. All the ROC curves were plotted on the same graph for visual comparison. The AUC scores for each model will also be printed as a numerical measure for comparison. This analysis facilitated an understanding of the relative strengths and weaknesses of the different models, aiding in model selection and decision-making.

Output:

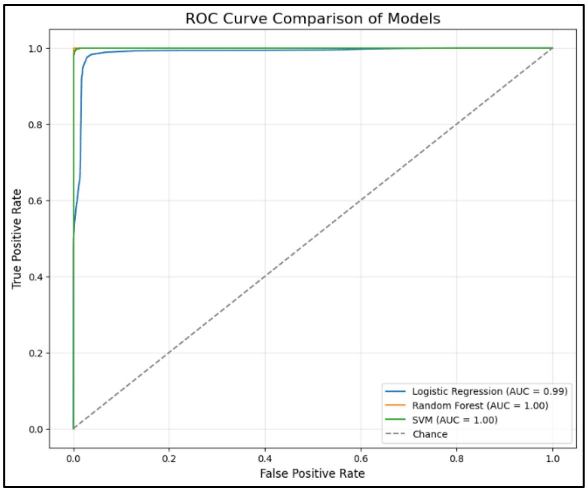


Figure 5: Visualizes ROC Curve Comparison of Models

This ROC curve comparison presents three classification models: Logistic Regression, Random Forest, and SVM. It plots the TPR against FPR at various settings of threshold values. The closer it goes to the upper-left corner, the better the performance. Performance has been quantified by AUC. In this case, the Random Forest and the SVM models show an AUC score of 1.00, indicating outstanding discrimination between the positive and negative classes. The Logistic Regression model also yields quite a good AUC of 0.99. From a visual point of view, this would indicate that the Random Forest and the SVM are the most capable classifiers for this dataset, while the Logistic Regression provides results that are slightly less discriminative.

Comparison of Models Accuracy

The executed code snippet in Python compared the performance of a few classification models by calculating their training and testing accuracy. Then, it created a histogram chart comparing the training and testing accuracies. Bars were colored differently to differentiate between the training and testing accuracies. Bars were annotated with their accuracy values. It finally analyzed the gap between the training and testing accuracies for probable overfitting. Models with big gaps between training and testing accuracies were flagged for probable overfitting, while models with smaller gaps were able to generalize well. This analysis helped in choosing the best model and also finding possible issues like overfitting.

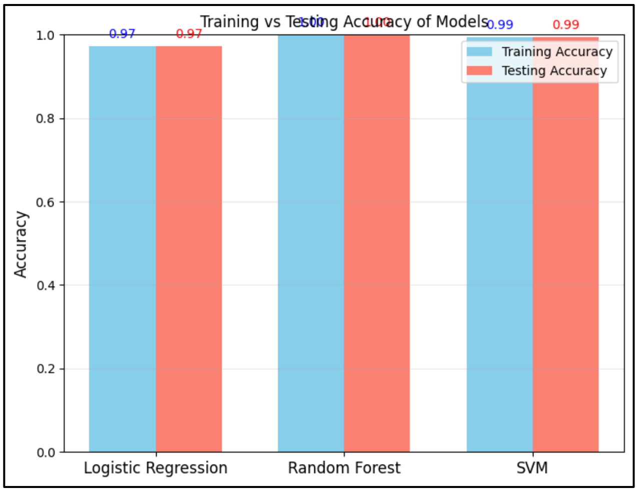


Figure 6: Depicts Training vs. Testing Accuracy of Models

The above bar chart shows the training and testing accuracies of three classification models: Logistic Regression, Random Forest, and Support Vector Machine. The blue bars represent the training accuracy, while the red bars represent the test accuracy. Both Random Forest and SVM have very similar training and test accuracies, indicating that they generalize well to unseen data and are not prone to overfitting. Logistic Regression presents slightly larger differences between the accuracy obtained for training and testing, which can be considered overfitting. Retrospectively, the Random Forest and SVM models seem to be robust and reliable, placing probably the Random Forest slightly above, given their performance is nearly perfect for training and testing.

Feature Importance and Analysis

Understanding the feature importance of a machine learning model sheds light on the data itself and can be used for better interpretability. With algorithms like Random Forest and Gradient Boosting, the feature importance score is handy in understanding the value of each feature in the prediction made by the model. By evaluating feature importance scores, the analyst can pinpoint the key features that substantially influence resource optimization. These features might include:

Historical Resource Utilization: Past trends in resource consumption can be used to predict future trends and optimize resource utilization.

Workload Patterns: This is similarly critical as it defines the workload nature, for instance, the peak times of use and which applications use what amount of resources.

External Factors: Weather conditions, holidays, or economic indicators are some of the factors that may influence the demand for resources and hence must be considered while performing optimization.

Hardware/Software Configuration: Specific hardware and software configurations at the Data Center level influence resource utilization and should be factored into the optimization models at work.

Handling Incomplete Data

Incomplete data is a prevalent challenge in real-world datasets, and it can substantially influence the performance of machine learning models. This can be addressed in several ways:

Imputation: This is the process of filling up empty values with estimated values. The most common of such imputation techniques include mean imputation, median imputation, and mode imputation, and some of their advanced versions include KNN imputation and multiple imputation.

Deletion: A straightforward approach is to remove instances with missing values. However, this can lead to significant data loss, especially if the missing values are not missing at random.

Model-based imputation: This would involve the training of a predictive model for observing the missing values given the rest of the features.

Advanced Techniques: More advanced techniques, such as multiple imputation and probabilistic imputation, account for uncertainty in the imputation process itself.

Implications for Data Center Management

The application of machine learning techniques holds huge potential for the reformation of resource management in US data centers. These models analyze a pattern in historical data to predict future resource demands, thus allowing optimized resource allotment and minimizing operational costs. It would also be possible to use machine learning algorithms in forecasting the periods of peak use so that the data center operators can pre-scale their resources and avoid service disruption. Besides that, predictive maintenance models show the possibility of hardware failures before they occur, which reduces both downtime and maintenance costs. Some recommendations for effectively integrating machine learning into data center operations could be:

Data Quality and Preparation: Basically, the quality of data and how well it is prepared is considered vital in training good machine learning models. It is very important to clean, pre-process, and feature-engineer with the intent of sustaining data integrity and relevance.

Model Selection and Training: The usage of a machine learning algorithm should be relevant to a certain use case and the nature of the data. Model training requires careful considerations together with hyperparameter tuning.

Continuous Monitoring and Evaluation: A machine learning model should be continuously monitored and evaluated regarding performance and other potential issues that can be resolved. Regular retraining may be required given changing data patterns and operational needs.

Collaboration and Skills: The use of machine learning within data centers requires collaboration among data scientists, IT professionals, and domain experts. A multidisciplinary approach will be able to ensure models are aligned with business objectives and technical constraints.

Ethics: The design and deployment of machine learning have to be considered in ethical dimensions, such as data privacy and bias. Operations of AI must be fair and accountable, embracing transparent and responsible AI practices.

Business Development Impact in the USA

Efficient resource usage management is critical for data centers, as it directly impacts operational costs, performance, and sustainability. This project provides significant business value in the following ways:

1. **Operational Efficiency:**

o Accurate classification of resource usage enables proactive management of workloads, preventing resource bottlenecks and over-utilization.

2. **Cost Optimization:**

o Identifying underutilized or overutilized resources helps in reallocation and scaling decisions, reducing unnecessary operational expenses.

3. Improved Decision-Making:

o Data-driven insights into resource patterns support better scheduling and allocation strategies, improving the reliability of data center services.

4. Sustainability:

o Efficient resource utilization reduces energy consumption and carbon footprint, aligning operations with environmental goals.

5. Fault Prevention:

o Predicting abnormal usage patterns allows for early detection of potential issues, minimizing downtime, and maintaining service-level agreements (SLAs).

Challenges and Limitations

The deployment of machine learning to data center resource optimization is not without its limitations and challenges. Quite a few large ethical worries remain concerning the use of operational data for analytics; there is an amount of concerned sentiment as to the privacy and security of this data. Also, quite an amount of informed decisions should be taken so that this does not end up with biases within that data, leading towards an outcome that can be unjust and biased. Other important reasons that may hamper the performance of the models of machine learning relate to data quality poor quality means inaccurate, incomplete, or noisy data. The prediction resulting from such data would thus be biased and unreliable. In machine learning, interpretability for complicated models is hard since one does not understand why particular decisions were made. Finally, the models trained on particular data centers will most likely be narrow in generalization when applied to other data centers that are characterized by different configurations and operational features.

Future Research Directions

To address the limitations, and challenges and further advance the domain of data center resource optimization, several future research directions can be explored. For instance, the use of even larger and more diverse data sets may be one venue to improve the models by providing higher accuracy and generalizability. This could be made possible through data center collaborations where anonymized and aggregated data can be shared, creating a larger dataset that can be representative. Another exciting research area is the development of real-time resource optimization and predictive maintenance techniques. Using advanced machine learning algorithms coupled with real-time streams of data, timely decisions can be made that will go a long way in improving resource utilization and system reliability. That implies developing sophisticated anomaly detection systems that can identify issues well in advance before they strike or predictive maintenance models that can forecast failures quite accurately.

CONCLUSION

This study aimed to explore how different ML techniques can be used for optimizing resource utilization by data centers in the US, focusing on strategies to handle incomplete data and implications for business development. The dataset was retrieved from the GitHub repository, which provided a rich dataset of resource usage metrics and operational data from U.S. data centers. It contained complex and fine-grained information that was necessary to optimize data center performance and deal with incomplete data challenges. A detailed description of the dataset and its key attributes was provided. It was designed for analyzing resource usage patterns in data centers, putting much emphasis on energy efficiency, workload distribution, and operational reliability. This integration of time-series data with sensor readings and performance logs provided a comprehensive overview of resource consumption and environmental conditions in data center operations. This dataset was curated for the engagement of machine learning models in the study and optimization of resource consumption along with the challenges of missing data. Analysis of resource utilization in US data centers was accomplished using the application of various models for machine learning, most notably, Logistic Regression, Random Forest, and Support Vector Machines; Retrospectively, the Random Forest and SVM models seem to be robust and reliable, placing the Random Forest slightly above, given their performance is nearly perfect for training and testing. The application of machine learning techniques holds huge potential for the reformation of resource management in US data centers. These models analyze a pattern in historical data to predict future resource demands, thus allowing optimized resource allotment and minimizing operational costs.

REFERENCES

1. Abdulazeez, D. H., & Askar, S. K. (2024). A Novel Offloading Mechanism Leveraging Fuzzy Logic and Deep Reinforcement Learning to Improve IoT Application Performance in a Three-Layer Architecture Within the Fog-Cloud Environment. IEEE Access.
2. Alam, M., Islam, M. R., & Shil, S. K. (2023). AI-Based Predictive Maintenance for US Manufacturing: Reducing Downtime and Increasing Productivity. International Journal of Advanced Engineering Technologies and Innovations, 1(01), 541-567.
3. Al Mukaddim, A., Mohaimin, M. R., Hider, M. A., Karmakar, M., Nasiruddin, M., Alam, S., & Anonna, F. R. (2024). Improving Rainfall Prediction Accuracy in the USA Using Advanced Machine Learning Techniques. Journal of Environmental and Agricultural Studies, 5(3), 23-34.
4. Al Mukaddim, A., Nasiruddin, M., & Hider, M. A. (2023). Blockchain Technology for Secure and Transparent Supply Chain Management: A Pathway to Enhanced Trust and Efficiency. International Journal of Advanced Engineering Technologies and Innovations, 1(01), 419-446.
5. Buiya, M. R., Laskar, A. N., Islam, M. R., Sawalmeh, S. K. S., Roy, M. S. R. C., Roy, R. E. R. S., & Sumsuzoha, M. (2024). Detecting IoT Cyberattacks: Advanced Machine Learning Models for Enhanced Security in Network Traffic. Journal of Computer Science and Technology Studies, 6(4), 142-152.
6. Buiya, M. R., Alam, M., & Islam, M. R. (2023). Leveraging Big Data Analytics for Advanced Cybersecurity: Proactive Strategies and Solutions. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 882-916.
7. Debnath, P., Karmakar, M., Khan, M. T., Khan, M. A., Al Sayeed, A., Rahman, A., & Sumon, M. F. I. (2024). Seismic Activity Analysis in California: Patterns, Trends, and Predictive Modeling. Journal of Computer Science and Technology Studies, 6(5), 50-60.
8. Ezeigweneme, C. A., Umoh, A. A., Ilojianya, V. I., & Adegbite, A. O. (2024). Telecommunications energy efficiency: optimizing network infrastructure for sustainability. Computer Science & IT Research Journal, 5(1), 26-40.
9. Feng, Y., Qi, Y., Li, H., Wang, X., & Tian, J. (2024, July). Leveraging federated learning and edge computing for recommendation systems within cloud computing networks. In Third International Symposium on Computer Applications and Information Systems (ISCAIS 2024) (Vol. 13210, pp. 279-287). SPIE.
10. Gazi, M. S., Nasiruddin, M., Dutta, S., Sikder, R., Huda, C. B., & Islam, M. Z. (2024). Employee Attrition Prediction in the USA: A Machine

Learning Approach for HR Analytics and Talent Retention Strategies. Journal of Business and Management Studies, 6(3), 47-59.

11. Gebreyesus, Yibrah, Damian Dalton, Sebastian Nixon, Davide De Chiara, and Marta Chinnici. "Machine learning for data center optimizations: feature selection using Shapley additive exPlanation (SHAP)." Future Internet 15, no. 3 (2023): 88.

12. Gong, Y., Huang, J., Liu, B., Xu, J., Wu, B., & Zhang, Y. (2024). Dynamic resource allocation for virtual machine migration optimization using machine learning. arXiv preprint arXiv:2403.13619.

13. Hasanuzzaman, M., Hossain, S., & Shil, S. K. (2023). Enhancing Disaster Management through AI-Driven Predictive Analytics: Improving Preparedness and Response. International Journal of Advanced Engineering Technologies and Innovations, 1(01), 533-562.

14. He, N., Yang, S., Li, F., Trajanovski, S., Zhu, L., Wang, Y., & Fu, X. (2023). Leveraging deep reinforcement learning with attention mechanism for virtual network function placement and routing. IEEE Transactions on Parallel and Distributed Systems, 34(4), 1186-1201.

15. Islam, M. R., Shawon, R. E. R., & Sumsuzoha, M. (2023). Personalized Marketing Strategies in the US Retail Industry: Leveraging Machine Learning for Better Customer Engagement. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 750-774.

16. Islam, M. R., Nasiruddin, M., Karmakar, M., Akter, R., Khan, M. T., Sayeed, A. A., & Amin, A. (2024). Leveraging Advanced Machine Learning Algorithms for Enhanced Cyberattack Detection on US Business Networks. Journal of Business and Management Studies, 6(5), 213-224.

17. Islam, M. Z., Islam, M. S., Al Montaser, M. A., Rasel, M. A. B., Bhowmik, P. K., & Dalim, H. M. (2024). EVALUATING THE EFFECTIVENESS OF MACHINE LEARNING ALGORITHMS IN PREDICTING CRYPTOCURRENCY PRICES UNDER MARKET VOLATILITY: A STUDY BASED ON THE USA FINANCIAL MARKET. The American Journal of Management and Economics Innovations, 6(12), 15-38.

18. Karmakar, M., Debnath, P., & Khan, M. A. (2024). AI-Powered Solutions for Traffic Management in US Cities: Reducing Congestion and Emissions. International Journal of Advanced Engineering Technologies and Innovations, 2(1), 194-222.

19. Khan, M. T., Akter, R., Dalim, H. M., Sayeed, A. A., Anonna, F. R., Mohaimin, M. R., & Karmakar, M. (2024). Predictive Modeling of US Stock Market and Commodities: Impact of Economic Indicators and Geopolitical Events Using Machine. Journal of Economics, Finance and Accounting Studies, 6(6), 17-33.

20. Krothapalli, B., Shanmugam, L., & Mohammed, S. B. (2023). Machine Learning Algorithms for Efficient Storage Management in Resource-Limited Systems: Techniques and Applications. Journal of Artificial Intelligence Research and Applications, 3(1), 406-442.

21. Kumari, S. (2024). Cloud Transformation for Mobile Products: Leveraging AI to Automate Infrastructure Management, Scalability, and Cost Efficiency. Journal of Computational Intelligence and Robotics, 4(1), 130-151.

22. Li, H., Wang, X., Feng, Y., Qi, Y., & Tian, J. (2024). Integration Methods and Advantages of Machine Learning with Cloud Data Warehouses. International Journal of Computer Science and Information Technology, 2(1), 348-358.

23. Ma, Y., Shen, Z., & Shen, J. (2024). Cloud Computing and Hyperscale Data Centers: A Comparative Study of Usage Patterns. Journal of Theory and Practice of Engineering Science, 4(06), 11-19.

24. Nasiruddin, M., Al Mukaddim, A., & Hider, M. A. (2023). Optimizing Renewable Energy Systems Using Artificial Intelligence: Enhancing Efficiency and Sustainability. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 846-881.

25. Okusi, O. (2024). Leveraging AI and machine learning for the protection of critical national infrastructure. Asian Journal of Research in Computer Science, 17(10), 1-11.

26. Pro-AI-Rokibul. (2024). Enhancing-Data-Center-Operations-with-Machine-Learning-Addressing-Missing-Resource-Usage-Data/README.md at main · proAIrokibul/Enhancing-Data-Center-Operations-with-Machine-Learning-Addressing-Missing-Resource-Usage-Data. GitHub. https://github.com/proAIrokibul/Enhancing-Data-Center-Operations-with-Machine-Learning-Addressing-Missing-Resource-Usage-Data/blob/main/README.md

27. Rahman, A., Debnath, P., Ahmed, A., Dalim, H. M., Karmakar, M., Sumon, M. F. I., & Khan, M. A. (2024). Machine learning and network analysis for financial crime detection: Mapping and identifying illicit transaction patterns in global black money transactions. Gulf Journal of Advance Business Research, 2(6), 250-272.

28. Rahman, M. K., Dalim, H. M., & Hossain, M. S. (2023). AI-Powered Solutions for Enhancing National Cybersecurity: Predictive Analytics and Threat Mitigation. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 1036-1069.

29. Shawon, R. E. R., Dalim, H. M., Shil, S. K., Gurung, N., Hasanuzzaman, M., Hossain, S., & Rahman, T. (2024). Assessing Geopolitical Risks and Their Economic Impact on the USA Using Data Analytics. Journal of Economics, Finance and Accounting Studies, 6(6), 05-16.

30. Shawon, R. E. R., Miah, M. N. I., & Islam, M. Z. (2023). Enhancing US Education Systems with AI: Personalized Learning and Academic Performance Prediction. International Journal of Advanced Engineering Technologies and Innovations, 1(01), 518-540.

31. Shawon, R. E. R., Rahman, A., Islam, M. R., Debnath, P., Sumon, M. F. I., Khan, M. A., & Miah, M. N. I. (2024). AI-Driven Predictive Modeling of US Economic Trends: Insights and Innovations. Journal of Humanities and Social Sciences Studies, 6(10), 01-15.

32. Shawon, R. E. R., Chowdhury, M. S. R., & Rahman, T. (2023). Transforming Urban Living in the USA: The Role of IoT in Developing Smart Cities. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 14(1), 917-953.

33. Sumon, M. F. I., Osiujjaman, M., Khan, M. A., Rahman, A., Uddin, M. K., Pant, L., & Debnath, P. (2024). Environmental and Socio-Economic Impact Assessment of Renewable Energy Using Machine Learning Models. Journal of Economics, Finance and Accounting Studies, 6(5), 112-122.

34. Shil, S. K., Chowdhury, M. S. R., Tannier, N. R., Tarafder, M. T. R., Akter, R., Gurung, N., & Sizan, M. M. H. (2024). Forecasting Electric Vehicle Adoption in the USA Using Machine Learning Models. Journal of Computer Science and Technology Studies, 6(5), 61-74.

35. Wang, Y., Bao, Q., Wang, J., Su, G., & Xu, X.

(2024). Cloud Computing for Large-Scale Resource Computation and Storage in Machine Learning. Journal of Theory and Practice of Engineering Science, 4(03), 163-171.

36. Zhang, Y., Liu, B., Gong, Y., Huang, J., Xu, J., & Wan, W. (2024, April). Application of machine learning optimization in cloud computing resource scheduling and management. In Proceedings of the 5th International Conference on Computer Information and Big Data Applications (pp. 171-175).